Apport de la fouille de données pour la prévention du risque suicidaire

Romain Billot*, Sofian Berrouiguet*,**, Mark Larsen***,
Michel Walter**, Jorge López Castroman****, Enrique Baca-García♯,
Philippe Courtet♯♯, Philippe Lenca*

*IMT Atlantique, Lab-STICC, UBL, F-29238 Brest, France
prenom.nom@imt-atlantique.fr
**CHRU de Brest, Pyschiatrie Adulte à Bohars, 29200 Brest, France
sofian.berrouiguet@gmail.com
***Black Dog Institute, University of New South Wales, Sydney, Australie
****CHRU de Nîmes, 30029 Nîmes, France
♯Hospital Universitario Fundacion Jimenez Diaz, Madrid, Espagne
♯♯CHRU de Montpellier, INSERM U1061, Montpellier, France

Résumé. Avec plus de 800 000 décès par an dans le monde, le suicide est la troisième cause de décès évitable. Il y a 20 fois plus de tentatives, impliquant de nombreuses hospitalisations, des coûts humains et sociétaux énormes. Ces dernières années, les modalités de collecte de données, sociologiques et cliniques, concernant les patients reçus en consultation après une tentative, ont connu de profonds changements liés aux outils numériques. Nous présentons les principaux résultats d'un processus complet de fouille de données sur un échantillon de suicidants de deux hôpitaux européens. Le premier objectif est d'identifier des groupes de patients similaires et le second d'identifier des facteurs de risque associés au nombre de tentatives. Des méthodes non supervisées (ACM et clustering) et supervisées (arbres de régression) sont appliquées pour y répondre. Les résultats mettent en lumière l'apport de la fouille de données à des fins descriptives ou explicatives.

1 Le suicide : un fléau de santé publique

On estime qu'il y a plus de $800\,000$ décès par suicide, par an dans le monde (WHO, 2014). Avec environ $10\,000$ décès en France par an, le suicide est la troisième cause de décès évitable. Le coût humain est donc colossal et bien souvent sous-estimé à cause des suicides non repérés. Ainsi, en France métropolitaine on considère une sous-estimation des décès de $9,4\%$. Le chiffre officiel de $9\,715$ décès pour l'année 2012 (première année où le nombre de décès est inférieur à $10\,000$) peut ainsi être ré-évalué à $10\,690$ (ONS, 2016). Par ailleurs, il est estimé qu'il y a environ vingt tentatives de suicide pour un décès par suicide, ce qui représente environ $200\,000$ tentatives en France par an. Les proches subissent souvent des conséquences sévères (Fauré, 2008) augmentant ainsi le coût humain. Le suicide est donc un problème de santé majeur dans toutes les sociétés avec des conséquences également financières

importantes (Smith, 2011). Malgré cela, comparativement à d'autres maladies, les efforts gouvernementaux, financiers et humains pour prévenir les maladies psychiatriques et le suicide en particulier sont assez récents et bien souvent jugés insuffisants (Montaigne, 2014; Lytle et al., 2016). L'Observatoire national du suicide, par exemple, n'a été créé en France qu'en 2013.

Si une première tentative représente un très grand facteur de risque (Finkelstein et al., 2015), d'autres critères doivent être pris en compte dans le processus de prise de décision thérapeutique et de prévention. Les facteurs de risque suicidaire ont déjà été étudiés parmi des populations de suicidants, qui sont une cible privilégiée pour l'élaboration de stratégies de prévention et d'intervention. Nous présentons un travail de fouille de données, réalisé avec des psychiatres, mené sur une population de suicidants. La section 2 recense quelques travaux, orientés données, menés pour analyser ce fléau et développer la prévention. Nous présentons le protocole clinique, les données, les méthodes déployées et les principaux résultats de notre étude dans la section 3. Enfin nous concluons dans la section 4.

2 Données et prévention du suicide

L'analyse de données sur le suicide n'est pas récente notamment grâce à l'attention portée par des sociologues. Durkheim (1897) a dégagé les causes du suicide et proposé une typologie des suicides, selon leurs causes, puis à l'aide d'une analyse statistique précise, l'auteur a montré que le suicide est un phénomène social normal. La statistique fait apparaître des régularités, déjà observées par Durkheim (1897), dans la fréquence des suicides, certaines évoluant avec les changements de rythmes de la vie sociale (Aveline et al., 1984).

Ces dernières années, les modalités de collecte de données, sociologiques et cliniques, concernant les patients reçus en consultation après une tentative, ont connu de profonds changements liés aux outils numériques. Cette collecte de données est augmentée par le développement et l'utilisation d'applications web et/ou mobile pour le suivi de patients (Berrouiguet et al., 2016a, 2017; Barrigón et al., 2017) ou encore par le suivi des médias sociaux, tels Facebook (Moreno et al., 2011) ou Twitter (Abboute et al., 2014), permettant également de considérer l'évolution de l'état des personnes suivies (Maigrot et al., 2016). Le nombre de variables décrivant les patients explose et dans une moindre mesure le nombre de patients suivis. Ainsi, de nouvelles opportunités apparaissent pour des analyses prenant en compte des dizaines ou encore des centaines de variables sur des échantillons de plus en plus grands pour étudier ce phénomène éminemment complexe. Un fort espoir est placé dans la fouille de données (Berrouiguet et al., 2016b; Ribeiro et al., 2016; Rakesh, 2017). Nous présentons ci-après quelques travaux intéressants et récents sans aucune exhaustivité.

L'identification de groupes de patients similaires a été l'objet de nombreux travaux avec bien souvent pour but d'identifier des facteurs de risque. Wolodzko et Kokoszka (2014a) identifient 7 groupes avec des risques accrus de conduite suicidaire dans un échantillon de 5 977 américains âgés de 15 à 54 ans. Dans une revue de la littérature, les mêmes auteurs distinguent 5 groupes et concluent à la nécessité de développer l'analyse de groupes sur des populations plus larges, représentatives et homogènes (Wolodzko et Kokoszka, 2014b). D'autres auteurs s'intéressent à la caractérisation de 418 patients asiatiques en distinguant les récidivistes des patients ayant réalisé une unique tentative (Choo et al., 2014). Enfin, notons l'étude de Lopez-Castroman et al. (2016) sur 1 009 patients qui identifie 3 groupes selon une échelle d'impulsivité et de fréquence de récidive. Les deux groupes sont classifiés ensuite par arbre de décision.

L'apprentissage supervisé, notamment pour l'évaluation du risque de passage à l'acte ou de récidive, est de plus en plus mis à contribution. Les procédures recommandent que chaque patient reçu dans les services hospitaliers après une tentative non fatale doit passer une évaluation psychologique avant sa sortie afin d'évaluer le risque de récidive (Chan et al., 2016). Il existe différents tests psychologiques. Delgado-Gómez et al. (2011) présentent une comparaison de l'échelle d'impulsivité BIS11 de Barratt version 11 (Patton et al., 1995) et le questionnaire IPDE-SQ de dépistage et d'évaluation du trouble de la personnalité (Loranger et al., 1994) afin d'évaluer leur capacité, une fois combinés avec quatre classifieurs, à distinguer des suicidants (345 individus) de non suicidants (534 individus). Les quatre classifieurs obtiennent de meilleurs résultats que la méthode traditionnelle de classification psychométrique, le SVM étant le meilleur. IPDE-SQ est jugée plus discriminante que BIS11. D'autres études, récentes, démontrent l'intérêt des méthodes supervisées pour l'évaluation du risque suicidaire parmi lesquelles celles de Tran et al. (2014), Glenn et Nock (2014), Kessler et al. (2015), Combes et al. (2016), Karmakar et al. (2016) et Walsh et al. (2017).

La plupart des études pointent qu'il est nécessaire de développer les analyses en augmentant la taille des échantillons, en intégrant de nouvelles variables dont des informations écologiques obtenues le plus objectivement possible dans le milieu naturel du patient, par exemple la qualité du sommeil. Elles mettent également l'accent sur la nécessité d'améliorer la qualité des données. Enfin notons qu'il est difficile d'extrapoler, donc de comparer, sauf sur quelques variables invariantes comme le sexe et l'âge, les résultats d'une population à une autre. Les études menées sont effectivement très dépendantes des populations de patients et des protocoles de collecte de données suivis, donc des données collectées (Lopez-Castroman et al., 2015).

3 Fouille de données pour la prévention du risque suicidaire

Nous présentons dans cette section les principaux résultats d'un processus complet de fouille de données sur un échantillon de suicidants reçus dans deux hôpitaux européens. Deux objectifs principaux discutés en concertation avec les psychiatres sont poursuivis. Le premier objectif est d'identifier des groupes de patients similaires et le second d'identifier des facteurs de risque associés au nombre de tentatives de suicide par patient. Le nombre de TS constitue un fort enjeu de santé publique car de nombreux patients sont des multi-récidivistes et les coûts humains et financiers sont colossaux.

3.1 Recrutement des patients et évaluation clinique

3.1.1 Recrutement des patients

La population est composée de patients âgés de 18 ans ou plus, reçus pour des tentatives de suicide dans deux hôpitaux universitaires, à Madrid (hôpital Ramon y Cajal), Espagne, et Montpellier (hôpital Lapeyronie), France, entre 1994 et 2006.

Les équipes médicales se sont accordées pour employer les mêmes méthodes cliniques et des procédures d'évaluation comparables. Les données sociodémographiques et cliniques des patients ont été fusionnées dans une base de données commune qui intègre également les résultats des questionnaires d'évaluation. La définition d'une tentative de suicide retenue est la suivante : *"a potentially self-injurious behavior with a nonfatal outcome, for which there*

is evidence (either explicit or implicit) that the person intended at some (nonzero) level to kill himself/herself." Les études ont été approuvées par les différents comités d'éthique et conduites selon les principes de la déclaration d'Helsinki, à propos des principes éthiques de la recherche médicale. Tous les participants ont signé un document de consentement après une explication précise des objectifs de l'étude et des procédures. Le jeu de données complet inclut des patients ne présentant pas d'historiques de tentatives de suicide, des donneurs de sang, et autres patients de "contrôle". Nous nous sommes concentrés sur la seule population des suicidants.

3.1.2 Procédure clinique

L'évaluation clinique des patients a été conduite aux urgences. Elle s'appuie sur des entretiens structurés avec notamment l'échelle Columbia (*Columbia Suicide History Form*). Cette échelle consiste en une série de questions permettant d'évaluer l'idéation suicidaire et les antécédents suicidaires chez les patients à risque. Les données ont été collectées grâce à l'application MEmind (Berrouiguet et al., 2015). L'application collecte des données sociodémographiques, de diagnostic, et pharmacologiques au sein du protocole d'évaluation. Les variables sociodémographiques incluent notamment l'âge au moment de l'épisode suicidaire, le genre, la profession, le statut marital, la situation professionnelle, le nombre d'enfants et le niveau d'éducation. Des informations sur les historiques familiaux quant à la question du suicide, l'âge de la première tentative, le caractère violent de l'acte, sont également récoltées. Les versions française et espagnole du questionnaire neuropsychiatrique international MINI (Lecrubier et al., 1997) ont été utilisées afin d'obtenir des diagnostics psychiatriques : troubles de l'humeur (par exemple troubles bipolaires, dépression), anxiété, troubles obsessionnels compulsifs, drogues et alcool, troubles psychotiques, alimentaires, somatiques, etc. Le psychiatre en charge du diagnostic complète à chaque fois les informations à l'aide du dossier médical et potentiellement à l'aide d'informations venant des proches. Le risque suicidaire a été évalué à l'aide de l'échelle SIS (*Suicide Intent Scale*, Beck et al. (1974)), une échelle de risque semi structurée à 15 items qui produit un score global de sévérité quant à l'intention suicidaire. La tentative en tant que telle est évaluée à l'aide de l'échelle RRRS (*Risk-Rescue Rating Scale*) qui est un questionnaire de 10 items qui mesure la sévérité de l'intention et du geste suicidaire au regard de la létalité et de la vraisemblance d'une intervention de sauvetage au moment du geste (Weisman et Worden, 1972). L'impulsivité du patient a été mesurée à l'aide de l'échelle BIS10 (*Barratt Impulsiveness Scale*) une série de 34 questions qui mesurent la propension du sujet à prévoir ses gestes, son comportement, dans diverses situations (Patton et al., 1995).

3.2 Construction de la base de données analysée

Bien que les données sont censées être collectées selon des processus identiques, une variabilité de la qualité est inévitable de par la diversité des équipes de recueil. De plus, le contexte particulier de l'accueil aux urgences de personnes très fragiles ne facilite pas une collecte exhaustive. Ainsi un processus robuste de qualification des données a été mené afin d'assurer une consistance et une complétude les plus grandes possibles de la base de données pour garantir des résultats aussi fiables que possible. La base de données originale contenait 2 802 patients et 263 variables. De nombreuses variables, notamment liées à la duplication de certains questionnaires (à un questionnaire correspond plusieurs dizaines de variables) codés sous différentes formes, sont redondantes et donc éliminées.

Puis, en accord avec les psychiatres, désireux d'avoir rapidement de premiers résultats, il a été décidé de ne garder que les variables qui satisfaisaient un taux de complétion minimum de 70%. La modélisation de la structure et de la typologie des données manquantes et l'utilisation de méthodes d'imputation seront réalisées ultérieurement. Ensuite, les 34 questions du questionnaire d'impulsivité de Barratt (BIS10, Patton et al. (1995)) ont été traitées afin de créer trois scores d'impulsivité : impulsivité motrice (11 questions), attentionnelle (11 questions), ou de non planification (12 questions). Pour chaque question du BIS10, le score prend une valeur entre 1 (faible impulsivité) et 4 (forte impulsivité). L'étendue théorique du score total est donc $[34 - 136]$ ($[11 - 44]$ pour les impulsivités motrice et attentionnelle, et $[12 - 48]$ pour la non planification). En appliquant le filtre strict de 70% de complétion, 26 variables sont conservées. Enfin, en accord avec les psychiatres, il a été décidé de ne conserver que les suicidants présentant un taux de complétion de 100% pour ces 26 variables, puis de supprimer 6 variables pour des questions de pertinence. Finalement, le processus de qualification conduit à un jeu de données final de 681 patients et 20 variables d'intérêt dont la qualité garantit des résultats significatifs et robustes. Le tableau 1 présente les principales caractéristiques sociologiques (à gauche) et cliniques (à droite) des 681 patients pour les 20 variables. De façon très synthétique, les patients sont plutôt jeunes (âge médian : 40 ans), plutôt des femmes, avec une situation professionnelle de salarié(e) et un statut marital varié, et ayant un historique de trouble mental, en particulier des épisodes sérieux de dépression ($70,3\%$) ou des troubles bipolaires (23%).

Caractéristiques sociologiques			
Variables qualitatives			
Variable	Modalité	n	%
Genre	Féminin	493	72,4
	Masculin	188	27,6
Statut marital	Célibataire	239	35,1
	Marié(e)	240	35,2
	Séparé(e)/Divorcé(e)	181	26,6
	Veuf/Veuve	21	3,1
Enfants	Non	272	39,9
	Oui	409	60,1
Niveau d'éducation	Faible	31	4,6
	Intermédiaire	368	54,0
	Élevé	282	41.4
Situation professionnelle	Salarié(e)	451	66,2
	Sans emploi	110	16,2
	En incapacité	41	6,0
	Retraité(e)	79	11,6
Variables quantitatives			
Variable	Médiane	Q_1	Q_3
Age	40,6	28,0	49,6

Caractéristiques cliniques			
Variables qualitatives			
Variable	Modalité	n	%
Trouble mental connu	Non	6	0,9
	Oui	675	99.1
Comportement suicidaire familial connu	Non	424	62,3
	Oui	257	37,7
Épisode dépressif connu	Non	199	29,2
	Oui	482	70,8
Trouble bipolaire connu	Non	521	76,5
	Oui	160	23,5
Trouble dysthymique connu	Non	651	95,6
	Oui	30	4,4
Trouble Compulsif Obsessionnel connu	Non	623	91,5
	Oui	58	8,5
Trouble alimentaire connu	Non	571	83,8
	Oui	110	16,2
Prise d'alcool/drogue par le passé	Non	465	68,3
	Oui	216	31,7
Prise de substance	Non	586	86,0
	Oui	95	14,0
Alcoolisme	Non	503	73,9
	Oui	178	26,1
Variables quantitatives			
Variable	Médiane	Q_1	Q_3
Nombre de TS	2	1	3
BIS10 impulsivité motrice	26	22	30
BIS10 attention	27	23	30
BIS10 non-planification	28	24	31

TAB. 1: Caractéristiques sociologiques et cliniques des 681 patients et des 20 variables retenus. Q1 et Q3 : 1er et 3ème quartiles, respectivement. BIS10 : *Barrat Impulsiveness Scale*.

3.3 Méthodes de fouille de données

Nous rappelons que, en concertation avec les psychiatres, nous cherchons à d'une part identifier des groupes de patients similaires et d'autre part identifier des facteurs de risque associés au nombre de tentatives de suicide (`TS` par la suite) par patient. Nous ne présentons pas les résultats de l'analyse statistique descriptive par manque de place.

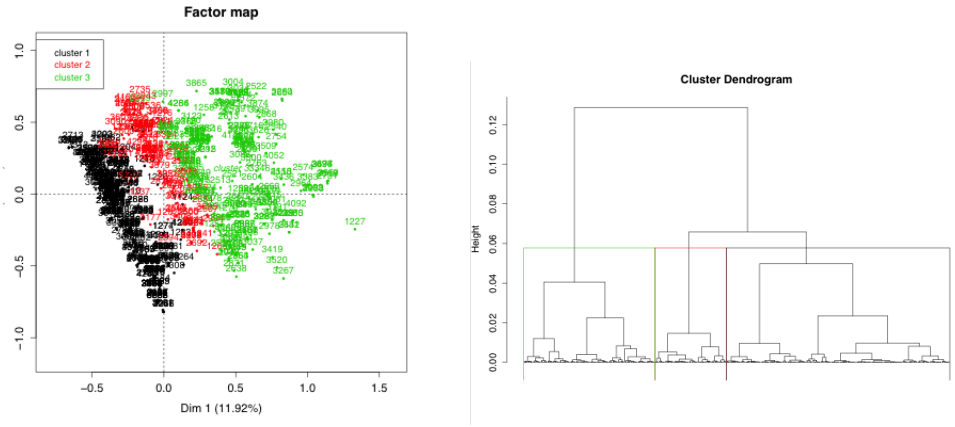
3.4 Résultats

3.4.1 Clustering des patients

Afin de mettre en lumière des groupes de patients partageant des caractéristiques similaires, nous avons procédé en deux étapes. Tout d'abord, une analyse des correspondances multiples (ACM) a été effectuée sur les seules variables qualitatives. Cette méthode factorielle, bien adaptée à l'analyse de questionnaires, représente les individus dans un nouvel espace où chaque dimension est une combinaison des variables de départ. Le nombre de dimensions retenues pour l'ACM est de 5. Les variables quantitatives comme l'âge ne sont pas utilisées pour le calcul des composantes principales, mais projetées ensuite sur le plan factoriel et utilisées pour l'interprétation. Cette méthode, habituellement utilisée en tant que technique de réduction de dimension, sert ici d'étape préalable à l'obtention de clusters robustes. Elle permet également, en transformant les variables qualitatives en variables continues, de représenter les individus dans un nouvel espace auquel on peut associer une métrique. Nous appliquons alors une Classification Hiérarchique sur Composantes Principales sur les individus dans le nouvel espace factoriel. Une classification ascendante hiérarchique permet de regrouper itérativement les paires de clusters les plus proches, en partant de singletons (chaque individu est seul dans son groupe) jusqu'à réunion en un unique cluster de tous les individus. La méthode construit un arbre binaire hiérarchique nommé dendrogramme, qui permet une interprétation visuelle des données et de la proximité entre individus via la hauteur de la branche les reliant. Elle permet d'identifier ainsi des groupes de patients partageant des facteurs de risque similaires et facilite les interactions avec les experts du domaine, mais non experts en fouille de données, afin de décider du nombre de clusters. Chaque cluster peut ensuite être interprété à travers la significativité de son association avec les modalités des différentes variables de départ (*v-test*).

La figure 1 montre (a) une projection en deux dimensions des individus sur le premier plan factoriel, (b) le dendrogramme. La structure de l'arbre (en terme de gain d'inertie dans la hiérarchie) et une discussion avec les experts en santé mentale ont mené à la sélection de trois clusters, également projetés sur le plan factoriel (a posteriori). Nous avons utilisé pour cela le package R `FactoMineR`. Une analyse approfondie des trois groupes a été réalisée pour la phase d'interprétation. Des tests d'association mettent en exergue les modalités sur ou sous représentées au sein des trois groupes. Les principales conclusions sont les suivantes : le groupe 1 correspond à un profil moyen plutôt féminin ($p < 0,001$), sans troubles bipolaires ni prises de drogues, substance, ou consommation d'alcool ($p < 0,001$). En revanche, cette population a en moyenne déjà subi des épisodes de dépression ($p < 0,001$) ou des troubles mentaux divers ($p < 0,05$). Le troisième groupe s'oppose au premier en exhibant un profil de patients plutôt masculin ($p < 0,001$), consommateur d'alcool, substances, ou drogues ($p < 0,001$). Les patients de ce groupe sont souvent célibataires et sans enfants ($p < 0,05$), et ne sont pas associés à un historique d'épisode dépressif connu ($p < 0,05$). Entre ces deux

groupes, on trouve le groupe 2, neutre en termes de genre, mais marqué par des personnes présentant une incapacité au travail ($p < 0,05$), un faible niveau d'éducation ($p < 0,05$) avec de possibles troubles bipolaires ($p < 0,001$) mais aucune consommation des produits évoqués précédemment ($p < 0,001$). Enfin, ces personnes ne présentent pas d'autres troubles de la santé mentale ou d'épisodes dépressifs connus ($p < 0,001$).



(a) Plan factoriel de l'ACM

(b) Arbre hiérarchique (dendrogramme)

FIG. 1: Approche de réduction de dimension : projection des individus statistiques sur les deux premières dimensions et dendrogramme menant au choix de trois groupes.

Ces premiers résultats montrent la capacité de la méthode à identifier des facteurs de risques déjà connus des praticiens comme le sexe, l'impulsivité, la prise de toxique. La méthode fait émerger également des facteurs de risque qui sont moins fréquemment explorés alors qu'ils semblent peser, d'après nos résultats, sur le risque de récidive. Au total, cette méthode couplée à des entretiens cliniques structurés permettrait d'obtenir pour chaque patient un niveau de risque précis sur lequel s'appuierait la décision de prise en charge.

3.4.2 Identification de facteurs de risques pour le nombre de `TS`

La deuxième phase de l'analyse a consisté à identifier des facteurs associés la multi-récidives. La variable d'intérêt est donc le nombre de `TS` que nous allons estimer par arbres binaires de régression, afin de faciliter l'interprétation des résultats par les psychiatres.

La construction d'un arbre binaire est fondée sur une séquence récursive de divisions des individus en sous-populations à l'aide de tests binaires (test d'égalité pour les variables qualitatives, test d'infériorité pour les variables quantitatives) sur les variables dites prédictives (les facteurs de risque). L'ensemble des individus est regroupé à la racine de l'arbre puis chaque division sépare chaque nœud en deux nœuds plus homogènes que le nœud parent. La variable retenue pour chaque division est celle minimisant l'écart quadratique à la moyenne de `TS` des individus dans les nœuds créés. L'arbre, de sa racine aux nœuds terminaux, hiérarchise ainsi les variables selon leur capacité à créer des groupes homogènes selon le nombre de `TS` .

Nous avons utilisé la méthode CART (Breiman et al., 1984) du package `Rpart` de R. Le paramètre de complexité, qui détermine le seuil minimal d'amélioration de l'erreur relative, pour réaliser une division, a été fixé à 0.01, et chaque nœud doit contenir au moins 10 individus. Ces choix ont été réalisés avec les psychiatres qui souhaitaient avoir des sous-groupes d'au moins une dizaine de personnes, plus facilement interprétables. Nous ne présentons ici que les résultats selon le genre. Les arbres ont été construits par validation croisée sur 10 échantillons (10-fold) et le paramètre de complexité final retenu, minimisant l'erreur relative globale, était de 0,019 pour l'arbre de genre féminin (figure 3), et 0,017 pour l'arbre de genre masculin (figure 2). Les arbres ont été interprétés et font sens pour les psychiatres.

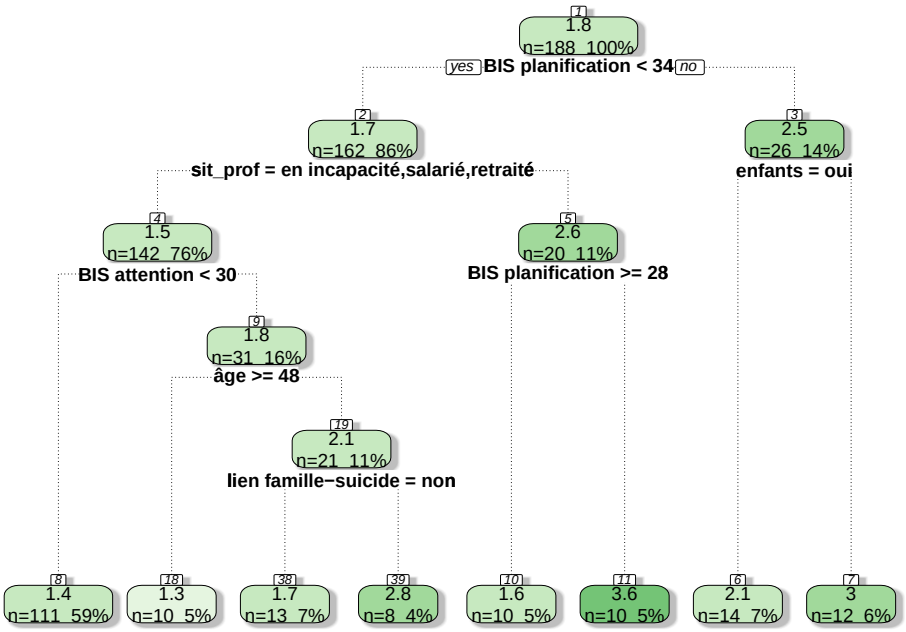


FIG. 2: Arbre de régression de la variable "nombre de TS" pour la population masculine.

L'arbre de régression des patients masculins fait apparaître des facteurs de risque propres aux hommes comme le statut professionnel ou un historique de comportement suicidaire dans la famille (nombre de `TS` moyen 2,8 vs. 1,7 sinon, $p < 0,05$). D'autre part, le fait de ne pas avoir d'enfants entraîne un sur-risque, qui n'apparaît pas chez les femmes. Pour les deux sous-populations, il est notable de voir que les scores d'impulsivité sont des facteurs explicatifs du nombre de `TS` . Pour les femmes, la notion de troubles alimentaires émerge, ce qui constitue un résultat important. En effet, pour les femmes salariées présentant une impulsivité motrice supérieure au score moyen, le nombre moyen de `TS` est de 2,9 pour les 68 femmes présentant des troubles alimentaires contre 2,3 pour les 202 autres ($p < 0,05$). Ce facteur de risque est peu questionné par les praticiens alors que nos résultats indiquent qu'il semble être important.

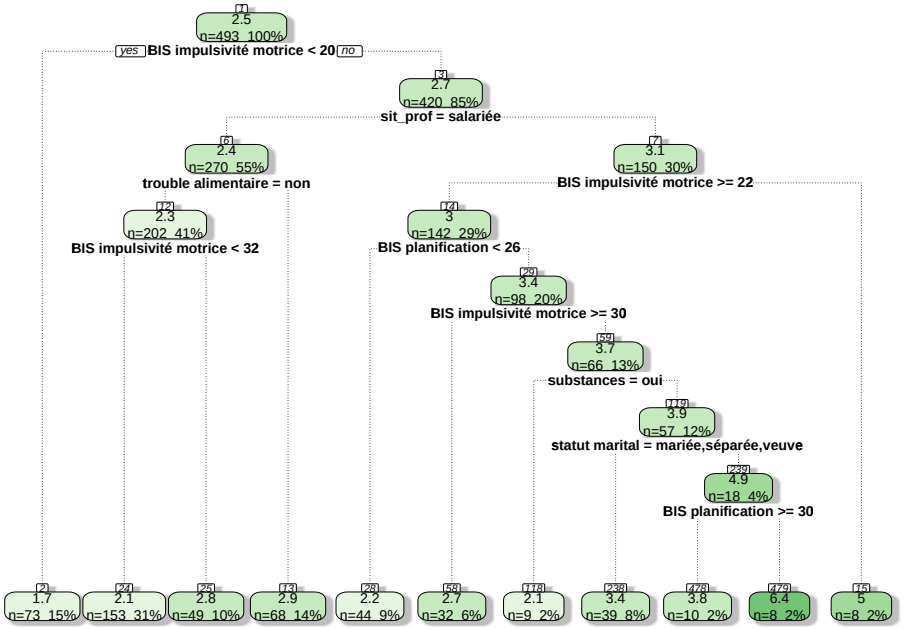


FIG. 3: Arbre de régression de la variable "nombre de TS" pour la population féminine.

4 Bilan, conclusion et perspectives

Une évaluation systématique de suicidants a permis de constituer une base de données de taille suffisante pour rendre significatif l'emploi de techniques de fouille de données. Nous avons évité d'employer le mot "prédiction". Il ne faut pas confondre évaluation du risque et prédiction du passage à l'acte. C'est sur le premier point que nous tentons modestement de contribuer, la prédiction d'un acte aussi complexe étant actuellement illusoire. Nous avons identifié des clusters de patients puis estimé le nombre de TS en fonction de facteurs de risque. Certains, comme les troubles alimentaires pour les femmes, sont aujourd'hui peu recherchés par les psychiatres participant à cette étude. Des investigations poussées sont nécessaires sur ce point pour revoir éventuellement les entretiens. Nos résultats, couplés à des entretiens structurés, permettraient d'obtenir pour chaque patient un niveau de risque précis sur lequel s'appuierait la décision de prise en charge. Le volume de données assure d'ores et déjà une représentativité statistique, mais ce travail met en évidence un problème de qualité des données dont les psychiatres ont pris conscience ainsi que des mesures à mettre en place. De nouvelles données, de meilleure qualité et contenant la temporalité des événements, sont en cours de constitution, permettront d'étudier la temporalité des TS et d'ouvrir de nouveaux champs d'investigations.

Remerciements. Nous tenons à remercier les relecteurs pour leurs remarques constructives et suggestions ayant permis d'améliorer la version finale de cet article.

Références

Abboute, A., Y. Boudjeriou, G. Entringer, J. Azé, S. Bringay, et P. Poncelet (2014). *Mining Twitter for Suicide Prevention*, pp. 250–253. Cham : Springer International Publishing.

Aveline, F., C. Baudelot, M. Beverraggi, et S. Lahlou (1984). Suicide et rythmes sociaux. *Economie et statistique 168*(1), 71–76.

Barrigón, M. L., S. Berrouiguet, J. J. Carballo, C. Bonal-Giménez, P. Fernández-Navarro, B. Pfang, D. Delgado-Gómez, P. Courtet, F. Aroca, J. Lopez-Castroman, et al. (2017). User profiles of an electronic mental health tool for ecological momentary assessment: MEmind. *International Journal of Methods in Psychiatric Research 26*(1), 9p.

Beck, R. W., J. B. Morris, and A. T. Beck (1974). Cross-validation of the suicidal intent scale. *Psychological reports 34*(2), 445–446.

Berrouiguet, S., M. L. Barrigón, S. A. Brandt, G. C. Nitzburg, S. Ovejero, R. Alvarez-Garcia, J. Carballo, M. Walter, R. Billot, P. Lenca, et al. (2017). Ecological assessment of clinicians' antipsychotic prescription habits in psychiatric inpatients: A novel web-and mobile phone–based prototype for a dynamic clinical decision support system. *Journal of medical Internet research 19*(1), 9p.

Berrouiguet, S., M. L. Barrigón, S. A. Brandt, S. Ovejero-García, R. Álvarez-García, J. J. Carballo, P. Lenca, P. Courtet, E. Baca-García, MEmind Study Group, et al. (2016a). Development of a web-based clinical decision support system for drug prescription: Non-interventional naturalistic description of the antipsychotic prescription patterns in 4345 outpatients and future applications. *PLoS ONE 11*(10), 16p.

Berrouiguet, S., R. Billot, P. Lenca, E. Baca-Garcia, B. Gourvennec, M. Simonnet, and P. Tanguy (2016b). Toward E-Health Applications for Suicide Prevention. In *CHASE 2016: IEEE First Conference on Connected Health: Applications, Systems and Engineering Technologies*, pp. 346–347. IEEE Computer Society.

Berrouiguet, S., P. Courtet, M. Perez-Rodriguez, M. Oquendo, and E. Baca-Garcia (2015). The memind project: a new web-based mental health tracker designed for clinical management and research. *European Psychiatry 30*(Supplement 1), 974.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. Stone (1984). *Classification and Regression Trees*. Wadsworth.

Chan, M. K., H. Bhatti, N. Meader, S. Stockton, J. Evans, R. C. O'Connor, N. Kapur, and T. Kendall (2016). Predicting suicide following self-harm: systematic review of risk factors and risk scales. *The British Journal of Psychiatry*, 7p.

Choo, C., J. Diederich, and I. S. adn Roger Ho (2014). Cluster analysis reveals risk factors for repeated suicide attempts in a multi-ethnic asian population. *Asian Journal of Psychiatry 8*, 38–42.

Combes, P., S. Combes, et M. Monziols (2016). Tentatives de suicide, prédire la récidive avec des techniques d'apprentissage statistique. In *Atelier sur l'Intelligence Artificielle et la Santé (Journées francophones d'Ingénierie des Connaissances)*, Montpellier, France, pp. 21p.

Delgado-Gómez, D., H. Blasco-Fontecilla, A. A. Alegria, T. Legido-Gil, A. Artés-Rodríguez, et E. Baca-Garcia (2011). Improving the accuracy of suicide attempter classification. *Artificial Intelligence in Medicine 52*(3), 165–168.

Durkheim, É. (1897). *Le Suicide : Étude de sociologie*. Félix Alcan, Paris.

Fauré, C. (2008). Suicide d'un proche : l'impact sur l'entourage. *Perspectives Psy 47*(4), 359–364.

Finkelstein, Y., E. M. Macdonald, S. Hollands, M. L. Sivilotti, J. R. Hutson, M. M. Mamdani, G. Koren, et D. N. Juurlink (2015). Risk of suicide following deliberate self-poisoning. *JAMA psychiatry 72*(6), 570–575.

Glenn, C. R. and M. K. Nock (2014). Improving the short-term prediction of suicidal behavior. *American Journal of Preventive Medicine 47*(3), S176–S180368–369.

Karmakar, C., W. Luo, T. Tran, M. Berk, and S. Venkatesh (2016). Predicting risk of suicide attempt using history of physical illnesses from electronic medical records. *JMIR mental health 3*(3), 10p.

Kessler, R., C. Warner, C. Ivany, and other authors (2015). Predicting suicides after psychiatric hospitalization in us army soldiers: The army study to assess risk and resilience in servicemembers (army starrs). *JAMA Psychiatry 72*(1), 49–57.

Lecrubier, Y., D. V. Sheehan, E. Weiller, P. Amorim, I. Bonora, K. H. Sheehan, J. Janavs, and G. C. Dunbar (1997). The Mini International Neuropsychiatric Interview (MINI). A short diagnostic structured interview: reliability and validity according to the CIDI. *European Psychiatry 12*(5), 224–231.

Lopez-Castroman, J., H. Blasco-Fontecilla, P. Courtet, E. Baca-Garcia, and M. A. Oquendo (2015). Are we studying the right populations to understand suicide? *World Psychiatry 14*(3), 368–369.

Lopez-Castroman, J., E. Nogue, S. Guillaume, M. C. Picot, and P. Courtet (2016). Clustering suicide attempters: Impulsive-ambivalent, well-planned, or frequent. *J Clin Psychiatry 77*(6), e711–e718.

Loranger, A., N. Sartorius, A. Andreoli, P. Berger, P. Buchheim, S. Channabasavanna, B. Coid, A. Dahl, R. Diekstra, B. Ferguson, L. Jacobsberg, W. Mombour, C. Pull, Y. Ono, and D. Regier (1994). The International Personality Disorder Examination. The World Health Organization/Alcohol, Drug Abuse, and Mental Health Administration International Pilot Study of Personality Disorders. *Archives of General Psychiatry 51*(3), 215–224.

Lytle, M., V. Silenzio, and E. Caine (2016). Are there still too few suicides to generate public outrage? *JAMA Psychiatry 73*(10), 1003–1004.

Maigrot, C., S. Bringay, et J. Azé (2016). Concept drift vs suicide : comment l'un peut prévenir l'autre ? In *Extraction et Gestion des Connaissances*, Volume E-30 of *RNTI*, Reims, France, pp. 219–230. Hermann-Éditions.

Montaigne, I. (2014). Prévention des maladies psychiatriques : pour en finir avec le retard français. Institut Montaigne.

Moreno, M. A., L. A. Jelenchick, K. G. Egan, E. Cox, H. Young, K. E. Gannon, and T. Becker (2011). Feeling bad on Facebook: depression disclosures by college students on a social networking site. *Depress Anxiety 28*(6), 447–455.

ONS (2016). SUICIDE - Connaître pour prévenir : dimensions nationales, locales et associatives. Observatoire National du Suicide.

Patton, J. H., M. S. Stanford, and E. S. Barratt (1995). Factor structure of the barratt impulsiveness scale. *Journal of clinical psychology 51*(6), 768–774.

Rakesh, G. (2017). Suicide prediction with machine learning. *The American Journal of Psychiatry Residents' Journal 12*(1), 15–17.

Ribeiro, J. D., J. C. Franklin, K. R. Fox, K. H. Bentley, E. M. Kleiman, B. P. Chang, and M. K. Nock (2016). Letter to the Editor: Suicide as a complex classification problem: machine learning and related techniques can advance suicide prediction - a reply to Roaldset (2016). *Psychological Medicine 46*(9), 2009–2010.

Smith, K. (2011). Trillion-dollar brain drain. *Nature News 478*, 15.

Tran, T., W. Luo, D. Phung, R. Harvey, M. Berk, R. L. Kennedy, and S. Venkatesh (2014). Risk stratification using data from electronic medical records better predicts suicide risks than clinician assessments. *BMC Psychiatry 14*(1), 76.

Walsh, C., J. D. Ribeiro, and J. C. Franklin (2017). Predicting risk of suicide attempts over time through machine learning. *Clinical Psychological Science 5*(3), 457–469.

Weisman, A. D. and J. W. Worden (1972). Risk-rescue rating in suicide assessment. *Archives of General Psychiatry 26*(6), 553–560.

WHO (2014). Preventing Suicide - A global imperative. World Health Organization.

Wolodzko, T. and A. Kokoszka (2014a). Characteristics of groups after the suicide attempt. Cluster analysis of National Comorbidity Survey (NCS) 1990-1992. *Psychiatr. Pol. 48*(6), 1253–1267.

Wolodzko, T. and A. Kokoszka (2014b). Classification of persons attempting suicide. A review of cluster analysis research. *Psychiatr. Pol. 48*(4), 823–834.

Summary

Over 800 000 people die due to suicide every year and it is estimated that for each suicide there may have been more than 20 others attempting suicide, involving huge human and societal costs. In recent years, digital tools have changed the way data are collected on patients. We present the main results of a comprehensive data mining process carried out on a sample of suicidal patients from two European hospitals. The first objective is to identify groups of similar patients and the second objective is to identify risk factors associated with the number of attempts. Unsupervised methods (ACM and clustering) and supervised methods (regression trees) are applied to address these two research objectives. The results highlight the high potential of data mining for descriptive or explanatory purposes.