Rapport du projet

Aimé Cazeel, Khadidiatou AW, Kokou Sossou

3 février 2020

Contents

1 Résumé **2**

2 Introduction **3**

2.1 Etat des lieux . . . 3

2.2 CBSM ou Cognitive Behavioral Stress Management . . . 3

2.3 Expérience . . . 3

3 Objectif **3**

4 Approche du problème **4**

4.1 Participants . . . 4

4.2 Méthodes d'expérimentation . . . 4

4.3 Limitations . . . 5

4.4 Objectifs de l'analyse . . . 5

4.5 Méthodes et critères pour l'analyse . . . 5

5 Premières analyses **8**

5.1 Données manquantes et premiers constats . . . 8

5.2 Valeurs extrêmes . . . 8

6 Imputation des données **8**

6.1 Imputation à T0 . . . 8

6.2 Imputation à T1 . . . 12

6.3 Bootstrap: comparaison de moyennes . . . 16

7 Construction de modèles **16**

7.1 Optimisation via Leave one Out . . . 16

7.2 Optimisation via Bootstrap . . . 18

7.3 Comparaisons . . . 19

8 Conclusion **20**

1 Résumé

Le CBSM ou Cognitive Behavioral Stress Management est un programme de gestion du stress qui mélange des exercices de relaxation, de restructuration cognitive et de dynamique de groupe.Si ce programme étudié principalement aux Etats-Unis s'avère efficace sur des maladies comme le cancer du sein ou le VIH, il n'existe cependant que peu d'études sur l'application du programme CBSM sur des patients atteints de maladies cardio-vasculaires et de son effet sur le stress, l'anxiété et les pensées intrusives.

Ainsi, l'objectif de ce document est d'étudier l'efficacité de ce programme auprès de patients atteints de maladies cardio-vasculaires. Pour cela, les patients ont été séparés en 2 groupes, un groupe témoin et un groupe participant au programme CBSM. Des mesures physiologiques ont étés relevés avant et après l'application du programme.

Une étude ultérieur ayant mis en valeur l'efficacité du programme concernant l'amélioration du stress perçu et de l'anxiété, nous cherchons ici à étudier l'efficacité du programme sur le stress ressenti.

Nous allons en premier lieu chercher à faire une imputation afin de conserver le maximum d'individus. En effet, à cause des codnitons de l'expérience, nous avons pu d'individus sans aucun problème.

Nous avons pu voir que plusieurs méthodes été possible pour l'imputation. La taille de nos données ne fut finalement pas spécialement contraignante et nons avons pu renvoyer des résultats satisfaisant.

Puis nous allons chercher des modèles de prédictions afin de peut-être réaliser une selection des patients rejoignant le programme. Nous pourrons voir qu'avec nos données, il semble difficile d'établir un modèle convenable. Même si nous utilisons différentes méthodes pour la recherche de predicteur efficace, nous avons au final une précision faible dans tout les cas. Il semble alors pour le moment difficile de prédire l'amélioration de l'état de santé d'un individus via de programme en se basant seulement sur des informations physilogique avant expérience.

2 Introduction

2.1 Etat des lieux

Les maladies cardiovasculaires sont un ensemble de troubles affectant le coeur et les vaisseaux sanguins. Première cause de mortalité dans le monde selon l'OMS, elles nécessitent souvent une prise en charge lourde comprenant soutien psychologique et médicaments. Ainsi, il est important de trouver des solutions efficaces pour aider les personnes exposées à ces maladies.

Si l'alimentation et le tabagisme sont des facteurs aggravant connus, de nombreuses études mettent en évidence l'existence d'un lien entre stress et maladies cardio-vasculaires. Ainsi, proposer des solutions agissant sur le stress et ne nécessitant pas une prise en charge lourde ou médicamenteuse peut permettre de soulager les personnes atteintes de problèmes cardiovasculaires.

Or, il existe de nombreuses méthodes pour faciliter la gestion du stress, dont le programme CBSM.

2.2 CBSM ou Cognitive Behavioral Stress Management

Le CBSM est un programme de gestion du stress qui mélange des exercices de relaxation, de restructuration cognitive et de dynamique de groupe.L'objectif est de permettre aux patients d'avoir accès à des connaisances sur eux-même, sur le stress et son impact, et sur les réactions psychologiques qu'il peut susciter. Il est constitué de plusieurs séances en groupe ainsi que d'exercices à réaliser chez soi.

Si ce programme étudié principalement aux Etats-Unis s'avére efficace sur des maladies comme le cancer du sein ou le VIH, il n'existe cependant que peu d'études sur l'application du programme CBSM sur des patients atteints de maladies cardio-vasculaires et de son effet sur le stress, l'anxiété et les pensées intrusives.

2.3 Expérience

Ainsi en 2016 une expérience a été réalisé sur des patients atteints de pathologies cardiaques. Ces patients ont suivi le programme CBSM et ont répondus à des questionnaires afin d'évaluer leurs états psychologiques. De plus des relevés physiologiques ont aussi été réalisés. Les questionnaires doivent permettre d'évaluer le ressenti des patients par rapport au stress, tandis que les relevés physiologique permettent d'évaluer l'impact physique des interventions.

3 Objectif

Ce document fait suite au travail rédigé par Franck D'ALESSANDRO mettant en avant l'efficacité du programme CBSM sur la diminution du stress perçu par les patients ainsi que leur anxiété, ainsi que le travail de Aimé CAZEEL confirmant une partie de ces résultats.

Le but de cette étude est donc d'analyser l'influence du programme sur le stress "physique" des patients (que l'on distinguera du stress perçu).

4 Approche du problème

4.1 Participants

Au départ, l'expérience porte sur 150 participants ayant développés une maladie cardiaque. 50 personnes participent au programme CBSM, 50 personnes participent à des séances la relaxation et 50 personnes sont des individus "contrôle" ne suivant pas de programme particulier (hormis les soins). Ces personnes proviennent de différent lieux dans l'aglomération de Grenoble :

- Service de réadaptation cardiaque de l'hôpital Sud (Echirolles)
- Institut cardio-vascualire du groupe hospitalier mutualiste de Grenoble
- Réseau des pathologies vasculaires GRANTED à Saint-Martin-d'Hères
- Service de cardiologie du CHU La Tronche
- Service de diabétologie du CHU La Tronche

Les patients sont recrutés par les équipes soignantes, les personnes acceptant de participés sont placés aléatoirement dans l'un des 3 groupes.

4.2 Méthodes d'expérimentation

Le programme CBSM est constitué de plusieurs séances (2h par semaine) avec en plus des excercices à réaliser chez soi. Après l'ensemble des séances, les patients sont invités à répondre à des questionnaires mesurant leur perception du stress puis des mesure physiologiques sont prises. Ces mesures sont prises avec un module BIOPAC MP 150 qui va permettre de relever plusieurs variables.

Les mesures et les réponses aux questionnaires sont pris à différents moments :

- T0 : avant le début des séances CBSM
- T1 : à la fin des 10 semaines d'interventions
- T2 : 6 mois après l'intervention

4.2.1 Mesure physiologique : HRV ou Heart Rate Variability

Les recherches en psychophysiologie intègrent de plus en plus d'étude sur la variabilité du rythme cardiaque (HRV). En effet, il existe un lien entre le système nerveux parasympathique (lié à la régulation cardiaque) et de nombreux phénomènes psychophysiologique. Le HRV est d'ailleurs utilisé pour prédire les risques de mortalité provenant de cause mental ou physique.

Un relevé du HRV est simple à mettre en place et sans douleur, d'où son utilisation répandue. Parmis les nombreuses variables étudiables, celles d'intérêts sont :

- RMSSD (Root Mean Square of Succesive differences) dont les variations sont dépendantes du tonus vagal (activité du nerf vague, composant du sytème parasympathique contrôlant les activités involontaires des organes).
- HF (High Frenquencies) dont les variations proviennent aussi du tonus vagal mais peuvent être influencé par la respiration.
- LF (Low Frequencies) ainsi que le rapport LF/HF, dont les variations dépendent de divers éléments dont le système sympathique (responsable du rythme cardiaque mais aussi de la contraction des muscles lisses) et le tonus vagal.

Bien que facile à relever, le HRV est sujet à des erreurs de mesures ou à des modifications de celui-ci dû à des facteurs externes pouvant le rendre difficile à étudier (caféine etc. . .)

Dans les études statistiques, le HRV est très souvent utlisé comme une variable les régressions ou les corrélations, permettant souvent de distinguer des groupes selons d'autres critères (comme des différences individuelles). Parfois, le HRV peut être considéré comme une variable dépendante en créant 2 groupes séparés par la médiane. A ce moment, on suppose que le HRV illustre des particularité individuelles (on sait par exemple que le controle vagal est partiellement héritable, ce qui peut en faire une information propre à chaque individus et non dépendantes de variables externes).

Concernant la distribution des variables liées au HRV, la question de la normalité des variables est discutée. Mais des études tendent à observer une non normalité de la distribution de ces variables. La transformation logarithmique est alors une procédure courante pour remédier à ce problème.

4.3 Limitations

4.3.1 Erreurs

Plusieurs problèmes apparaissent dans notre méthodologie :

- Si au départ, nous devions avoir 3 groupes, au final, seulement 2 groupes existent effectivement : les groupes CBSM et CONTROLE. Ces deux groupes sont de tailles différentes. De plus, le groupe CONTROLE réalise des exercices de relaxation.
- Des erreurs de mesures peuvent fausser nos résultats (erreurs de manipulation).De plus, nous faisons face à des individus sous médication ayant une pathologie cardiaque, les chances d'obtenir des valeurs aberantes sont grandes.

4.3.2 Durées de l'expérimentation

Les individus de l'expérience ont été exposé au programme pendant 10 semaines, sans obligations d'être présent à toute les séances, ni obligation à réaliser les exercices à faire chez soi, cela limite donc l'influence du programme sur nos patients.

Enfin, cette étude est basée sur le bénévolat. Hormis la volonté des patients, il n'y avait que peu d'obligations de poursuivre l'étude. Ainsi, nous observons une très grande absence de réponse pour les temps T2 (plusieurs mois après expérience). Nous avons aussi des patients absents lors des premières mesures, mais présent après etc.

Au final, au vu du faible de nombre de réponse pour le temps T2, nous avons décider de limiter nos analyses à T0 et T1, ne nous permettant pas de constater des résultats sur le long terme.

4.4 Objectifs de l'analyse

Ainsi, nos objectifs sont donc les suivants :

- Corriger les valeurs aberantes de certains patients
- Déterminer un modèle de prédiction de l'amélioration de l'état d'un individu.

4.5 Méthodes et critères pour l'analyse

4.5.1 Imputation

L'imputation est le processus de remplacement des données manquantes par des valeurs substituées. Plusieurs techniques d'imputations existent.

- la moyenne(mean): remplace l'ensemble des valeurs manquantes d'une variable par la moyenne de cette variable.
- Predictive mean matching(pmm): Le principe est simplement de chercher l'individus complet le plus proche de l'individus à qui il manque une valeur puis de remplacer la valeur manquante par celle de l'individus proche. C'est une méthode assez populaire et qui se présente comme assez robuste.
- Bayesian Linear Regression(norm): approche de la régression linéaire dans laquelle l'analyse statistique est entreprise dans le contexte de l'inférence bayésienne.
- random forest(rf): utilise le random forest pour imputer les valeurs manquantes

Afin de pouvoir évaluer une méthode, on aura tendance par la suite à minimiser certains critères comme :

- le RMSE (Root Mean Squared Error) ou l'erreur quadratique moyenne, il s'agit simplement de l'espérance de laa racine carré des carrés de la différence entre estimation et valeure réelle.
- le Mean absolute error qui est simplement la valeur absolue de la différence entre prediction et réalité. On peut aussi voir cette mesure selon le pourcentage d'erreur et on parlera alors de MAPE (Mean Absolute Pourcentage Error).

On observera aussi les statistiques descriptives avant et après les imputations pour vérifier que nous n'avons pas trop altérer nos variables.

4.5.2 Modélisation

Par la suite, nous cherchons à créer un modèle pour déterminer si une personne pratiquant le programme CBSM pourra observer une amélioration physique. Cela pourra permettre de trier à l'avance les personnes. Voici une liste non exhaustive des méthodes utilisées.

4.5.2.1 Regression Logistique

La régression logistique est un modèle de régression binomiale. Il s'agit d'un cas particulier du modèle de régression linéaire et est souvent utilisé en apprentissage automatisé. Elle se base sur le LOGIT et est estimé via le maximum de vraissemblance.

4.5.2.2 SVM

Les séparateurs à vaste marge sont des techniques d'apprentissage supervisé destinées à résoudre des problèmes de discimination et de regression. Le principe est de chercher une frontière entre nos groups qui maximisent la marge, c'est à dire la distance entre la frontière et les éléments les plus proches.

4.5.2.3 Random Forest

Les forêts d'arbres de décisions sont des méthodes d'apprentissage basées sur les concepts de bagging (ou bootstrap) et de sous-espaces aléatoire. Le principe est d'effectuer un apprentissage d'arbre de décision sur des sous-ensembles de données.

L'arbre de décision est une méthode d'apprentissage qui cherche à découper nos données en sous-ensemble par des critères maxisimisant (selon les algorithmes) la différence entre les sous-ensembles. On sépare nos données selon une variable puis l'on réintère ce processus pour chaque sous-groupe.

Ces méthodes sont assez courantes parmis les méthodes d'apprentissage supervisé.

4.5.2.4 Leave-One-Out

La validation croisée est une méthode d'estimation de la fiabilité basée sur une technique d'échantillonage. Le principe est de découper nos données d'entrainements en plusieurs partition. A chaque itération, on utilisera une partition comme test et les autres comme données pour entrainer le modèle.

Le leave-one-out est un cas particulier de la validation croisée où l'on considère chaque individus comme une partition. Pour des jeux de données restreints, il permet de réaliser une validation croisée sans avoir un problème lors de l'entrainement (où l'on risquerai d'avoir très peu de données en réduisant plus encore la taille de l'échantillon d'apprentissage).

4.5.2.5 Bootstrap

Avec le peu de données que nous risquons d'avoir au final, nous avons besoin d'une méthode nous assurant de contrôler la "robustesse" de nos méthodes. Le bootstrap est une méthode qui se base sur le reéchantillonage avec remise. Il s'agit de considérer nos données comme la distribution de la population (pour nous, les personnes atteintes de maladie cardiaque) et donc de tirer plusieurs échantillons au sein de cette distribution.

Dans notre cas, le bootstrap est un moyen de contourner notre manque de données. Néanmoins, cette méthode dépend beaucoup de nos données de départ pour être efficace. Il faudra utiliser celle-ci avec parcimonie.

5 Premières analyses

5.1 Données manquantes et premiers constats

Le premier problème dans notre jeu de données est la présence de données manquantes. Certaines personnes n'ont des données que concernant le temps T0 ou T1.

En effet, sur les 105 individus du jeu de données de départ, 23 n'ont pas de valeurs à T0, 48 n'ont pas de données à T1 et 57 n'ont pas de données à T0 et T1.

Si nous pouvons par la suite discuter de l'intérêt de conserver des individus à qui il manque des données à T0 ou T1, il semble relativement logique de ce débarasser des individus sans données à T0 et T1 (il s'agit des personnes présentes seulement au temps T2), qui seront sans intérêt dans notre cas.

Cela nous laisse alors avec 90, dont 9 sans données à T0 et 33 sans données à T1.

5.2 Valeurs extrêmes

Le second problème est la présence de valeurs abérantes.

En effet, sur les 81 individus du jeu de données de départ avec des valeurs à T0, 15 ont au moins une valeur extrême, et sur les 57 individus avec des valeurs à T1, 6 ont au moins une valeur extrême.

6 Imputation des données

Notre but premier est de remplacer les valeurs extrêmes chez nos individus afin de conserver un maximum d'individus pour la suite de l'étude. Pour cela, nous traitons nos données en 2 parties : nous allons premièrement imputer nos données à T0 avec l'ensemble de individus ayant des données à T0 puis nous allons imputer en T1 de la même façon.

6.1 Imputation à T0

A T0, nous avons environ 81 observations et 10 variables avec 35 valeurs manquantes. Vu la taille de nos données, nous ne pouvons pas nous permettre de supprimer ces valeurs manquantes, donc nous allons procéder à une imputation.

6.1.1 Répartition des NA par variables

Regardons le taux de valeurs abérantes pour chaque varialbe à T0:

	% NA
T0_Mean_RR_ms	2.47
T0_STD_RR_ms	7.41
T0_Mean_HR_1_min	4.94
T0_STD_HR_1_min	4.94
T0_RMSSD_ms	7.41
T0_VLF_ms2	1.23
T0_LF_ms2	2.47
T0_HF_ms2	4.94
T0_Total_power_ms2	4.94
T0_Frequence_Respiratoire	2.47

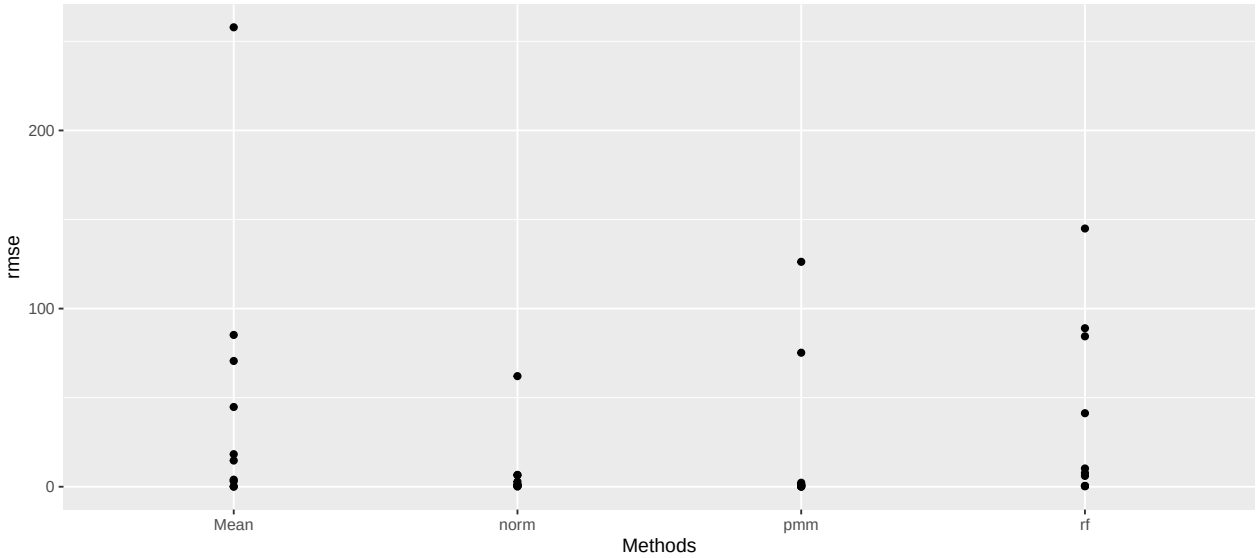
On observe que de manière générale, les variables n'ont pas un taux trop important de valeurs manquantes sauf deux variables (RMSSD et STD).

Nous avons plusieurs choix possibles de méthode pour l'imputation. Nous allons donc devoir comparer l'efficacité de ces méthodes sur nos données.

6.1.2 Comparaison des méthodes

Pour les comparer, nous allons donc volontairement retirer des valeurs parmis les données connus. Considérons notre tableau de données sans les NA, puis créons artificiellement et aléatoirement 5% de valeurs manquantes.

	mean	pmm	rf	norm	BestRmse
T0_Mean_RR_ms	18.26	0.50	14.70	7.68	0.50
T0_STD_RR_ms	2.24	0.63	1.02	2.82	0.63
T0_Mean_HR_1_min	0.28	0.10	0.35	0.07	0.07
T0_STD_HR_1_min	0.19	0.13	0.21	0.30	0.13
T0_RMSSD_ms	3.96	6.08	3.09	10.25	3.09
T0_VLF_ms2	75.23	6.48	126.26	0.44	0.44
T0_LF_ms2	88.97	44.76	144.96	85.22	44.76
T0_HF_ms2	62.08	0.07	6.66	0.07	0.07
T0_Total_power_ms2	257.90	41.27	70.60	84.43	41.27
T0_Frequence_Respiratoire	0.88	1.26	1.27	1.13	0.88



En faisant une comparaison des quatre méthodologies d'imputations, nous remarquons que les méthodes **norm** et **pmm** minimisent au mieux les variables. Mais pour en choisir qu'une à la fin nous allons vérifier les statistiques descriptives des données avant et après imputation et surtout pour la variable RMSSD en phase T1 qui servira plus tard pour la prédiction.

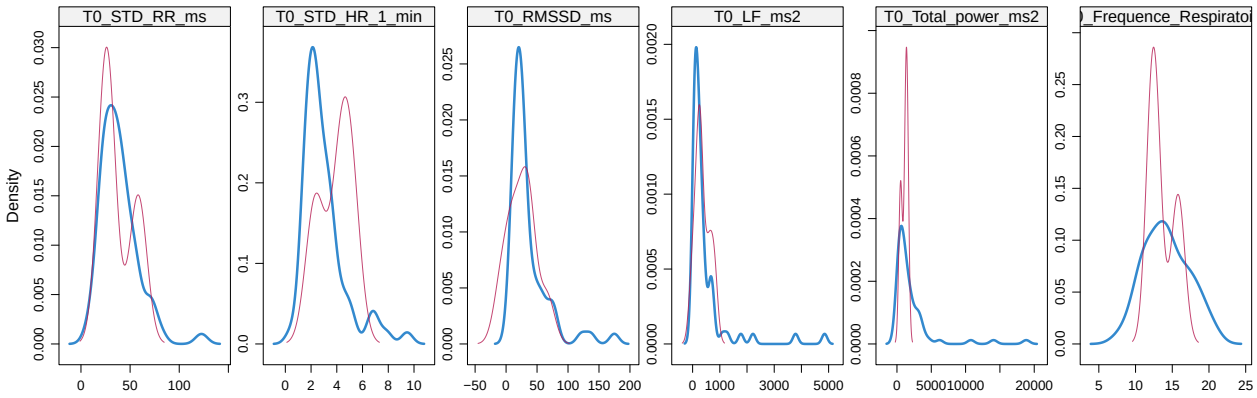
6.1.3 Comparaison des statistiques descriptives des méthodes et de quelques variables pour les méthodes norm et pmm

6.1.3.1 Données originales:

	T0_Mean_RR_ms	T0_STD_RR_ms	T0_Mean_HR_1_min	T0_STD_HR_1_min	T0_RMSSD_ms
	Min. : 654.9	Min. : 7.107	Min. :50.01	Min. :0.4276	Min. : 4.853
	1st Qu.: 827.8	1st Qu.: 25.279	1st Qu.:60.42	1st Qu.:1.8964	1st Qu.: 17.550
	Median : 908.1	Median : 35.295	Median :66.84	Median :2.4959	Median : 26.012
	Mean : 914.6	Mean : 38.304	Mean :66.93	Mean :2.9802	Mean : 34.630
	3rd Qu.: 997.4	3rd Qu.: 45.589	3rd Qu.:72.59	3rd Qu.:3.4310	3rd Qu.: 40.421
	Max. :1200.8	Max. :122.626	Max. :91.79	Max. :9.4619	Max. :174.752

6.1.3.2 Données imputées par norm

	T0_Mean_RR_ms	T0_STD_RR_ms	T0_Mean_HR_1_min	T0_STD_HR_1_min	T0_RMSSD_ms
	Min. : 654.9	Min. : 7.107	Min. :50.01	Min. :0.4276	Min. : -9.546
	1st Qu.: 827.8	1st Qu.: 25.387	1st Qu.:60.42	1st Qu.:1.9783	1st Qu.: 16.861
	Median : 912.4	Median : 35.308	Median :66.85	Median :2.4959	Median : 24.875
	Mean : 914.9	Mean : 38.643	Mean :66.93	Mean :3.0289	Mean : 34.136
	3rd Qu.: 997.4	3rd Qu.: 47.137	3rd Qu.:72.59	3rd Qu.:3.4342	3rd Qu.: 38.238
	Max. :1200.8	Max. :122.626	Max. :91.79	Max. :9.4619	Max. :174.752

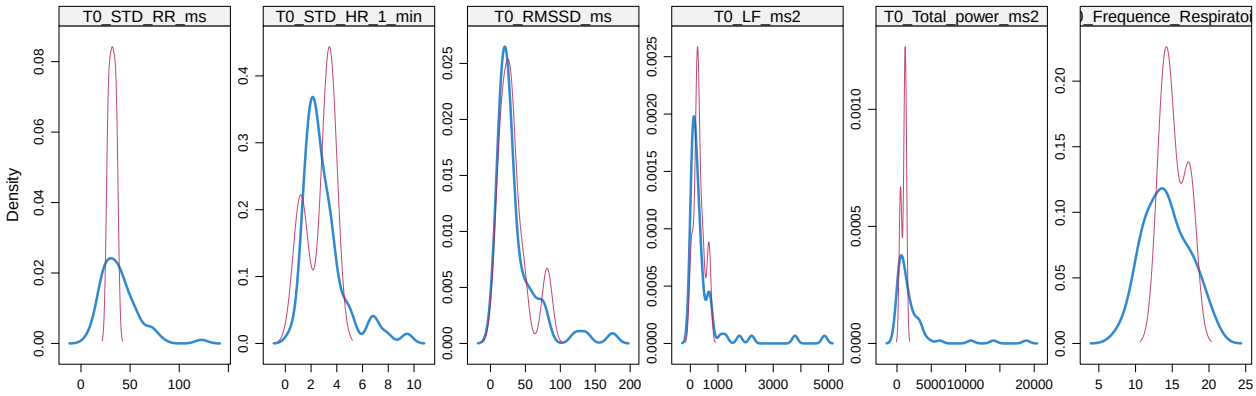


En bleu la densité réelle, en rouge la densité après imputation.

Nous pouvons observer que la méthode norm renvoie une répartition similaire à la répartition d'origine pour la variable LF et potentiellement STD. Pour les autres variables ce n'est pas le cas, même si on observe bien des moyennes très proche.

6.1.3.3 Données imputées par pmm

	T0_Mean_RR_ms	T0_STD_RR_ms	T0_Mean_HR_1_min	T0_STD_HR_1_min	T0_RMSSD_ms
	Min. : 654.9	Min. : 7.107	Min. :50.01	Min. :0.4276	Min. : 4.853
	1st Qu.: 827.8	1st Qu.: 25.524	1st Qu.:60.42	1st Qu.:1.8964	1st Qu.: 17.084
	Median : 908.1	Median : 35.308	Median :66.57	Median :2.4959	Median : 25.737
	Mean : 914.5	Mean : 38.420	Mean :66.92	Mean :2.9739	Mean : 35.022
	3rd Qu.: 997.4	3rd Qu.: 45.589	3rd Qu.:72.59	3rd Qu.:3.4251	3rd Qu.: 42.498
	Max. :1200.8	Max. :122.626	Max. :91.79	Max. :9.4619	Max. :174.752



En bleu la densité réelle, en rouge la densité après imputation.

Nous pouvons observer que la méthode renvoie une bonne répartition pour les variable RMSSD et LF mais ce n'est pas autant le cas avec les autres variables.

En regardant les statistiques descriptives, nous remarquons que les valeurs de la méthode pmm se rapproche plus des vraies valeurs. D'autant plus que la méthode Norm nous a donné des valeurs négatives. Nous allons donc utiliser le pmm pour notre imputation.

6.2 Imputation à T1

T1_Mean_RR_ms	T1_STD_RR_ms	T1_Mean_HR_1_min	T1_STD_HR_1_min
976.8550	50.7064	61.5833	3.1299
973.7792	62.8008	61.9898	6.3224
1238.2258	18.7991	48.4687	0.8212
1016.0228	35.9671	59.1288	2.1256
895.1647	31.4268	67.1114	2.4192
795.9072	22.4034	75.4457	2.1376

A T1, nous avons environ 57 observations et 10 variables avec 22 valeurs manquantes.

6.2.1 Répartition des NA par variables

Regardons le taux de valeurs manquantes pour chaque variable à T1:

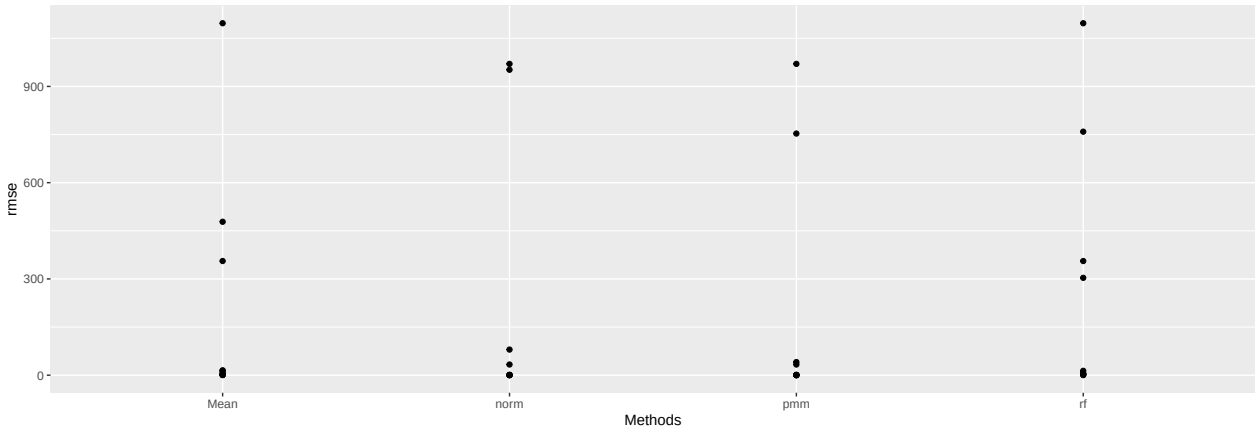
	% NA
T0_Mean_RR_ms	2.47
T0_STD_RR_ms	7.41
T0_Mean_HR_1_min	4.94
T0_STD_HR_1_min	4.94
T0_RMSSD_ms	7.41
T0_VLF_ms2	1.23
T0_LF_ms2	2.47
T0_HF_ms2	4.94
T0_Total_power_ms2	4.94
T0_Frequence_Respiratoire	2.47

Vu la taille de nos données, nous pouvons dire que les taux de valeurs manquantes ne sont pas négligeables.

6.2.2 Comparaison des résultats

Considérons notre tableau de données sans les NA, puis créons artificiellement et aléatoirement 5% de valeurs manquantes. Nous obtenons ces résultats:

	mean	pmm	rf	norm	BestRmse
T1_Mean_RR_ms	13.63	13.63	15.10	5.48	5.48
T1_STD_RR_ms	0.03	0.03	0.83	0.50	0.03
T1_Mean_HR_1_min	0.82	0.82	3.09	0.74	0.74
T1_STD_HR_1_min	0.10	0.10	0.25	0.19	0.10
T1_RMSSD_ms	0.30	0.30	3.27	3.90	0.30
T1_VLF_ms2	33.21	33.21	40.91	79.80	33.21
T1_LF_ms2	355.80	355.80	303.44	7.15	7.15
T1_HF_ms2	970.33	970.33	952.02	753.12	753.12
T1_Total_power_ms2	1096.94	1096.94	478.19	758.99	478.19
T1_Frequence_Respiratoire	1.08	1.08	0.18	0.63	0.18



Nous remarquons que les méthodes **norm** et **pmm** ont les RMSE les plus faibles et donnent donc de meilleurs résultats. Mais pour en choisir qu'une à la fin nous allons voir les statistiques descriptives des données avant et après imputation et surtout pour la variable RMSSD en phase T1 qui servira plus tard pour la prédiction.

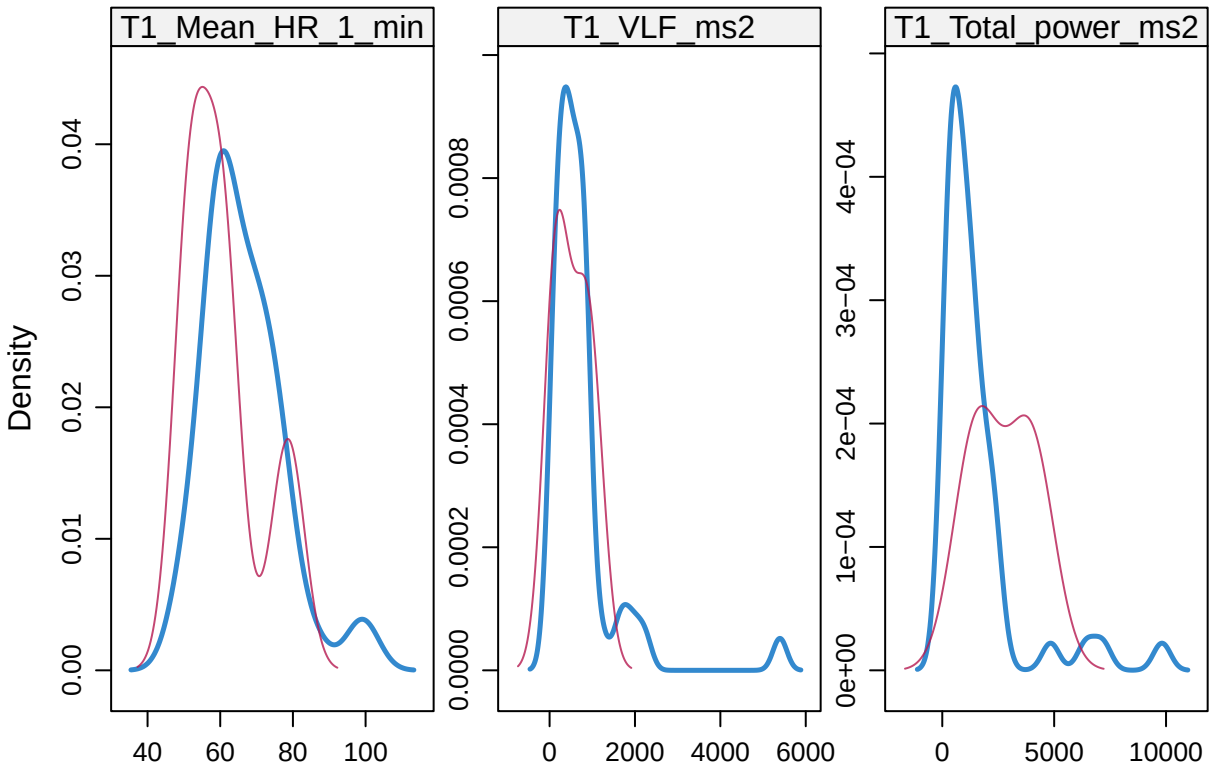
6.2.3 Comparaison des statistiques descriptives de quelques variables pour les méthodes norm et pmm

6.2.3.1 Données originales:

	T1_Mean_RR_ms	T1_STD_RR_ms	T1_Mean_HR_1_min	T1_STD_HR_1_min	T1_RMSSD_ms
	Min. : 598.9	Min. : 7.107	Min. : 48.47	Min. : 0.4276	Min. : 4.618
	1st Qu.: 827.8	1st Qu.: 26.849	1st Qu.: 58.97	1st Qu.: 1.8960	1st Qu.: 14.343
	Median : 952.0	Median : 35.540	Median : 63.10	Median : 2.2972	Median : 25.749
	Mean : 937.1	Mean : 41.142	Mean : 65.84	Mean : 3.2437	Mean : 36.865
	3rd Qu.:1019.0	3rd Qu.: 50.265	3rd Qu.: 73.08	3rd Qu.: 3.2837	3rd Qu.: 36.300
	Max. :1238.2	Max. :120.334	Max. :100.24	Max. :12.6799	Max. :178.357

6.2.3.2 Données imputées par norm

	T1_Mean_RR_ms	T1_STD_RR_ms	T1_Mean_HR_1_min	T1_STD_HR_1_min	T1_RMSSD_ms
	Min. : 598.9	Min. : 7.107	Min. : 48.47	Min. : 0.2095	Min. :-19.08
	1st Qu.: 827.8	1st Qu.: 26.849	1st Qu.: 59.05	1st Qu.: 1.8122	1st Qu.: 14.14
	Median : 952.0	Median : 35.540	Median : 63.10	Median : 2.2972	Median : 25.75
	Mean : 937.4	Mean : 41.212	Mean : 65.90	Mean : 3.2070	Mean : 36.11
	3rd Qu.:1019.0	3rd Qu.: 50.265	3rd Qu.: 73.08	3rd Qu.: 3.2837	3rd Qu.: 36.30
	Max. :1238.2	Max. :120.334	Max. :100.24	Max. :12.6799	Max. :178.36

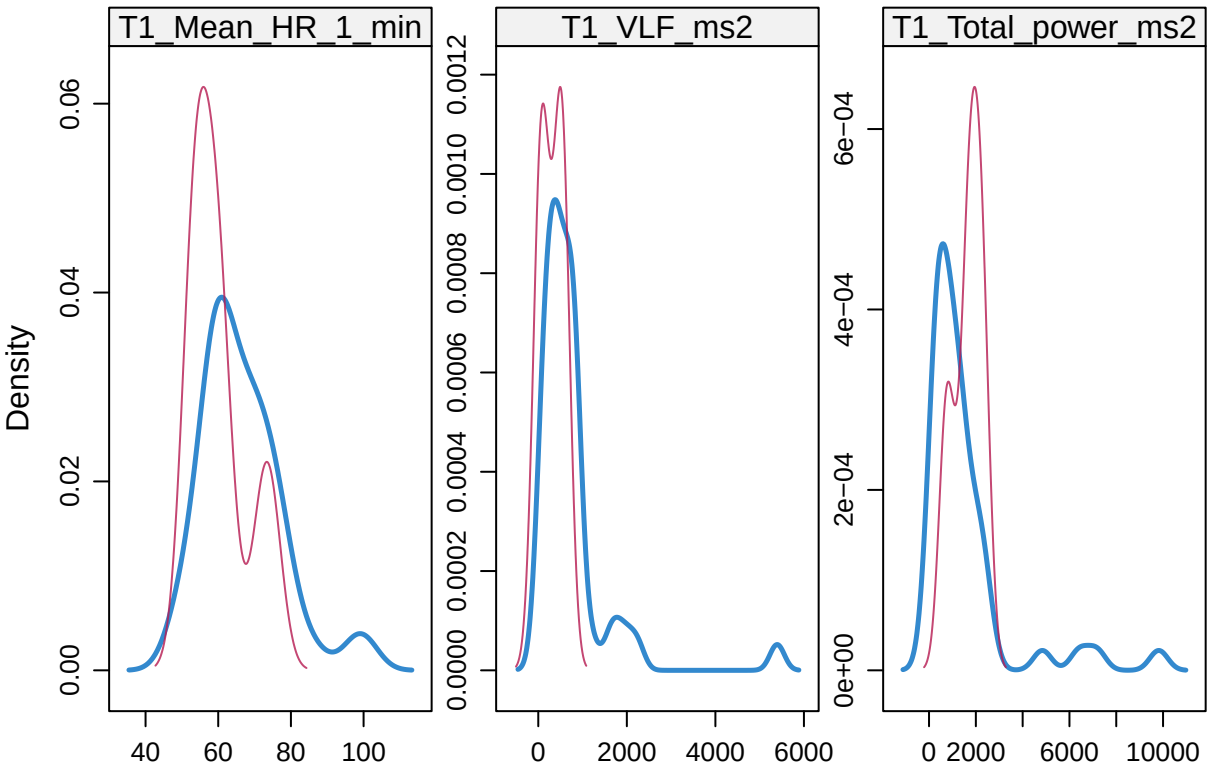


En bleu la densité réelle, en rouge la densité après imputation.

Nous pouvons observer que la méthode norm renvoie une répartition assez similaire à la répartition d'origine pour la variable MEAN HR et potentiellement VLF. Pour les autres variables ce n'est pas le cas, même si on observe bien des moyennes très proche.

6.2.3.3 Données imputées par pmm

	T0_Mean_RR_ms	T0_STD_RR_ms	T0_Mean_HR_1_min	T0_STD_HR_1_min	T0_RMSSD_ms
	Min. : 654.9	Min. : 7.107	Min. :50.01	Min. :0.4276	Min. : 4.853
	1st Qu.: 827.8	1st Qu.: 25.524	1st Qu.:60.42	1st Qu.:1.8964	1st Qu.: 17.084
	Median : 908.1	Median : 35.308	Median :66.57	Median :2.4959	Median : 25.737
	Mean : 914.5	Mean : 38.420	Mean :66.92	Mean :2.9739	Mean : 35.022
	3rd Qu.: 997.4	3rd Qu.: 45.589	3rd Qu.:72.59	3rd Qu.:3.4251	3rd Qu.: 42.498
	Max. :1200.8	Max. :122.626	Max. :91.79	Max. :9.4619	Max. :174.752



En bleu la densité réelle, en rouge la densité après imputation.

Nous pouvons observer que la méthode renvoie une assez bonne répartition pour des variables mais ce n'est pas autant le cas pour la variable MEAN HR.

En regardant les statistiques descriptives, nous remarquons que les valeurs de la méthode pmm se rapproche plus des vraies valeurs. Globalement, les résultats sont très proche de la réalité et même si parfois, l'estimation bayesienne donne de meilleur résultat, nous pouvons observer que la méthode pmm est dans l'ensemble plus efficace.

Conclusion : Notre jeu de données sera imputé par la méthode de la moyenne prévisionnelle qui donne de meilleurs résultats en T0 et T1.

Pour la suite nous allons combiner nos données en T0 et T1 afin de créer nos modèles, nous nous retrouvons donc avec 48 observations et 20 colonnes.

6.2.4 Comparaison avant et après imputation

RMSSD Avant Imputation:

	T0_RMSSD_ms	T1_RMSSD_ms
	Min. : 4.853	Min. : 4.618
	1st Qu.: 20.021	1st Qu.: 14.851
	Median : 28.035	Median : 25.749
	Mean : 44.796	Mean : 39.956
	3rd Qu.: 57.952	3rd Qu.: 44.460
	Max. :214.604	Max. :178.357
	NA's :4	NA's :3

RMSSD Après Imputation:

	T0_RMSSD_ms	T1_RMSSD_ms
	Min. : 4.853	Min. : 4.618
	1st Qu.: 20.021	1st Qu.: 17.094
	Median : 28.035	Median : 28.166
	Mean : 44.678	Mean : 40.341
	3rd Qu.: 57.952	3rd Qu.: 51.035
	Max. :214.604	Max. :178.357

En observant par exemple pour la variable RMSSD avant imputation et après imputation, nous obtenons une moyenne de 44.79 en T0 et 39.95 en T1 contre 44.67 en T0 et 40.34 , les valeurs sont donc assez similaires.

6.3 Bootstrap: comparaison de moyennes

Pour terminer nos analyses, nous souhaitons identifier de nouveau si il y a eu une amélioration général de l'état de santé des patients. Nous allons donc comparer les moyennes de notre échantillon. Au vu du faible de nombre de données, afin de parfaire nos estimations de la moyenne, nous avons utilisé la méthode du Bootstrap. Nous regardons surtout la variable RMSSD qui est un bon indicaeur de cette amélioration. Nous avons donc tiré 200 échantillons, et pour chacun d'eux, avons calculé les moyennes en T0 et T1 et appliqué un Test de Wilcoxon sur ces moyennes obtenues. Nous avons donc comme hypothèses:

- H0: T0>T1
- H1: T0<T1

M0_boot	M1_boot	pval_boot
44.40285	39.95523	0.2798285

Nous avons une p-value= 0.2798285 qui est supérieure à 0.05, donc on conserve H0 au seuil de 5%; il y a bien une amélioration.

7 Construction de modèles

Maintenant que nous avons imputé nos variables, nous pouvons chercher à créer un modèle pour prédire si une personne a vu sa condition physique s'améliorer. Pour cela, nous n'utilisons que les variables du temps T0.

Avec moins d'une cinquantaine d'individus dans nos résultats finaux, nous décidons de simplement chercher un modèle avec l'ensemble de nos données. Il faudra alors espérer avoir plus d'individus dans le futur afin de tester nos modèles.

7.1 Optimisation via Leave one Out

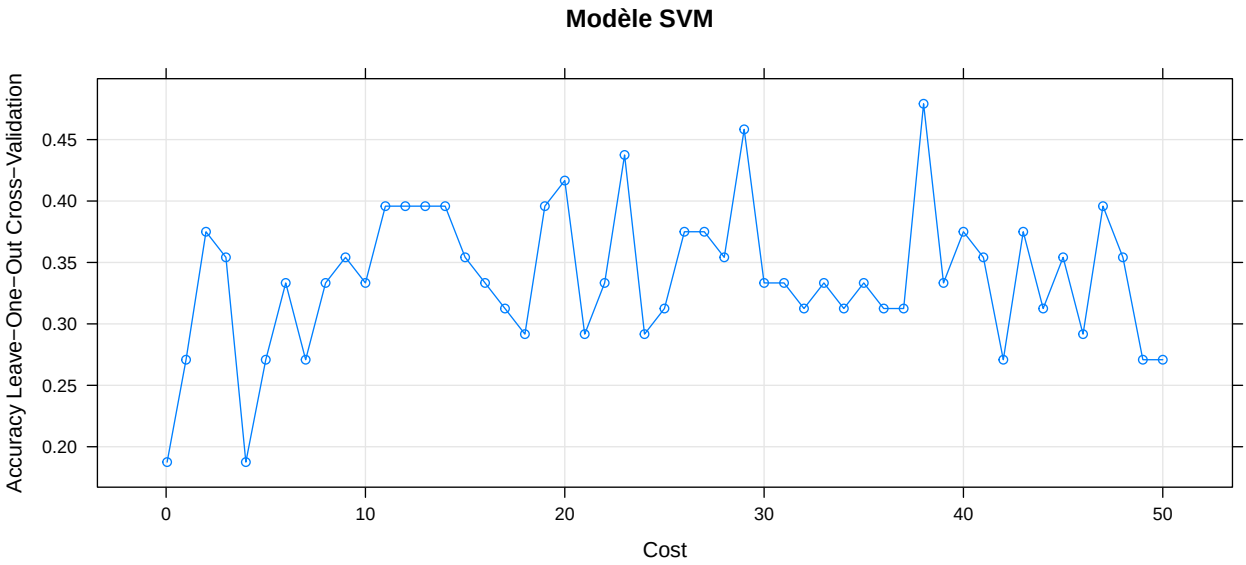
Nous allons chercher ici à optimiser nos modèles via la validation croisée. La méthode de validation croisée Leave one out permet de faire des tests avec peu d'individus.

7.1.1 Regression logistique

La première méthode que nous allons utiliser ici est la regression logistique descendante.

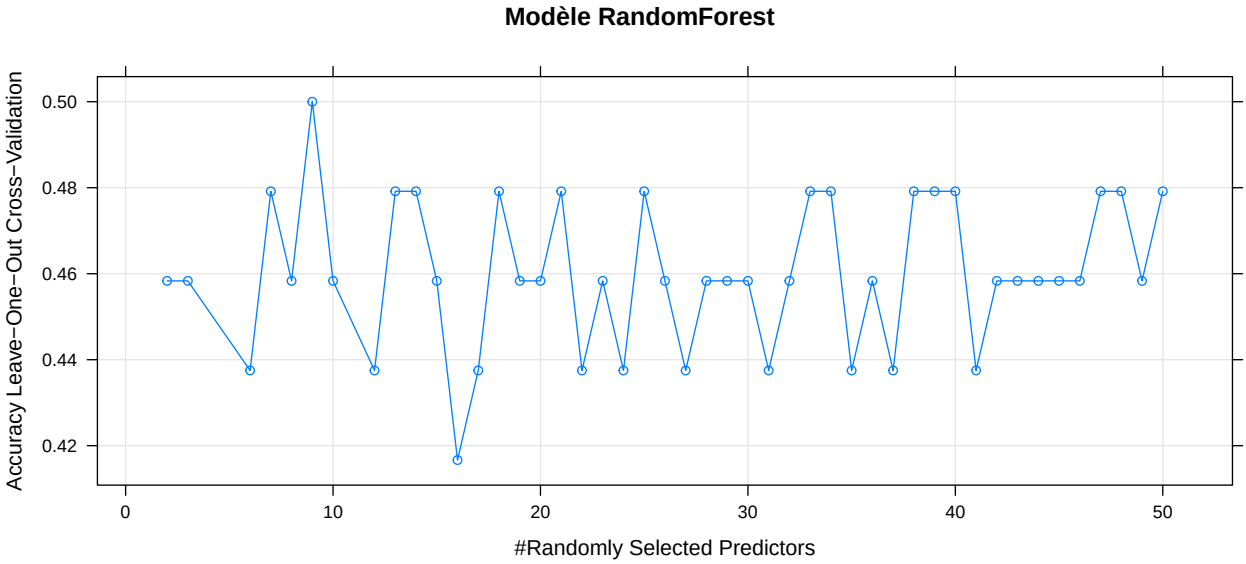
La précision de notre modèle est d'environ 0.5208333, ce qui est légèrement meilleur qu'un modèle renvoyant toujours la même prédiction avec nos données. Les variables conservée sont les variables RMSSD, VLF, HF, Mean HR, STD RR. Les variables conservées sont principalement celles que l'on considère souvent comme assez importantes dans la litterature scientifique. Parmis ce type de variable, seul LF a été écartée.

7.1.2 SVM



Le SVM ici a dans l'ensemble un accuracy proche ou inférieur à 50%. Vu nos données, cela signifie qu'il est moins bon qu'un predicteur prédisant toujours la même valeur.

7.1.3 randomForest



Optimiser la méthode de forêt aléatoire avec la validation croisée "Leave One Out" n'est pas évidente. En effet, l'évaluation via un seul individu ne permet pas d'obtenir de résultat robuste dans ce cas là car le résultat dépend trop des variables choisies. De plus, l'accuracy, qu'importe le nombre de variable, semble se stabiliser vers 50% ou moins, ce qui est actuellement moins efficace qu'un predicteur prédisant toujours la même variable dans notre cas.

7.2 Optimisation via Bootstrap

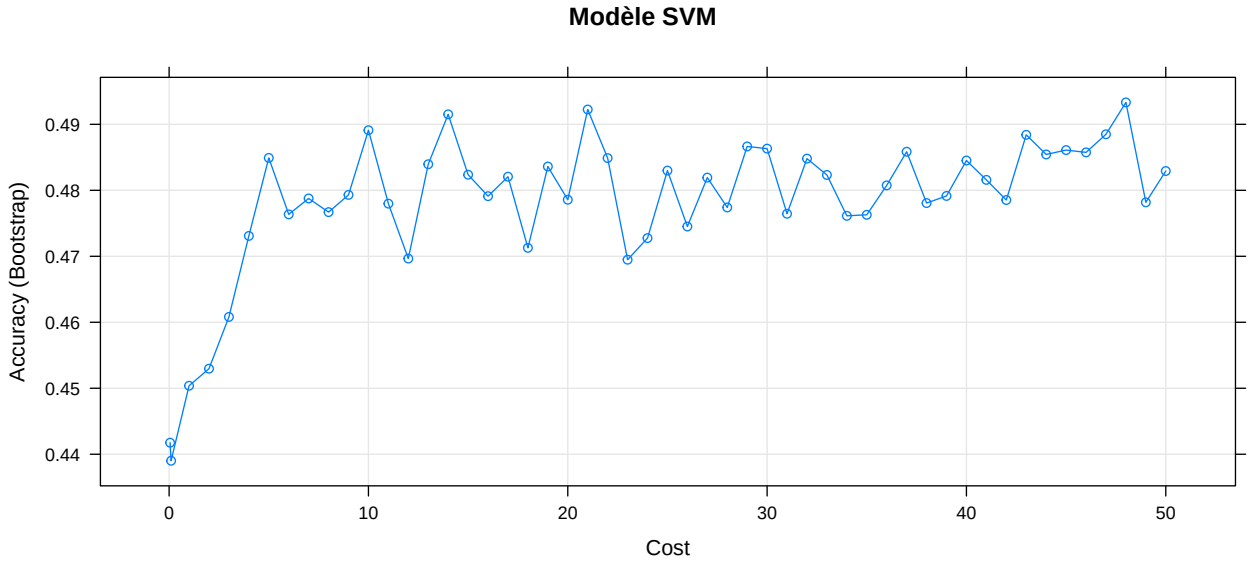
Ici, nous allons utiliser une autre méthode, le bootstrap pour essayer de contourner notre problème de manque de donnée.

7.2.1 Regression logistique

La précision de notre modèle est d'environ 0.5176385, ce qui est légèrement meilleur qu'un modèle renvoyant toujours la même prédiction avec nos données. Les variables conservée sont les variables RMSSD, VLF, HF, Mean HR, STD RR. Les variables conservées sont principalement celles que l'on considère souvent comme assez importantes dans la litterature scientifique. Parmis ce type de variable, seul LF a été écartée.

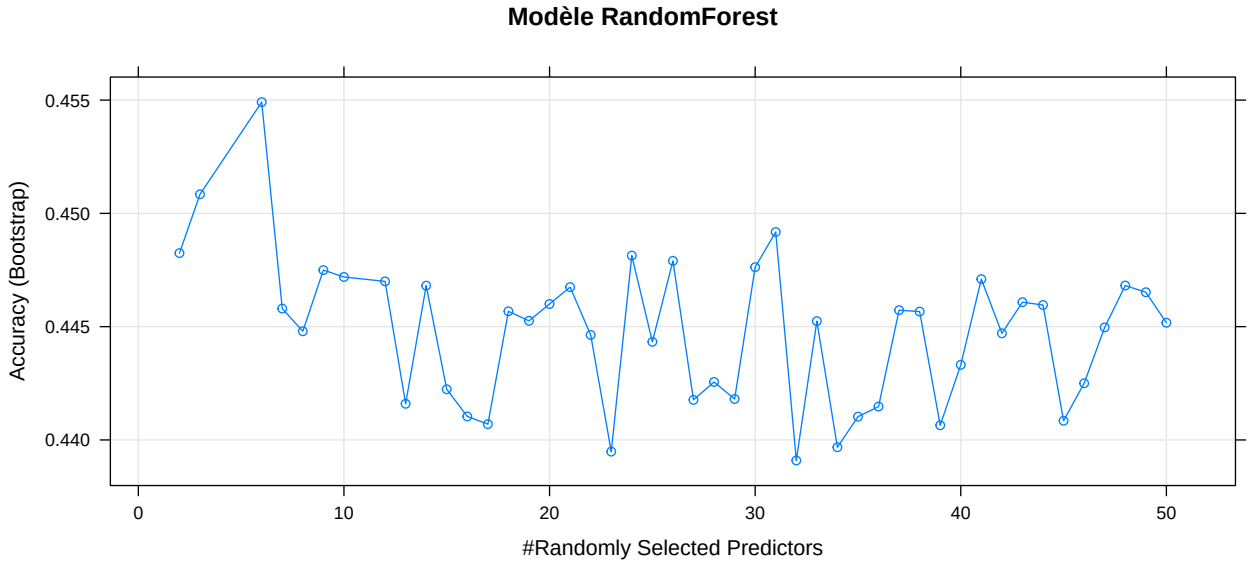
Au final, on retrouve pratiquement le même résultat qu'avec le leave one out.

7.2.2 SVM



Le SVM ici a dans l'ensemble un accuracy proche ou inférieur à 50%. Vu nos données, cela signifie qu'il est moins bon qu'un predicteur prédisant toujours la même valeur.

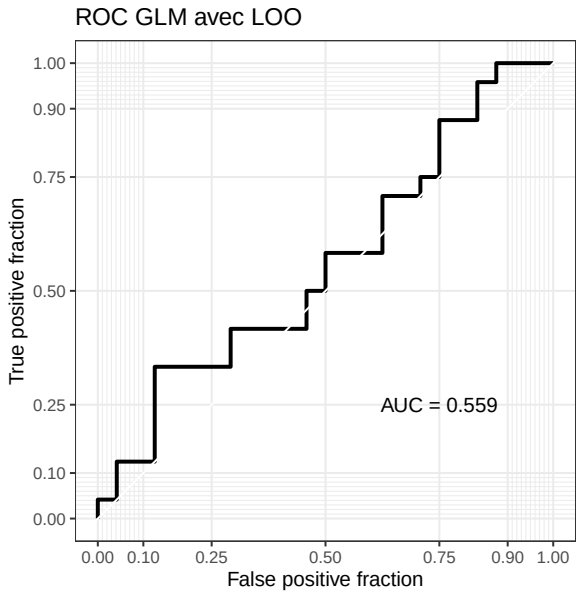
7.2.3 randomForest

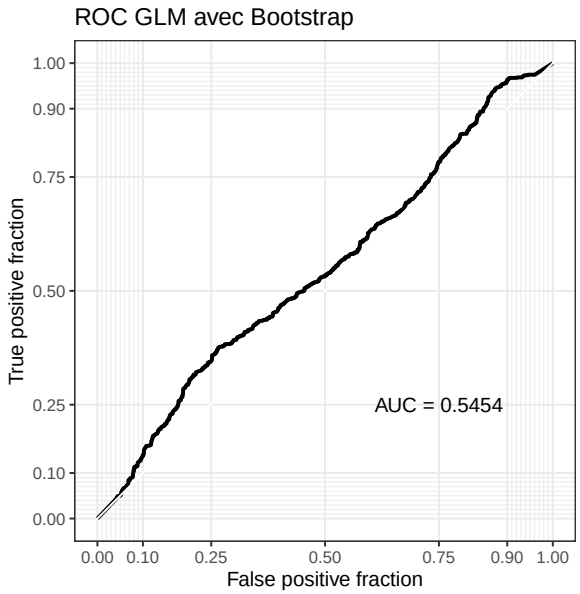


L'accuracy, qu'importe le nombre de variable, semble se stabiliser vers 50% ou moins, ce qui est actuellement moins efficace qu'un predicteur prédisant toujours la même variable dans notre cas.

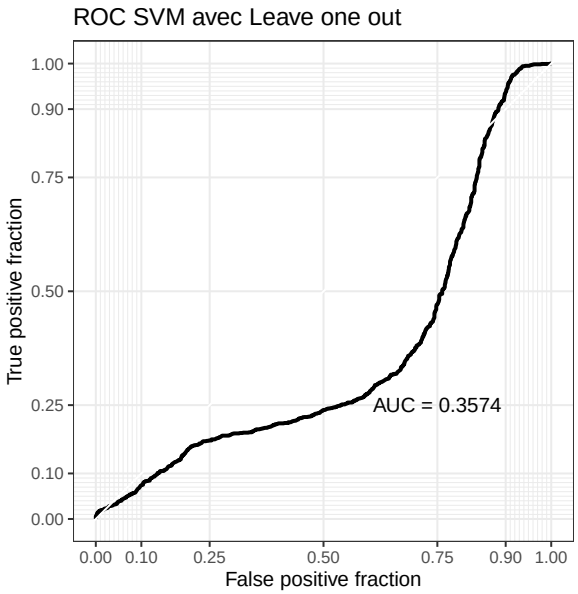
7.3 Comparaisons

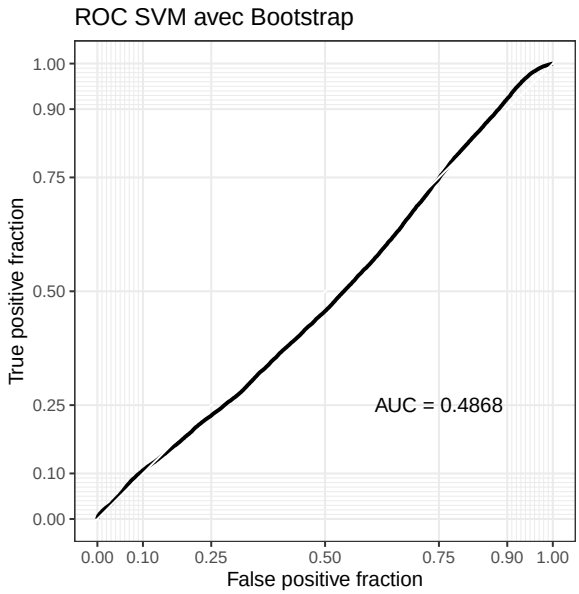
Nous avons vu que quelques soit les méthodes d'évaluation et de prédiction que nous utilisons, dans l'ensemble, nous obtenons souvent des résultats semblables, avec une précision proche de 50%. Nous allons maintenant terminer en comparant nos méthodes selon d'autres critères afin d'arrêter notre choix. Pour cela, nous regardons les courbes ROC.



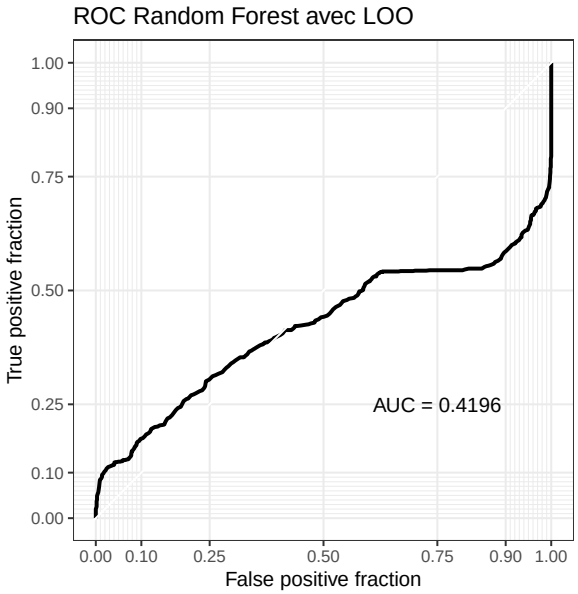


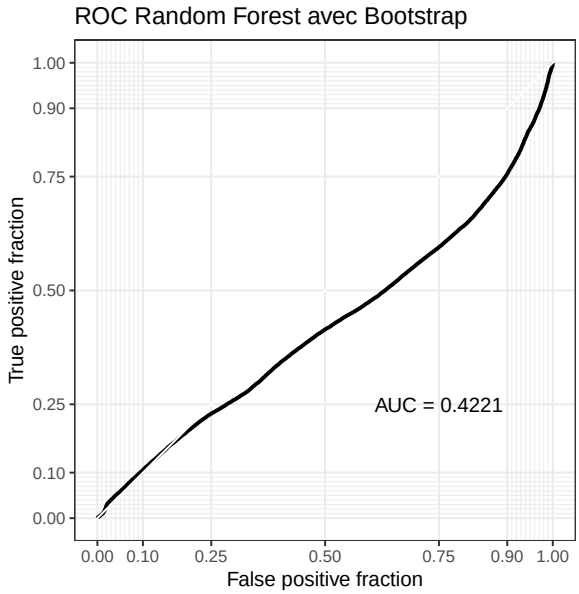
Dans l'ensemble, on peut supposer que l'AUC de la regression logistique est d'environ 0.55 .





Dans l'ensemble, on peut supposer que l'AUC du SVM est d'environ 0.48 . C'est moins qu'avec la regression logistique.





Dans l'ensemble, on peut supposer que l'AUC du random Forest est d'environ 0.43, c'est encore inférieur au SVM et à la regression logistique.

8 Conclusion

Concernant l'imputation, nous avons pu voir que plusieurs méthodes etaient possibles. La taille de nos données ne fut pas spécialement contraignante et nons avons pu renvoyer des résultats satisfaisants. En sachant que les relevés physiologiques sont souvent soumis à des erreurs quand ils concernent des personnes après des problèmes cardiaques, nous pouvons aisement imaginer que l'imputation permettra de compenser cela dans des expériences futures.

Concernant les modèles de prédictions, nous pouvons voir qu'avec nos données, il semble difficile d'établir un modèle convenable. Même si nous utilisons différentes méthodes pour la recherche de predicteur efficace,

nous avons au final une précision faible dans tout les cas. Il semble alors pour le moment difficile de prédire l'amélioration de l'état de santé d'un individus via le programme en se basant seulement sur des informations physilogiques avant expérience.

Enfin, pour conclure, il est à noter qu'il est possible que la méthode du RandomForest puisse renvoyer de meilleures prédictions. En effet, il suffirait d'inverser le modèle et nous aurions alors un modèle de prédiction avec une précision plus grande. Ainsi, on peut imaginer qu'avec beaucoup plus de données et surement d'itération, il est possible de voir apparaître un modèle acceptable.