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

Simon Bussy[1] †, Raphaël Veil[2,3], Vincent Looten[2,3], Anita Burgun[2,3], Stéphane Gaïffas[1,4], Agathe Guilloux[5], Brigitte Ranque[6,7], Anne-Sophie Jannot[2,3]

[1] LPSM, UMR 8001, Sorbonne University, Paris, France
[2] APHP, Département d'Informatique Biomédicale et de Santé Publique, HEGP, Paris, France
[3] INSERM, UMRS 1138, Eq22, Centre de Recherche des Cordeliers, Université Paris Descartes, Paris, France
[4] CMAP, UMR 7641, École Polytechnique CNRS, Paris, France
[5] LAMME, Université Evry, CNRS, Université Paris-Saclay, Paris, France
[6] INSERM, UMRS 970, Université Paris Descartes, Paris, France
[7] APHP, Département de Médecine Interne, HEGP, Paris, France

Résumé : La production d'un algorithme d'intelligence artificielle (IA) à partir de données de vie réelle réside dans l'intégration et la transformation d'un très grand nombre de covariables (grande dimension), puis dans la construction d'un modèle d'apprentissage. Dans cet article, nous allons détailler les différentes étapes de ce processus à travers une illustration sur des données de soin réutilisées pour prédire la réadmission à l'hôpital après une crise vaso-occlusive chez les patients atteints de drépanocytose. Dans notre étude, les covariables sont extraites d'un entrepôt de données et sont interprétées grâce à l'utilisation d'outils d'extraction d'information textuelle et de transformation de variables longitudinales. Un total de 174 covariables ont alors été créées à partir du dossier patient. Différentes méthodes d'apprentissage ont ensuite été comparées pour finalement proposer l'algorithme final. Cette étude illustre d'une part la nécessité de disposer d'un grand nombre d'outils pour réutiliser les données de soin, et d'autre part l'importance de considérer différentes méthodes d'apprentissage pour construire un algorithme de prédiction performants.

Mots-clés : design d'algorithme d'IA, données longitudinales, prédiction en grande dimension.

1 Introduction

Dans le domaine de la santé, de nombreux progrès sont attendus de l'intelligence artificielle, à savoir la production et la mise en œuvre d'algorithmes performants, notamment pour mieux prédire le pronostic des patients, qu'il s'agisse de guérison, décès, réhospitalisation, ou encore récidive, et ainsi proposer un traitement personnalisé pour chaque patient. Les données utilisées pour produire ces algorithmes en médecine sont des données de vie réelle et/ou des connaissances expertes. Les données de vie réelle sont utilisées pour correspondre au mieux à la pratique clinique quotidienne et pour produire des algorithmes pouvant s'appliquer à tout type de patient. L'utilisation de ces données de vie réelle pose de nombreuses difficultés : la quantité d'information est gigantesque, elles sont stockées dans les dossiers médicaux électroniques (*electronic health record* en anglais, EHR) qui caractérisent le patient et son suivi, et elles nécessitent souvent d'être interprétées. En effet, l'information pertinente pour le pronostic n'est généralement pas identifiée *a priori*. Tout l'enjeu est donc à la fois de construire et sélectionner des covariables pertinentes, puis d'élaborer un modèle pronostique performant. Dans cet article, nous allons détailler les étapes suivies pour répondre à ces questions à travers une illustration portant sur la réadmission à l'hôpital après une crise vaso-occlusive (CVO) chez les patients atteints de drépanocytose.

La drépanocytose est la maladie monogénique la plus fréquente dans le monde. Elle est responsable de CVO répétées, qui sont des épisodes douloureux aigus, entraînant une augmentation de la morbidité et de la mortalité (Bunn, 1997). Le traitement des CVO est avant tout symptomatique : il s'agit d'hydrater le patient et de calmer les douleurs grâce aux opiacés. Lorsque les douleurs et les besoins en antalgiques s'estompent, la crise est considérée

†. Contact: simon.bussy@gmail.com

comme terminée et le patient peut sortir. Bien qu'il existe des études sur les facteurs de risque de complications précoces, très peu d'entre elles ont spécifiquement abordé la question de la réhospitalisation précoce (Brousseau *et al.*, 2010). Nous considérons comme précoce toute réadmission survenant au cours des 30 premiers jours suivant la sortie d'hospitalisation, ce choix de seuil étant standard dans les études qui traitent de la drépanocytose (Brousseau *et al.*, 2010). On sait que les réadmissions à l'hôpital sont responsables d'énormes coûts, et sont également une mesure de la qualité des soins. Les hôpitaux ayant des ressources limitées, l'identification des patients à haut risque de réadmission est une question primordiale et des modèles prédictifs sont souvent considérés pour tenter d'y remédier.

Pour l'intégration et la transformation des données de vie réelle, nous détaillerons différentes options et outils de visualisation mis en place pour interpréter au mieux les données longitudinales. Pour la construction du modèle, nous discuterons des deux aspects (performance de la prédiction et sélection des covariables) pour toutes les méthodes considérées, avec un accent particulier mis sur la méthode de régularisation Elastic-Net (Zou & Hastie, 2005). La régularisation est apparue comme un thème dominant dans la sélection de variables à partir de données de grande dimension. Nous considérerons les méthodes suivantes : la régression logistique (LR) (Hosmer Jr *et al.*, 2013) et les SVM (Schölkopf & Smola, 2002) avec noyau linéaire, toutes deux pénalisées avec la régularisation Elastic-Net pour éviter le surapprentissage (Hawkins, 2004). Les forêts aléatoires (RF) (Breiman, 2001), le gradient boosting (GB) (Friedman, 2002) et les réseaux de neurones (NN) (Yegnanarayana, 2009) seront également considérés. Enfin, nous utiliserons le modèle de Cox (Cox, 1972), de CURE (Farewell, 1982) (qui considère une fraction de la population comme n'étant plus soumise à un risque de réadmission), et le modèle de mélange C-mix récemment développé (Bussy *et al.*, 2018) ; tous trois étant également pénalisés par la régularisation Elastic-Net.

2 Matériel et méthodes

2.1 Cas d'étude

Il s'agit d'une étude de cohorte rétrospective monocentrique faite sur 286 patients de l'Hôpital Européen Georges Pompidou (HEGP), dont le service de médecine interne fait partie de l'un des 3 centres experts parisiens de la drépanocytose chez l'adulte. L'HEGP dispose par ailleurs d'un entrepôt de données de santé (EDS) dont l'architecture repose sur le logiciel libre I2B2 permettant d'accéder facilement aux données par des requêtes SQL. L'EDS contient des données de routines, notamment les numéros déidentifiés des séjours et des patients, des données socio-démographiques, différents comptes-rendus textuels, les diagnostics (CIM-10), les procédures (CCAM), les résultats biologiques, les données de prescriptions (CPOE), ainsi qu'une terminologie de référence pour les codages de ces informations (LOINC).

Dans notre étude, les critères d'inclusion sont l'admission dans le service de médecine interne pour CVO (CIM-10 : D57-0) entre le 1er janvier 2010 et le 31 décembre 2015. Les critères d'exclusion sont le diagnostic d'addiction aux opiacés (CIM-10 : F11) et/ou la prescription de Buprénorphine ou de Méthadone, pour éviter tout facteur de confusion quant aux traitements et à la durée du séjour. Nous avons sélectionné aléatoirement un séjour par patient pour éviter le biais de surreprésentation des patients ayant de très nombreux séjours.

2.2 Variables

Nous avons extrait les variables suivantes pour l'ensemble des patients de l'étude :
- 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.

Pour faciliter l'extraction des données à partir des comptes-rendus, qui sont stockées dans l'EDS sous forme de texte libre (données non structurées), nous avons utilisé un outil développé en interne appelé FASTVISU (Escudié *et al.*, 2015). Ce logiciel, connecté avec l'EDS, permet de vérifier rapidement la présence de certains mots-clés définis au préalable à l'aide d'expressions régulières ; plusieurs mots-clés pouvant se rapporter à un concept particulier. En fonction du contexte, l'utilisateur peut ensuite renseigner la présence d'une comorbidité, d'un antécédent, ou la valeur de l'hémoglobine de base du patient par exemple. Certaines données étaient absentes des compte-rendus : les antécédents non-mentionnés ont été considérés comme « absent » ; l'hémoglobine de base à été imputée par la dernière valeur d'hémoglobine avant la sortie du 1er séjour d'hospitalisation du patient considéré.

Nous avons aussi dérivé de nouvelles covariables définies par les experts du domaine à partir des variables extraites :

— chaque variable catégorielle a été binarisée,
— 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.

Pour ces 2 dernières, lorsque les patients n'avaient pas reçu d'opiacé et/ou d'oxygène pendant le séjour, nous avons considéré que ces traitements avaient été donnés au moins aux urgences conformément au protocole habituel. Nous avons donc imputés ces variables par le délai entre l'hospitalisation et la sortie.

Pour chaque variable longitudinale (qui dépend du temps), le nombre de points et leur position sur la série temporelle diffèrent pour chaque patient. Nous avons alors utilisé la méthode suivante pour décrire et visualiser la trajectoire de chacune de ces variables en stratifiant les patients en deux groupes suivant leur délai de réadmission (avec le seuil de 30 jours) :

1. création d'une grille temporelle entre la première et la dernière mesure de la variable considérée, en fixant comme t_0 la date d'hospitalisation (les valeurs mesurées aux urgences sont donc placées à $t < t_0$),
2. création d'un nuage de point global (de tous les patients) avec le temps en abscisse et les valeurs de la variable en ordonnées,
3. création d'une trajectoire moyenne avec intervalle de confiance, selon la procédure suivante :
   — un spline est ajusté pour chaque trajectoire individuelle (pour inférer une représentation fonctionnelle "lisse"),
   — on calcule les valeurs prises par le spline sur la grille définie au point 1.,
   — on obtient une matrice pour chaque variable avec une ligne par patient (le nombre de colonnes correspondant à la taille de la grille),
   — on calcule la valeur moyenne ainsi avec intervalle de confiance gaussien à 95% pour chaque temps de la grille (colonne de la matrice).

Par ailleurs, au-delà de l'aspect visualisation, nous avons extrait des covariables dérivées des différentes séries temporelles (sur les 48 dernières heures des séjours) pour être utilisées par les méthodes d'apprentissage :

— dernière valeur disponible avant la sortie,
— pente de la régression linéaire ajustée aux données,
— paramètres d'un processus gaussien (GP) ajusté aux données.

Les GPs sont connus pour bien ajuster et expliquer des données EHR (Pimentel *et al.*, 2013). Nous avons utilisé des GPs avec fonction moyenne linéaire et une somme de noyau (un noyau constant qui modifie la moyenne, un noyau à fonctions de base radiales et un noyau blanc pour expliquer la composante de bruit du signal).

3 Construction de l'algorithme

Nous considérons d'abord les méthodes LR et SVM, toutes deux pénalisées avec la régularisation Elastic-Net. Un avantage de cette méthode de régularisation est sa capacité à effectuer une sélection de covariables (partie lasso) et donc identifier les covariables les plus prédictives, mais également à gérer la corrélation potentielle entre covariables (partie ridge). L'utilisation de cette pénalité permet de surcroît d'identifier d'éventuels facteurs de confusion. Pour toute les méthodes utilisant la régularisation Elastic-Net, le paramètre de pénalisation est choisi par cross-validation (grid-search).

Nous considérons également d'autres algorithmes d'apprentissage tels que les RF et le GB. Pour ces deux algorithmes, tous les hyper-paramètres sont estimés en utilisant une procédure de cross-validation avec recherche aléatoire (Kohavi *et al.*, 1995) (par exemple pour les RF : le nombre d'arbres dans la forêt, la profondeur maximale de l'arbre ou le nombre minimum d'échantillons requis pour diviser un nœud interne). Pour ces deux méthodes, l'importance des covariables est mesurée en utilisant le critère de Gini (Menze *et al.*, 2009), définie comme la diminution totale de l'impureté du nœud pondérée par la probabilité d'atteindre ce nœud (approximée par la proportion d'échantillons atteignant ce nœud) moyenné sur tous les arbres de l'ensemble. Enfin, nous considérons un NN, un perceptron avec une couche cachée, entrainé à l'aide d'un algorithme d'optimisation stochastique, avec comme fonction d'activation des unités linéaires rectifiées (ReLU) pour obtenir une activation sparse (Glorot *et al.*, 2011) et être capable de comparer la sélection des covariables avec les autres méthodes régularisées. Le terme de régularisation ainsi que le nombre de neurones dans la couche cachée sont également choisis par cross-validation avec recherche aléatoire.

Pour tous ces modèles, nous avons utilisé les implémentations de référence de la bibliothèque `scikit-learn` (Pedregosa *et al.*, 2011). Nous avons aussi utilisé le modèle de Cox pénalisé (Simon *et al.*, 2011), le modèle CURE (Farewell, 1982) qui considère qu'une fraction de la population n'est pas sujette à un risque de réadmission (avec une fonction logistique pour la partie incidence et un modèle de survie paramétrique), et le modèle C-mix (Bussy *et al.*, 2018), conçu pour déceler des sous-groupes de population à risque de réadmission variable dans un contexte de survie en grande dimension.

Les données ont été divisées aléatoirement en un échantillon d'apprentissage et un échantillon de validation (30%) pour l'estimation de la performance de prédiction mesurée par l'AUC (Area Under the Curve) dans notre étude (Bradley, 1997). Pour les méthodes de survie (Cox, CURE et C-mix), la prédiction est issue de la fonction de survie estimée pour chaque méthode, évaluée au seuil de 30 jours.

4 Résultats

4.1 Visualisation des trajectoires des variables temporelles

La méthode décrite permet de comparer l'évolution des variables qui dépendent du temps, en fonction du délai de réadmission (avant ou après 30 jours). Pour la visualisation, nous avons alors exclu les patients qui présentent des facteurs de confusion potentiels (e.g. transfusion ou infection pour supprimer les variations en rapport avec une autre affection qu'une CVO). Deux variables, qui sont liées entre elles, montrent des trajectoires clairement différentes selon le délai de réadmission : l'hémoglobine et l'hématocrite. Elles sont représentées Figure 1.

Pour être plus précis, on observe les écarts suivants :

— la valeur moyenne d'hémoglobine diminue de 9 g/dL à 7 g/dL environ pour le groupe "réadmission précoce", tandis qu'il reste stable autour de 9 g/dL pour les autres patients,

— la valeur moyenne d'hématocrite diminue de 25% à 20% environ pour le groupe "réadmission précoce", tandis qu'il reste stable autour de 25% pour les autres patients.

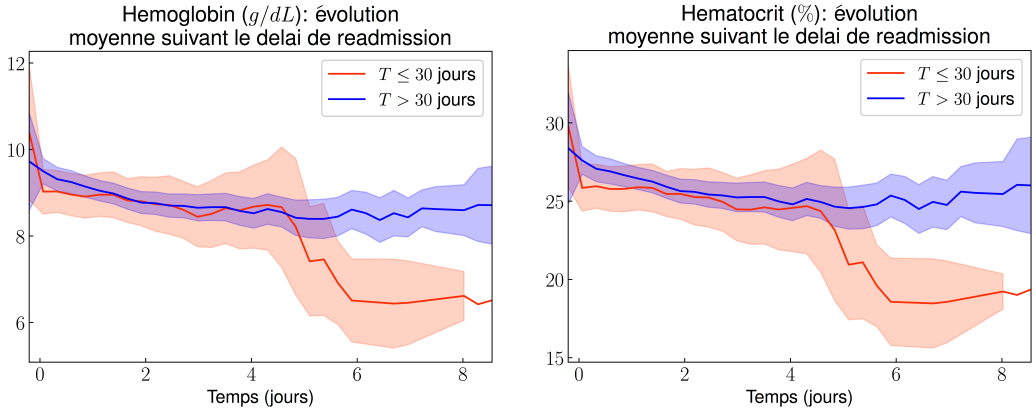

FIGURE *1 – La trajectoire de ces variables chez les patients qui reviennent aux urgences dans les 30 jours qui suivent leur sortie d'hospitalisation montre clairement une diminution autour du 5ème jour d'hospitalisation, tandis que la trajectoire est stable chez les autres patients.*

4.2 Performance des algorithmes et sélection des covariables prédictives

La Table 1 donne les performances en prédiction des différents algorithmes. Le C-mix

TABLE *1 – Comparaison de la performance de prédiction en terme d'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**

obtient les meilleures performances. L'importance des différentes variables dans chaque algorithme varie considérablement, comme le montre la Figure 2. Parmi les 20 variables sé-

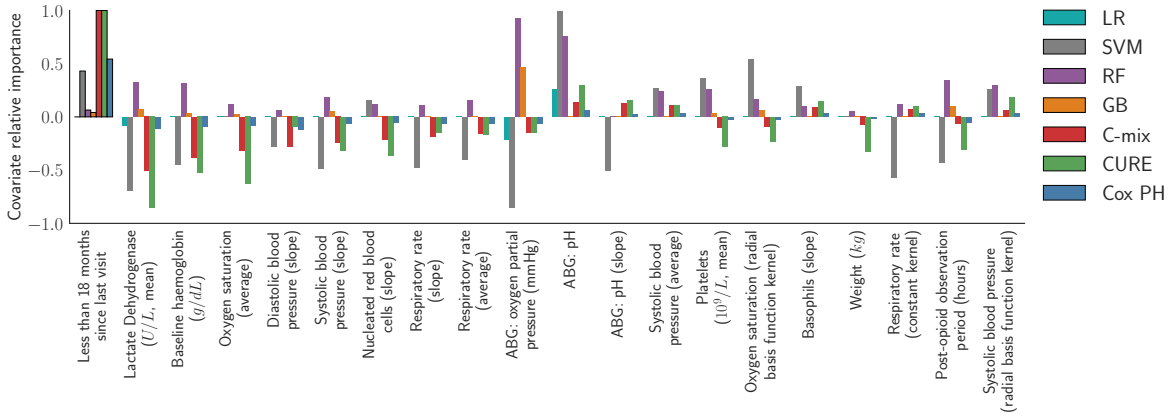

FIGURE *2 – Comparaison de l'importance des 20 premières covariables (ordonnées suivant le C-mix, le modèle le plus performant).*

lectionnées par le C-mix, 9 sont des variables dérivées de l'évolution temporelle de variables longitudinales (6 pentes et 3 hyper-paramètres de processus gaussien).

5 Discussion

Cette étude illustre les difficultés rencontrées lors de la réutilisation de données de soin pour construire un algorithme d'IA. Cela implique de disposer non seulement d'outils pour stocker et réorganiser les données afin de pouvoir les extraire facilement (comme l'entrepôt) ; mais aussi d'outils facilitant l'extraction textuelle, ou encore permettant de modéliser l'information longitudinale. Dans notre étude, ceci nous a permis d'extraire 174 covariables à partir du dossier patient. Les méthodes de construction d'algorithme donnent dans notre illustration des performances très différentes avec des variables sélectionnées variant d'une méthode à une autre. Cela illustre la nécessité d'avoir une approche exhaustive sur les méthodes disponibles pour construire un algorithme final performant pour la tâche de prédiction d'intérêt.

Dans cette étude, l'extraction de données a été effectuée sans *a priori* sur la pertinence de chaque covariable. Par exemple, nous avons extrait toutes les covariables biologiques qui ont été mesurées pendant le séjour d'un patient, sans présumer de leur importance sur le risque de réadmission, bien que des choix aient été faits pour modéliser ces variables de façon pertinente. Après avis d'expert, les variables sélectionnées font sens d'un point de vue clinique, mettant en évidence la capacité des méthodes de régularisation à identifier les covariables cliniquement pertinentes. Les covariables les plus importantes sont liées à la gravité de la maladie drépanocyaire sous-jacente (par exemple, fréquence de crise, hémoglobine de base), et les paramètres biologiques de crise classiquement monitorés (par exemple, le lactate déshydrogénase). Les différences de sélection suivant les modèles semblent être liées aux hypothèses sous-jacentes de chaque méthode : plutôt des variables en rapport avec la gravité de la maladie sous-jacente pour les modèles de survie, et des variables liées à la crise pour les autres modèles. Ceci souligne, à nouveau, la nécessité d'utiliser plusieurs méthodes pour avoir une interprétation des covariables.

Notre étude met en exergue les étapes pour développer des algorithmes performants dans le champ de la médecine personnalisée. Les algorithmes développés prédisent un risque de réadmission qui pourraient aider les médecins à décider si un patient spécifique peut ou non sortir de l'hôpital. Néanmoins, la plupart des covariables sélectionnées ont été dérivées de données brutes ou non structurées, ce qui rend difficile la mise en œuvre de ces algorithmes directement à partir du dossier de soins électronique dans notre hôpital. Il serait aussi nécessaire d'estimer la robustesse de l'approche et de prévoir une validation sur une cohorte externe ou sur une cohorte prospective. De plus, il n'est pas possible actuellement d'implémenter ce type d'algorithme dans le dossier de soin. Il serait donc nécessaire de travailler avec les entreprises produisant les dossiers de soins pour que de telles fonctionnalités soient offertes.

Références

BRADLEY A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, **30**(7), 1145–1159.

BREIMAN L. (2001). Random forests. *Machine learning*, **45**(1), 5–32.

BROUSSEAU D. C., OWENS P. L., MOSSO A. L., PANEPINTO J. A. & STEINER C. A. (2010). Acute care utilization and rehospitalizations for sickle cell disease. *Jama*, **303**(13), 1288–1294.

BUNN F. H. (1997). Pathogenesis and treatment of sickle cell disease. *New England Journal of Medicine*, **337**(11), 762–769.

BUSSY S., GUILLOUX A., GAÏFFAS S. & JANNOT A.-S. (2018). C-mix : A high-dimensional mixture model for censored durations, with applications to genetic data. *Statistical Methods in Medical Research*, **0**(0), 0962280218766389.

COX D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, **34**(2), 187–220.

ESCUDIÉ J.-B., JANNOT A.-S., ZAPLETAL E., COHEN S., MALAMUT G., BURGUN A. & RANCE B. (2015). Reviewing 741 patients records in two hours with fastvisu. In *AMIA Annual Symposium Proceedings*, volume 2015, p. 553 : American Medical Informatics Association.

FAREWELL V. T. (1982). The use of mixture models for the analysis of sureval data with long-term survivors. *Biometrics*, **38**(4), 1041–1046.

FRIEDMAN J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, **38**(4), 367–378.

GLOROT X., BORDES A. & BENGIO Y. (2011). Deep sparse rectifier neural networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, p. 315–323.

HAWKINS D. M. (2004). The problem of overfitting. *Journal of chemical information and computer sciences*, **44**(1), 1–12.

HOSMER JR D. W., LEMESHOW S. & STURDIVANT R. X. (2013). *Applied logistic regression*, volume 398. John Wiley & Sons.

KOHAVI R. *et al.* (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, p. 1137–1145 : Stanford, CA.

MENZE B. H., KELM B. M., MASUCH R., HIMMELREICH U., BACHERT P., PETRICH W. & HAMPRECHT F. A. (2009). A comparison of random forest and its gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC bioinformatics*, **10**(1), 213.

PEDREGOSA F., VAROQUAUX G., GRAMFORT A., MICHEL V., THIRION B., GRISEL O., BLONDEL M., PRETTENHOFER P., WEISS R., DUBOURG V. *et al.* (2011). Scikit-learn : Machine learning in python. *Journal of machine learning research*, **12**(Oct), 2825–2830.

PIMENTEL M., CLIFTON D. A., CLIFTON L. & TARASSENKO L. (2013). Modelling patient time-series data from electronic health records using gaussian processes. In *Advances in Neural Information Processing Systems : Workshop on Machine Learning for Clinical Data Analysis*, p. 1–4.

SCHÖLKOPF B. & SMOLA A. J. (2002). *Learning with kernels : support vector machines, regularization, optimization, and beyond*. MIT press.

SIMON N., FRIEDMAN J., HASTIE T. & TIBSHIRANI R. (2011). Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of statistical software*, **39**(5), 1.

YEGNANARAYANA B. (2009). *Artificial neural networks*. PHI Learning Pvt. Ltd.

ZOU H. & HASTIE T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, **67**(2), 301–320.