Méthodologie d'analyse et de surveillance pour la prévention des arrêts maladie

Tom Duchemin

HESAM UNIVERSITÉ

le cnam

ÉCOLE DOCTORALE Sciences et Métiers de l'Ingénieur
Laboratoire MESuRS

THÈSE

présentée par : **Tom DUCHEMIN**
soutenue le : **12 novembre 2020**

pour obtenir le grade de : **Docteur d'HESAM Université**

préparée au : **Conservatoire national des arts et métiers**

Discipline : **26 - Mathématiques appliquées et application des mathématiques**

Spécialité : **Sécurité Sanitaire**

Méthodologie d'analyse et de surveillance pour la prévention des arrêts maladie

THÈSE dirigée par :
Mme Mounia N. HOCINE MCF HDR, Le Cnam

Jury

Mme Virginie RONDEAU	Directrice de Recherche, Inserm-Université de Bordeaux	Président
M. Yann LE STRAT	PhD HDR, Directeur DATA, Santé Publique France	Rapporteur
Mme Marie ZINS	PU-PH, Inserm-UVSQ	Rapporteur
Mme Mounia N. HOCINE	MCF HDR, Le Cnam	Examinateur
M. Geert MOLENBERGHS	Professeur, Katholieke Universiteit Leuven	Examinateur
Mme Catherine POLLAK	PhD, Directrice de pôle, DREES	Examinateur
M. Gilbert SAPORTA	Professeur émérite, Le Cnam	Examinateur
M. Raphaël SOULIGNAC	Directeur *Data Science*, Malakoff Humanis	Invité

THÈSE

Remerciements

Je souhaiterais remercier toutes celles et ceux qui ont contribué de près ou de loin aux travaux présentés dans ce mémoire.

Je souhaiterais tout d'abord remercier Mounia Hocine qui a dirigé cette thèse. J'ai beaucoup apprécié travailler avec toi pendant ces trois années et j'ai beaucoup appris. Merci pour ton temps et tes conseils toujours pertinents.

Je souhaiterais remercier mes collaborateurs de Malakoff Humanis qui ont permis à cette thèse d'exister et de se dérouler dans les meilleures conditions. Merci à Radowan Lounissi qui m'a accompagné pendant ces trois ans et qui m'a permis de comprendre très vite le contexte de ce projet, merci pour tout le temps passé sur ce projet et, surtout, merci pour tous les bons moments passés ici ! Merci à Raphaël Soulignac de m'avoir accueilli dans l'équipe *Data Science* et de m'avoir suivi pendant tout ce temps : les temporalités de la thèse et de l'entreprise semblent parfois inconciliables et pourtant tu as réussi à les unir pour le mieux. Merci à Stéphane Barde pour ses précieux conseils et le temps passé sur ce projet. Enfin, merci à Anne-Sophie Godon qui a permis de lancer cette thèse.

Je souhaiterais ensuite remercier les co-auteurs et toutes celles et ceux qui ont fortement collaboré aux travaux de cette thèse. Merci à William Dab dont les conseils bienveillants ont beaucoup nourri ces travaux. Merci à Angela Noufaily de m'avoir accueilli deux semaines à l'université de Warwick et d'avoir suivi nos travaux autour de la surveillance : c'était un plaisir de travailler avec toi et j'ai beaucoup appris. Merci à Avner Bar-Hen d'avoir participé, d'une manière non négligeable, à nos interrogations méthodologiques. Merci enfin à Laura Temime et Kévin Jean qui ont toujours su formuler des remarques pertinentes et qui ont permis à cette thèse de se dérouler pour le mieux au Cnam.

Je souhaiterais remercier tous les membres du jury qui ont accepté de donner de leur temps précieux

pour évaluer ces travaux. Merci à Yann Le Strat et à Marie Zins de rapporter cette thèse. Merci à Geert Molenberghs, à Catherine Pollak, à Virginie Rondeau et à Gilbert Saporta de l'examiner.

Je souhaiterais remercier les membres du comité de suivi de thèse, Avner Bar-Hen, Vincent Lefieux, Nikolaos Limnios et Gilbert Saporta, qui m'ont permis d'avancer grâce à leurs conseils pertinents.

Je souhaiterais aussi remercier tous les membres du laboratoire MESuRS : Audrey, David, Hélène, Isabelle(s), Jérôme, Jonathan, Rania, Narimane, Oumou, Paul, Pearl et ceux que j'ai oublié. J'ai passé trois très bonnes années de thèse et c'est en grande partie grâce à vous. Merci pour l'environnement plus que favorable pour une thèse et merci pour tous les bons moments passés ! Merci aussi pour les quelques kilos pris pendant les Tea Times.

Je souhaiterais remercier tous les gens que j'ai pu côtoyer chez Malakoff Humanis. J'ai vécu pendant ces trois ans énormément de changements dans l'entreprise (suite à une fusion et à la construction d'un pôle *Data*) et j'ai croisé tellement de gens que j'en oublierai forcément : pour ne vexer personne, je ne citerai donc personne ou presque. Merci à toute l'équipe *Data Science* pour tous ces bons moments : l'environnement de travail est extrêmement positif et j'ai pu apprécier tous nos échanges et ces tonnes de viennoiseries englouties. Merci à l'équipe Marketing Stratégique et surtout à Francis Massie de m'avoir accueilli pendant les premiers mois de ma thèse. Merci à Thibaut Tan de m'avoir accueilli dans son bureau du Cadran. Merci à toutes les équipes travaillant sur le *Diagnostic et Protection du Capital Humain* pour tous ces échanges qui ont enrichi mes travaux et m'ont même appris bien plus.

Je souhaiterais enfin remercier ma mère, ma soeur, Emma, ma famille et mes amis qui m'accompagnent tous les jours.

Résumé

Alors que les arrêts maladie sont le signe d'un mal-être croissant chez les salariés et qu'ils pèsent un coût certain pour la collectivité, la numérisation et le partage systématique des données offrent de belles opportunités pour leur prévention. Nous avons ainsi profité de cette opportunité pour développer un éventail d'outils de prévention basés sur des méthodes d'analyse statistique. Dans un premier temps, ces travaux de thèse proposent une analyse des mécanismes expliquant les arrêts maladie chez le salarié. L'analyse d'une enquête nationale a premièrement permis d'identifier et de hiérarchiser leurs principaux facteurs déterminants grâce à l'algorithme des forêts aléatoires. Ensuite, une analyse de données administratives a identifié des trajectoires d'absence pouvant mener à des arrêts graves grâce à des analyses séquentielles et à de la modélisation multi-état. Dans un second temps, des outils ont été développés afin d'identifier des situations anormales d'arrêt maladie à l'échelle de l'entreprise. Une typologie d'entreprise a premièrement été construite afin de produire des valeurs repère pour que les entreprises évaluent précisément leur situation. Un algorithme de détection des pics d'absence, adapté de modèles de surveillance épidémiologique, a enfin été développé pour pouvoir identifier automatiquement les entreprises en excès.

Mots-clés : santé au travail, prévention, arrêt maladie, surveillance, forêt aléatoire, statistiques.

Abstract

At a time when sick leave is a sign of growing ill-being for workers and a cost burden for the society, the systematic digitalization and distribution of data offers great opportunities for its prevention. We have therefore taken advantage of this opportunity to develop a range of prevention tools based on statistical analysis methods. In a first part, this work proposes an analysis of the mechanisms explaining sick leave among workers. The analysis of a national survey has first identified and prioritised their main determinants using random forest. Then, an analysis of administrative data had helped to identify absence trajectories that could lead to serious sick leaves thanks to sequential analyses and multi-state modelling. In a second step, tools were developed to identify abnormal situations of sick leave at company level. A company typology was first built to produce benchmark values for companies to accurately assess their situation. Finally, an algorithm for identifying absence peaks, adapted from epidemiological surveillance models, was finally developed to automatically identify companies in difficulty.

Keywords : occupational health, prevention, sick leave, surveillance, random forest, statistics.

Table des matières

Liste des tableaux

Table des figures

Production scientifique

Articles dans des revues internationales à comité de lecture

— Duchemin T., Bar-Hen A., Lounissi R., Dab W., Hocine M.N., Hierarchizing Determinants of Sick Leave, Insight From a Survey on Health and Well-Being at the Workplace. *Journal of Occupational and Environmental Medicine*. 2019 ; 61(1). https://dx.doi.org/10.1097/JOM.0000000000001643

— Duchemin T., Bar-Hen A., Lounissi R., Dab W., Hocine M.N., Response to Predictors of Long-Term Sick Leave in the Workplace. *Journal of Occupational and Environmental Medicine*. 2019 ; 61(1). https://dx.doi.org/10.1097/JOM.0000000000001726

— Duchemin T., Hocine, M.N., Modeling sickness absence data : a scoping review. PLoS ONE 15(9) : e0238981. 2020. *Plos One*, 2020. https://doi.org/10.1371/journal.pone.0238981

— Duchemin T., Noufaily A., Hocine M.N., A statistical algorithm for outbreak detection in a multi-site setting : the case of sick leave monitoring. (soumis)

— Duchemin T., Bastard J., Ante-Testard P.A., Assab R., Daouda O.S., Duval A., Garsi, J.-P., Lounissi R., Nekkab N., Neynaud H., Smith D.R.M., Dab W., Jean K., Temime L., Hocine M.N., Monitoring sick leave data for early detection of influenza outbreaks. (en révision) https://doi.org/10.1101/2020.05.28.20115782

Conférences et colloques

— IBC 2018 : International Biometrics Conference, Barcelona, Spain.
Communication orale "Analysing sickness absence data using semi-Markov models" avec Mounia N. Hocine.

— JdS 2018 : Journée des Statistiques, Palaiseau, France.

Communication orale "Prédire l'absence au travail pour raison de maladies : méthodologies et résultats issus d'une base de données de Malakoff Médéric" avec Marianne Fabre, Radowan Lounissi et Mounia N. Hocine

— Preventica 2019, Paris, France.
Communication orale "Comment analyser les données d'absentéisme pour mieux le comprendre et le maîtriser ?" avec Mounia N. Hocine

— IBC 2020 : International Biometrics Conference, Seoul, South Korea.
Communication orale "A statistical algorithm for sick leave outbreak detection at the workplace" avec Angela Noufaily et Mounia N. Hocine

— JdS 2020 : Journée des Statistiques, Nice, France.
Communication orale : "Modèle de détection d'anomalies pour données longitudinales : application aux arrêts maladie" avec Angela Noufaily et Mounia N. Hocine.

Chapitre 1

Contexte et enjeux

Contenu

1.1 Les arrêts maladie

La cause des arrêts maladie est déjà sous-entendue dans leur énoncé : les salariés ont des arrêts maladie car ils sont malades. Cette simple phrase ne suffit malheureusement pas à constituer une thèse et la vérité s'avère être plus complexe. Les arrêts maladie ne sont pas le fruit d'un processus totalement aléatoire et de nombreux déterminants, complexes, peuvent les expliquer et permettre de les prévenir. Ce mémoire tente de comprendre ce phénomène en identifiant déterminants, processus et comportements qui témoignent d'un risque accru d'arrêts maladie.

1.1.1 Définition

Avant d'entrer dans le cœur du sujet, définissons ce que nous entendons par arrêts maladie.

L'arrêt maladie est défini légalement comme une absence du travail pour raison d'accident ou de maladie non-professionnels qui doit être constatée par un médecin. Le salarié absent doit justifier son arrêt dans les 48 heures auprès de son employeur et de l'assurance maladie en leur présentant un certificat médical.

L'arrêt maladie ne doit pas être confondu avec d'autres types d'arrêts comme les congés parentaux ou les accidents du travail et maladies professionnelles. Sauf mention explicite, nous ne parlerons que d'arrêts maladie dans ce mémoire.

Les arrêts maladie peuvent être prescrits par le médecin pour différents motifs, qui sont indiqués sur le certificat médical. Dans les travaux de ce mémoire, les arrêts maladie seront analysés sans distinction de motif puisque l'information n'est pas disponible dans les données utilisées. Nous souhaiterions aussi mentionner que les arrêts maladie sont définies dans les textes légaux comme des arrêts pour maladie non-professionnelle : cette distinction est cependant en pratique plus complexe puisque la déclaration d'une maladie professionnelle est un processus complexe et des maladies professionnelles pourraient donc être le motif réel de certains arrêts maladie.

L'arrêt maladie suspend le contrat de travail et donne droit à une indemnisation du salarié. La procédure d'indemnisation peut différer selon les salariés mais est, en général, partagée entre trois acteurs :

1. **L'assurance maladie** verse tout d'abord aux salariés des Indemnités Journalières (IJ) équivalentes à 50% du salaire journalier de base (avec un plafonnement à 1,8 fois le montant du Smic) après

un délai de carence de trois jours pour la plupart des salariés du secteur privée (le délai est d'un jour pour les salariés de la fonction publique et il n'y a pas de délai de carence pour les salariés travaillant en Alsace-Moselle).

2. **L'employeur** complète ensuite cette indemnité en fonction des accords collectifs de branche et de sa convention collective. Au minimum, l'employeur doit compléter l'indemnité de base pour atteindre 90% du salaire de base au-delà du huitième jour d'arrêt si le salarié a plus d'un an d'ancienneté. Ce complément est maintenu pendant le premier mois de l'arrêt puis est diminué progressivement.
3. Enfin, un **assureur complémentaire** peut compléter l'indemnisation des arrêts de travail lorsque ce dernier dépasse une durée que l'on nomme franchise. Le montant et la durée de cette indemnisation vont dépendre des contrats signés entre les employeurs et les organismes de protection sociale.

1.1.2 Conséquences des arrêts maladie

Pour le salarié, les arrêts maladie ont été mis en place afin de leur offrir un temps de repos et leur assurer un revenu lorsqu'ils ne sont pas en capacité de travailler : la première conséquence est donc de toute évidence positive. Cependant, certaines conséquences négatives peuvent être identifiées [1]. Premièrement, l'arrêt maladie peut avoir des conséquences financières puisque le salarié n'est pas toujours indemnisé à la hauteur de son salaire. Ensuite, un arrêt de durée trop longue peut mener à une situation d'isolement et d'inactivité qui peuvent créer des troubles psychologiques [2, 3].

Les arrêts maladie ont aussi des conséquences financières claires sur les acteurs chargés d'indemniser le salarié. En 2017, l'assurance maladie évalue à 7,4 milliards d'euros le montant des IJ versées pour les arrêts maladie [4]. En 2014, le *Alma Consulting Group* évalue le coût direct de l'absence pour les entreprises à 45 milliards d'euros en France [5] : le coût direct inclue le maintien du salaire du salarié absent, son remplacement éventuel et la perte de valeur ajoutée entraînée par cette absence. La méthodologie utilisée pour calculer ce coût montre que l'impact de l'absence est donc plus que financière puisqu'elle a aussi un impact sur l'organisation.

Tenter de prévenir les arrêts maladie a ainsi un impact positif sur les finances des salariés, de l'Etat et des entreprises mais aussi sur la santé des salariés. Il faut de plus ajouter que l'objectif de cette prévention n'est pas d'empêcher les salariés de s'arrêter, mais de prévenir les causes de ces arrêts :

empêcher les salariés de s'arrêter aurait bien sûr un impact désastreux sur leur santé, ce qui pourrait d'ailleurs causer des arrêts encore plus graves dans le futur.

1.1.3 Causes des arrêts maladie

Comme énoncer tautologiquement en introduction, si les salariés sont en arrêt maladie c'est parce qu'ils sont malades. Les causes de cette maladie sont cependant diverses et peuvent provenir de plusieurs sources. Il faut aussi ajouter que, pour une même maladie, un salarié peut décider ou non de prendre un arrêt de travail. Nous présenterons dans une section dédiée une revue plus complète des causes d'arrêts maladie 2.2.1 mais, pour comprendre l'objectif de ces travaux, dressons un bref panorama de ces causes.

Le graphique en Figure 1.1 présente un graphe synthétique décrivant les différents processus déterminant l'occurrence d'un arrêt maladie. Si nous simplifions, nous pouvons identifier trois catégories de déterminants pour ces arrêts :

1. les déterminants individuels, relatifs aux salariés : sa santé [6], son hygiène de vie[7], son âge [8, 9], *etc.*
2. les déterminants professionnels, relatifs aux entreprises : la pénibilité physique de l'emploi [10, 11], la pression exercée sur les salariés [12], *etc.*
3. les déterminants exogènes, relatifs à tout le reste : maladies saisonnières [13], cadre légal [14] , pandémie [15], *etc.*

Notre objectif sera d'explorer différentes méthodologies pour identifier et prévenir ces différents déterminants d'arrêts de travail.

1.2 Terrain d'étude

Les travaux de cette thèse se sont déroulés dans le cadre d'un contrat CIFRE entre le Cnam et Malakoff Humanis qu'il convient d'expliquer pour saisir les objectifs et le déroulé des travaux présentés.

1.2.1 Entreprise et laboratoire d'accueil

L'entreprise Malakoff Humanis est un groupe de protection sociale (c'est-à-dire un organisme qui met en œuvre des régimes de retraites complémentaires et des couvertures de protection sociale com-

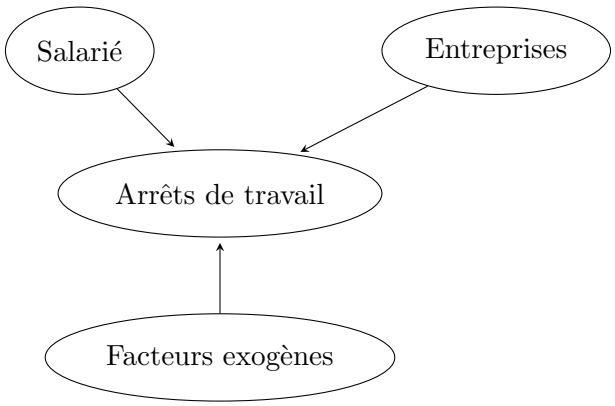

FIGURE 1.1 – Sources des causes potentielles d'arrêt maladie

plémentaire), paritaire (piloté par les partenaires sociaux), mutualiste et à but non-lucratif. Malakoff Humanis est né suite au rapprochement entre Malakoff Médéric et Humanis, deux anciens groupes de protection social, le 1er janvier 2019 : les travaux présentés dans cette thèse ont débuté chez Malakoff Médéric, avant ce rapprochement.

Malakoff Humanis est concerné par la problématique des arrêts de travail puisqu'elle propose des contrats de prévoyance qui couvrent le risque d'incapacité pour les entreprises et les particuliers. L'incapacité d'un salarié est l'impossibilité temporaire à travailler (contrairement à l'invalidité, qui est une impossibilité permanente). Un salarié est dit en incapacité de travail lorsqu'il est en arrêt maladie pour une durée supérieure à une franchise déterminée par un contrat qui relie l'organisme de prévoyance et l'entreprise. Cette franchise est en général de 3 mois pour les contrats de Malakoff Humanis. L'organisme de prévoyance indemnise ainsi le salarié en complément de l'assurance maladie et de l'entreprise. Outre la problématique d'assurance, Malakoff Humanis propose de nombreuses études et de nombreux services pour comprendre et maîtriser l'absentéisme des entreprises et c'est principalement dans ce contexte que s'inscrivent ces travaux.

La chaire Entreprise et Santé et le laboratoire MESuRS En 2015 a été inaugurée la chaire *Entreprise et Santé*, fruit d'un partenariat entre le Cnam et Malakoff Médéric. L'objectif de cette chaire est d'accompagner les entreprises dans la maîtrise des enjeux de santé au travail en leur proposant des formations mais aussi en développant des axes divers de recherche et de prospective.

L'axe principal de recherche de la chaire est de mieux connaître l'efficacité des actions de prévention dans les entreprises. L'efficacité de ces actions s'évalue par la réduction des expositions au facteur de

risque, par l'amélioration des comportements favorables à la santé, par la diminution de la mortalité et de la morbidité mais aussi par la réduction de l'absence au travail pour raisons de santé.

Nos travaux s'insèrent dans cette dernière problématique et font suite à des travaux déjà engagés au laboratoire MESuRS. Des stagiaires ont notamment déjà commencé à analyser et à mettre au propre les données présentées dans ce mémoire.

1.2.2 Enjeux actuels

Les enjeux de cette thèse croisent donc les enjeux de cette Chaire et les enjeux d'un organisme de prévoyance comme Malakoff Humanis.

Un premier enjeu est de pouvoir approfondir la connaissance des arrêts maladie : qui sont les salariés absents ? pourquoi sont-ils absents ? comment réduire leur risque d'absence ? Cette connaissance permettra de nourrir des plans d'action et d'informer les entreprises sur les causes probables de leur absentéisme maladie.

Le second enjeu est de proposer des outils innovants pour maîtriser l'absentéisme maladie des salariés et donc des entreprises.

Les travaux présentés se basent sur deux sources de données : une enquête représentative de la population française, le Baromètre Santé et bien-être au Travail (BST) qui permet d'alimenter les travaux sur les causes des arrêts maladie ; des données administratives, les Déclarations Sociales Nominatives (DSN), décrivant l'ensemble des trajectoires d'absence des salariés et qui permettent de mettre en place des outils de suivi et de compréhension de l'absence au travail chez les entreprises.

Ces deux axes d'étude devraient donc nourrir des réflexions générales autour de l'analyse des données d'arrêt maladie mais devraient aussi nourrir un outil de suivi de l'absentéisme et de la santé au travail des salariés proposé par Malakoff Humanis nommé le *Diagnostic et la Protection du Capital Humain*. Cet outil prend la forme d'une application qui propose un tableau de bord de l'absentéisme de l'entreprise et qui fournit aussi des informations sur différents éléments du capital humain des entreprises, et notamment sur la santé au travail (exposition à des facteurs de risques psychosociaux, pénibilité physique, *etc.*). Le capital humain est défini par l'OCDE comme "l'ensemble des connaissances, qualifications, compétences et caractéristiques individuelles qui facilitent la création du bien-être personnel, social et économique" [16] et, comme il ne s'agit pas du sujet de ces travaux, ce terme

ne sera probablement plus utilisé dans ce mémoire.

L'objectif des travaux est de nourrir le tableau de bord de l'absence dont l'objectif est de caractériser l'absentéisme des entreprises et de sélectionner des informations pertinentes et personnalisées à partir des données à disposition. Ces deux objectifs sont très ouverts et nous allons tenter de les préciser dans les paragraphes suivants.

1.3 Objectif des travaux et axes d'analyse

1.3.1 Objectif des travaux

L'objectif principal de cette thèse est de développer des outils statistiques pour la surveillance des arrêts maladie. Dans ce mémoire, nous allons entendre la surveillance dans deux acceptions :

1. la surveillance est premièrement entendu au sens large. Étymologiquement, la surveillance est l'acte de *sur-veiller*, soit d'observer un phénomène avec un supplément d'attention. Nous appelons donc surveillance la description des facteurs qui définissent un phénomène d'intérêt (dans notre cas, les arrêts maladie).
2. la surveillance est deuxièmement entendu en son sens statistique ou épidémiologique. En épidémiologie, la surveillance est le processus de recueil, d'analyse et de partage de données de santé dans un objectif de santé public [17]. Nous définissons ainsi la surveillance comme le suivi d'un phénomène dans le temps. Ce suivi permet d'identifier des moments critiques où ce phénomène présente un comportement anormal pour ensuite lever des alertes qui informeront les différents acteurs en jeu.

Ces deux définitions s'enrichissent l'une et l'autre. La première acception permet de comprendre et d'expliquer les comportements anormaux identifiées par la seconde ; la seconde acception permet d'identifier les lieux ou instants où des actions sont nécessaires et ces actions peuvent être mises en place grâce à la première acception.

1.3.2 Plan du mémoire

Pour répondre à cet objectif, nous allons logiquement procéder en deux étapes.

1.3.2.1 Premier axe : identifier les signaux d'arrêts maladie

Dans un premier temps, nous développerons la première acception de la surveillance. Notre objectif sera d'analyser les facteurs entrant en jeu dans la prise d'arrêt maladie : quels sont les déterminants des arrêts ? est-ce qu'une trajectoire passée peut nous informer sur des trajectoires futures ?

Pour appréhender le problème, nous commencerons par présenter des éléments de méthodologie statistique couramment utilisés pour analyser les données d'absence par le biais d'une revue de littérature (*scoping review*) : comment appréhender cette donnée ? quels modèles s'offrent à nous ?

Nous présenterons ensuite les résultats d'une analyse de données d'enquête qui tente d'identifier et de hiérarchiser les déterminants des arrêts maladie en fonction de leur durée. Nous discuterons en particulier des arrêts de longue durée, qui sont un sujet de préoccupation pour beaucoup de salariés et d'entreprises. Ces données d'enquête proviennent du *Baromètre Bien-être et Santé au Travail*, une enquête représentative de la population française que nous présenterons par la suite.

Après avoir étudié les arrêts sous le prisme de leur durée, nous proposerons ensuite une analyse différente en étudiant les trajectoires d'absence : observe-t-on des trajectoires typiques d'arrêts maladie ? certaines trajectoires peuvent-elles être des signaux d'arrêts graves dans le futur ? Nous utiliserons, pour ces analyses, des données administratives décrivant des entreprises assurées par Malakoff Humanis.

1.3.2.2 Deuxième axe : détecter les excès d'arrêts maladie

Dans un second temps, nous développerons la seconde acception de la surveillance. Notre objectif sera ici d'identifier des situations où le volume d'arrêts maladie semble anormalement élevé au sein des entreprises.

Nous présenterons tout d'abord des facteurs expliquant les différences d'absentéisme entre les entreprises, une étape primordiale pour saisir les enjeux d'une telle surveillance.

Dans un second temps, nous proposerons un algorithme de surveillance statistique de l'absentéisme des entreprises. Nous présenterons pour cela une revue des méthodes utilisées en surveillance et nous proposerons une adaptation d'un algorithme couramment utilisé en pratique, le modèle dit de Farrington[18]. Ces analyses seront effectuées sur les données administratives déjà présentées précédemment.

Avant de conclure, nous ferons une brève excursion vers les causes d'arrêt que nous avons plutôt mises à part dans nos travaux : les causes exogènes. Nous expliquerons que la surveillance des données d'arrêts maladie peut permettre de surveiller d'autres phénomènes comme la grippe.

1.3.2.3 Synthèse des travaux et perspectives

La conclusion de ce mémoire tentera enfin une synthèse de ces travaux : ces deux axes d'analyse pourraient en effet nourrir un système global de surveillance des arrêts de travail et les résultats observés permettent d'élaborer un bref guide de bonnes pratiques. Nous terminerons par une analyse des limites et perspectives.

Chapitre 2

Identifier les signaux d'arrêts maladie

Contenu

2.1 Introduction : analyser les arrêts maladie, un exercice complexe

2.1.1 Rappel de l'objectif

Dans ce chapitre, nous traiterons de l'analyse des données d'arrêt maladie à l'échelle individuelle pour identifier les processus qui mènent à ces arrêts. L'objectif est d'identifier, à partir de données administratives et de données d'enquête, des facteurs actionnables pour prévenir les arrêts maladie mais aussi des signaux faibles dans les trajectoires d'absence pour identifier les salariés les plus susceptibles de s'arrêter. En pratique, ces travaux pourraient permettre d'élaborer des plans d'action sur des populations soigneusement définies.

En introduction, nous tenterons d'identifier les obstacles méthodologiques et d'établir un état de l'art des méthodes statistiques appropriées pour l'analyse des arrêts maladie. Ensuite, nous étudierons une base d'enquête pour identifier les facteurs déterminants des arrêts maladie. Nous proposerons ensuite une analyse des trajectoires d'absence pour identifier des signaux faibles d'arrêts. Nous conclurons enfin par une discussion des travaux effectués en évaluant notamment la prédictibilité des arrêts de travail.

2.1.2 Les arrêts maladie : un phénomène complexe à analyser

L'analyse des arrêts maladie soulève de nombreux problèmes méthodologiques. Les arrêts maladie sont en effet un phénomène multidimensionnel car, à l'échelle du salarié, nous pouvons les définir selon deux critères :

1. Premièrement, les arrêts peuvent être définis selon leur durée. La durée de l'arrêt décrit sa *gravité*.
2. Deuxièmement, les arrêts peuvent être définis selon *leur fréquence*, c'est-à-dire le nombre d'arrêts de travail pendant une période donnée.

Ces deux dimensions sont importantes car elles peuvent apporter des informations différentes. Les arrêts longs sont plus graves et ont des déterminants plus marqués que les arrêts courts comme nous le montrerons plus tard. Un salarié avec une répétition d'arrêts maladie courts aura des conséquences plus négatives sur l'organisation du travail. Un effort préalable doit donc être effectué avant l'analyse afin d'identifier l'indicateur adéquat à la problématique.

Les données d'arrêt maladie, en fonction de l'indicateur choisi, présentent de plus de nombreuses

particularités statistiques. Si l'on analyse les arrêts maladie comme des données de comptage (nombre d'arrêts dans l'année par exemple), les données sont concentrées en zéro (*zero-inflated*) puisque seul 36% des salariés ont un arrêt dans l'année [19]. Lorsque l'on analyse les données d'arrêts comme des données de durée, ces dernières sont censurées puisque nous ne savons pas *a priori* la durée des tous les arrêts de travail lorsqu'ils commencent. Ces données sont aussi tronquées puisque nous ne savons pas le passé des salariés (avant son arrivée dans l'entreprise d'intérêt ou même, dans notre cas, avant que l'entreprise ait souscrit à un contrat de prévoyance). Cela pose problème puisque, comme nous le montrerons, les salariés ayant déjà eu des arrêts maladie ont une probabilité beaucoup plus forte d'en avoir à nouveau.

Les sources d'arrêts maladie sont de plus très hétérogènes. Comme nous l'avons illustré en Figure 1.1, les arrêts de travail peuvent être déterminés par des facteurs individuels : l'âge, le sexe, l'état de santé, la vie personnelle, *etc.* Les arrêts de travail peuvent être aussi déterminés par des facteurs liés à l'entreprise et au travail comme la pénibilité physique et mentale de l'emploi. Enfin, des déterminants exogènes ont un impact sur les arrêts de travail : maladie saisonnière (comme la grippe), phénomènes météorologique, grève, *etc.* Nous effectuerons une revue plus précise de ces déterminants dans une section ultérieure.

Une autre problématique est la corrélation très forte qui existe entre tous ces paramètres qui nous poseront des problèmes méthodologiques et qui rendent une analyse causale extrêmement délicate ;

Afin de résoudre ces problématiques, nous proposons ensuite une revue de littérature (*scoping review*) afin d'identifier les différents angles d'analyse statistique de ces données d'arrêts de travail.

2.1.3 Publication 1 : revue de littérature, PLOS One 2020

L'article présenté dans cette section a été publié par *PlosOne* le 15 septembre 2020 [98].

Résumé de l'article en français : L'identification des déterminants des arrêts maladie est un enjeu fort pour la société et pour les entreprises qui pourraient réduire les coûts associés mais aussi améliorer la qualité de vie des salariés. Ces déterminants sont très nombreux et leur identification des difficile à cause de la structure complexe de ces données. Le choix d'outils statistiques adéquats est ainsi un véritable défi.

Pour faciliter cet exercice, nous avons conduit une revue détaillée de la littérature selon la méthode

PRISMA de la *scoping review*. Notre objectif est d'identifier les différents axes d'analyse statistique des arrêts maladie. Nous avons lancé des requêtes dans des bases de données traitant de domaines divers (Medline, World of Science, Science Direct, Psycinfo et EconLit) afin d'identifier les publications utilisant des outils statistiques pour expliquer ou prédire les arrêts de travail à l'échelle du salarié.

Nous avons sélectionné 469 articles parmi les 5150 recueillis dans les bases de données, publiés entre 1981 et 2019. Au total, nous avons identifiés trois types d'analyse. Les principales analyses se concentre sur des indicateurs *univariés* des arrêts de travail et principalement sur des données de comptage. ces analyses représentent 438 articles. La deuxième catégorie analyse des données bivariées : durée et fréquence des arrêts de travail sont étudiées simultanément grâce à des modèles multi-états ou à des analyses trajectoires. Quatorze articles ont choisi cet axe d'analyse. Enfin, une dernière catégorie d'analyse utilise des techniques provenant de l'analyse causale comme les modèles à équations structurelles (22 articles).

Notre revue a montré que les arrêts maladie ont inspiré des méthodes très diverses d'analyse. La majorité de ces analyses ont étudié les arrêts de travail comme une donnée de comptage (soit comme la fréquence des arrêts, soit comme le nombre de jours d'arrêts pendant une période donnée) et des travaux exploratoires devraient être poursuivis pour comprendre simultanément la durée et la fréquence des arrêts de travail. La majorité des articles n'évaluait pas la pertinence de la méthodologie choisie qualitativement ou quantitativement : présenter un indicateur montrant la capacité d'explication ou de prédiction du modèle semble pourtant très important dans ce contexte où, le nombre de facteurs déterminants est très important. Les modèles évalués présentent d'ailleurs, comme attendu, des capacité d'explication assez faibles.

Modeling sickness absence data: a scoping review

Tom DUCHEMIN[1,2]*, Mounia N. HOCINE[1]

[1] Laboratoire Modélisation, Epidémiologie et Surveillance des Risques Sanitaires, Conservatoire national des arts et métiers, Paris, France.

[2] Malakoff Médéric Humanis, Paris, France.

*Corresponding author:

Tom DUCHEMIN,

tom.duchemin@cnam.fr (TD)

ABSTRACT

The identification of sick leave determinants could positively influence decision making to improve worker quality of life and to reduce consequently costs for society. Sick leave is a research topic of interest in economics, psychology, health and social behaviour. The question of choosing an appropriate statistical tool to analyse sick leave data can be challenging. In fact, sick leave data have a complex structure, characterized by two dimensions: frequency and duration, and involve numerous features related to individual and environmental factors. We conducted a scoping review to characterize statistical approaches to analyse sick leave data in order to synthesise key insights from the extensive literature, as well as to identify gaps in research. We followed the PRISMA methodology for scoping reviews and searched Medline, World of Science, Science Direct, Psycinfo and EconLit for publications using statistical modeling for explaining or predicting sick leave at the individual level. We selected 469 articles from the 5150 retrieved, dated from 1981 to 2019. In total, three types of model were identified: univariate outcome modeling using for the most part count models (438 articles), bivariate outcome modeling (14 articles), such as multistate models and structural equation modeling (22 articles). The review shows that there was a lack of evaluation of the models as predictive accuracy was only evaluated in 18 articles and the explanatory accuracy in 43 articles. Further research based on joint models could bring more insights on sick leave spells, considering both their frequency and duration.

1. INTRODUCTION

Understanding sick leave (SL) is a crucial issue for workers and their employers. Identification of determinants of sick leave could help decision makers set up appropriate prevention policies to improve the quality of life of workers and reduce costs for employers[1,2]. However, potential determinants are diverse, and their identification may be extremely difficult. SL may be related to both individual and professional environments [3,4], and indeed, the literature covering the topic is very wide-ranging. Studies can be found in journals of public health, epidemiology, sociology, economics, and psychology [5–8]. In addition, the modeling of SL is made difficult by certain features of the collected data. For instance, SL data can be zero-inflated, over-dispersed, censored, truncated, or highly seasonal, to give a few examples. Furthermore, SL is a two-dimensional variable characterised by both its frequency (number of SL spells over a given period) and its duration (length of a SL spell). Thus, the choice of appropriate statistical tools is very important.

Here, we document the state of the art on statistical methods for modeling SL data. This may help identify major trends and gaps in the scientific literature that could guide researchers towards better modeling. We pay careful attention to the afore-mentioned issues, summarize the results from the literature, and describe the best-adapted statistical approaches to deal with the properties of SL data. To proceed, we followed the *scoping review* methodology recently published by PRISMA[9]. While reviews on the determinants of SL have been previously published[4,10–13], this is, to our knowledge, the first review providing an overview of the various statistical tools that can be used to identify SL determinants.

2. METHODS

2.1. Literature search

We performed a scoping literature review to assess how SL is modeled. A scoping approach was preferred to a systematic approach in order to describe trends and practices in different fields of research, and formulate new research possibilities. The recently published *PRISMA* (Preferred

Reporting Items for Systematic reviews and Meta-Analysis) extension for scoping reviews *(PRISMA-ScR)*[9] was used.

To perform a comprehensive search, we systematically searched five databases from different fields which may deal in SL analyses: *MEDLINE, World of Science, Science Direct, PsycInfo* and *EconLit*. We formulated queries on three items: *(i)* model (keywords like *explaining*, *predicting*, *factor*, *determinant*, *risk*, *model*, and *classification*), *(ii)* absence (keywords like *sickness absence*, *absenteeism*, *sickness spell*, *sick leave* and *sick-leave*) and *(iii)* work (keywords like *working*, *worker*, *employee* and *adult*). The *MEDLINE* query was:

(predic$[Title/Abstract] OR risk$[Title/Abstract] OR classification$[Title/Abstract] OR regression$[Title/Abstract] OR explain$[Title/Abstract] OR determinant$[Title/Abstract] OR factor$[Title/Abstract] OR model$[Title/Abstract])

AND

(work$[Title/Abstract] OR employee$[Title/Abstract] OR adult$[Title/Abstract])

AND

(sickness absen$[Title/Abstract] OR sickness spell$[Title/Abstract] OR sick leave$[Title/Abstract] OR sick-leave$[Title/Abstract] OR absenteeism[Title/Abstract])

Searches were performed in September 2019 over the full databases.

The review was then performed in two steps. The first consisted in a review of titles and abstracts by a single reviewer (T.D.). Articles were included if they met the following criteria:

1. Original articles published in peer-reviewed journals (which excludes theses, book chapters, conference communications, reviews, etc.).
2. Involve statistical models describing any outcome related to the occurrence of SL at the individual level. Hence, any models describing aggregated data were not considered relevant.
3. Explicitly mention sickness absence in a working population as an outcome of a statistical model.

When a reference could not be included with certainty, its full text was obtained and a second screening was performed. This second step consisted of a review of the articles' contents by two reviewers (T.D. and M.H.).

The first step was performed by a single reviewer because the amount of retrieved material was overwhelming: this first step is therefore not infallible, and mistakes may have been made. However, we have tried to reduce these errors by eliminating only articles that seemed clearly inappropriate during this first step, and by keeping all items that may have caused any doubt.

For the sake of clarity, articles retrieved from the review are referenced with the prefix A in the *Results* section and can be found in the Electronic Supplementary Material. This document also provides all the results from this scoping review.

2.2. Resource extraction and analysis

The informations extracted from each included article consisted of:

1. *Metadata:* publication year, journal title, authors;
2. *Dataset characteristics:* study population, country of the study, first year of the study;
3. *Statistical methods:* statistical model, definition of the outcome, criteria for the evaluation of the models, predictive ability of the model (AUC or C-Index), consideration of the multilevel nature of SL data.

We only retained the value of the evaluation criterion for predictive models because the AUC and C-Index are comparable; even if the statistical methodologies are different. Note that explicative criteria such as R^2 are dependent on the model and are thus not relevant for model comparison in our scoping review, where we describe different methods. Finally, descriptive analyses were performed on the extracted data to examine frequencies and trends of statistical approaches in the literature.

3. RESULTS

3.1. Study selection

A total of 5150 articles were retrieved from *MEDLINE*, 226 from *Science Direct*, 419 from *World of Science*, 173 from *PsycInfo*, and 15 from *EconLit*. After screening abstracts and titles, we excluded 3536 publications that did not meet our inclusion criteria. By examining the remaining 877 in detail, a further 416 were excluded. Thus, a total of 469 articles were retained for the scoping literature review. Figure 1 shows the PRISMA flow diagram for the study's selection method.

Fig.1 - PRISMA diagram of the selection process for study inclusion in the review.

The first article modeling SL identified in our selection was published in 1981 [A358]. While SL articles go back to at least 1958[14], the earliest ones only used descriptive tools, not statistical models. Figure 2 shows the number of published peer review articles retained per year: very few articles modeling SL with statistical tools were published before 1998. Since then, the number has tended to increase over time.

Fig.2 - Number of articles included in the scoping literature review from each year

Retained articles were mainly published in occupational health journals, as shown in Table 1. However, as SL is an interdisciplinary research theme, journals in the fields of public health, economics, social science, general medicine, psychology, and generalist journals, are also represented. We note that almost half of the databases used for analyses comes from Scandinavian countries, with 231 articles publishing studies on databases from Finland, Denmark, Sweden, and Norway. Databases from the Netherlands and the US are analyzed in 64 and 28 publications, respectively. Databases from remaining countries correspond to less than 20 articles per country.

Journal	**Number (%)**
Journal of Occupational and Environmental Medicine	38 (9,5%)
Occupational and Environmental Medicine	33 (8,25%)
International Archives of Occupational and Environmental Health	26 (6,5%)
Scandinavian Journal of Work, Environment & Health	26 (6,5%)
BMC public health	24 (6%)
Scandinavian Journal of Public Health	21 (5,25%)
European Journal of Public Health	19 (4,75%)
Occupational Medicine (Oxford, England)	19 (4,75%)
Social Science & Medicine	11 (2,75%)
Journal of Occupational Rehabilitation	12 (3%)

Table 1 - Number of retrieved articles per peer reviewed journal.

Retrieved publications used two different kinds of data source for their analyses: *(i)* administrative databases or registers describing certified SL of workers, and *(ii)* questionnaires describing declared SL. Across the 469 studies, the formers are found in 346 studies, the latter in 421. In 302 articles, register and questionnaire were linked together. One article also used data from a meta-analysis.

A total of 425 publications focused on all-cause SL, while the remaining few focused on precise causes of SL. Among those, 11 with low back pain-related SL, 12 dealt with musculoskeletal disorder-related SL, 12 with depressive or mental disorder-related SL, and 10 with other cause-specific SL such as voice problems, work-related SL, or respiratory complaints. The retrieved articles many different populations with specific job, specific pathologies or more generally specific characteristics. The most studied population is a general population without specific characteristics (231 articles). A very large number of articles also study healthcare workers (75 articles) and employees in the public sector, healthcare workers excluded (43 articles).

3.2. Statistical methods for modeling sick leave

Three main categories of statistical approach for modeling SL arise from the results of the review, as illustrated in Table 2:

1. *Modeling a univariate SL outcome:* based on regression models, this allows researchers to evaluate the effect of potential predictors. In this approach, SL data of various forms are considered, involving a number of statistical approaches, as described below. A total of 372 publications included at least one of those methods:
    a. *Count data:* SL data can be defined as the number of days (or hours) of absence, or the number of SL spells during a given time interval. To model count data, the authors use Poisson models (84 articles,) [A1-A84], negative binomial models (50 articles) [A1-A7,A85-A121], zero-inflated negative binomial models (9 articles) [A120-A128], hurdle models (10 articles) [A117-A120,A129-A132], and zero-inflated Poisson models (2 articles) [A128,A133]. These methods are described in 141 articles of those retained in the present review.

b. *Time-to-event data:* SL data can be defined in terms of time to subsequent SL. This approach is used in 86 publications. Of these, 77 use Cox models [A17,A26,A120-A208]. Four publications use models derived from Cox models: the Andersen-Gill model [A209,A210], frailty modeling [A67], and competing risk models [A211]. Other publications use parametric models [A46,A212,A213] or other nonparametric models [A214,A215].

c. *Categorical data:* SL data can be defined in terms of experiencing at least one SL spell, or another predefined categorical SL event, over a given time interval. SL is described using a binary outcome in 185 publications, and with a multiple category outcome in 24 publications. For binary SL outcomes, 181 models use a logit link function [A9,A80,A90,A96,A102,A149,A216-A376], 3 a probit link function [A6,A376,A377], and 2 use GLM with an unspecified canonical link [A378,A379]. For multinomial SL outcomes, 15 publications use multinomial logistic regression [A188,A232,A380-A389] and 8 use ordinal logistic regression [A391-A397].

d. *Continuous indicators:* SL data can also be defined as the number of days, spells, or hours, the sickness absence rate, the average duration of spells, etc. A total of 57 publications involve a continuous SL indicator. Amon them, 51 use linear regression [A4,A20,A65,A90,A98,A108,A177,A331,A398-A434], 2 use joinpoint regression [A435,A436], 1 uses Tobit regression [A437], and 2 use GLM with an unspecified link function [A4,A438]. Another article uses a deep learning method, a fuzzy network model, and compares it to machine learning method [A439].

2. *Modeling a bivariate SL outcome:* based on joint models, this strategy allows researchers to evaluate the effect of potential predictors on both SL duration and frequency. Seventeen publications use a joint approach, modeling simultaneously the duration and frequency of SL. Eight of these use multi-state modeling [A189,A440-A445], while the nine others use trajectory analysis [A13,A42,A385,A446-A449].

3. *Structural equation modeling (SEM):* based on a factorial analysis and a linear regression model, this allows testing, without quantifying, the causal association between predefined latent variables and SL. In this approach, SL is defined as a continuous outcome: frequency, duration, or a combination of both. A total of 23 publications use this approach [A277,A416,A450-A469], mostly in journals involving qualitative research such as *the Journal of Organizational Behavior* [A458-A461].

<table>
<tr><th colspan="3">Outcome of interest</th><th>Model</th></tr>
<tr><td rowspan="4">Modeling a univariate SL outcome (n = 438)</td><td colspan="2">Count data (n = 132)</td><td>- Poisson (84)
- Negative binomial (44)
- Zero-inflated negative binomial (9)
- Hurdle model (8)
- Zero-inflated Poisson (2)</td></tr>
<tr><td colspan="2">Time-to-event (n = 86)</td><td>- Cox (77)
- Andersen-Gill (2)
- Frailty model (1)
- Competing risk model (1)
- Other parametric (3) or nonparametric (2)</td></tr>
<tr><td rowspan="2">Categorical data (n = 190)</td><td>Binary (n = 171)</td><td>- Logit (167)
- Probit (3)
- GLM with unspecified link (2)</td></tr>
<tr><td>Multinomial (n = 20)</td><td>- Multinomial logistic (12)
- Ordinal logistic (8)</td></tr>
<tr><td colspan="2">Rate/Continuous indicator (n = 57)</td><td>- Linear regression (45)
- Joinpoint (2)
- Tobit (1)
- Fuzzy network and machine learning methods (1)
- GLM with unspecified link (2)</td></tr>
<tr><td colspan="3">Modeling a bivariate SL outcome (n = 14)</td><td>- Multistate model (7)
- Trajectory analysis (7)</td></tr>
<tr><td colspan="4">Structural equation modeling (n = 22)</td></tr>
</table>

Table 2 - Methods retrieved from the scoping review.

We finally found that 52 articles evaluate the accuracy of these models in terms of predictability or explainability. Among those 5752 18 evaluated the predictive capacity of their models using AUC or the C-Index. Obtained AUCs ranged from 0.53 to 0.88. Further, 43 publications among those 57 proposed an evaluation of the models using R^2 or pseudo R^2 for regression models, or with specific criteria for conceptual models (e.g., AGFI, RMSEA, IFI, NNFI, CFI). Moreover, 51 of the 469 articles

took into account the multilevel nature of SL data by using hierarchical modeling or by including mixed effects to describe both individual and work levels.

4. DISCUSSION

This scoping review has highlighted the increase in statistical modeling of sick leave data over the previous 40 years. We identified different statistical approaches which could be gathered into three categories: models for univariate SL outcome, models for bivariate SL outcome, and structural equation modeling. Logistic regression and count data models are the most popular approaches: logistic regression gives intuitive results thank to odd ratios when the problem involves a binary outcome; count data models are also often used to analyse variables such as number of day of SL or number of spells per year. Most of the time, the publications focus on non-specific population, but many articles focus on the case of healthcare workers and civil servants, who are very affected by sickness absenteeism. The way outcomes are encoded can vary greatly (occurrence of a spell, frequency, duration, etc.). Finally, the predictive performance of SL models also varies a lot from one publication to the next: when provided, the AUC and C-Index values range from 0.53 to 0.88.

This review was faced with the difficulty of dealing with the huge number of publications that exist on SL. We have tried to be as exhaustive as possible by processing data from very numerous sources and we retrieved a huge amount of publications. Consequently, the first step of the review was carried out by a single reviewer and we may have failed to identify all relevant articles during this first step of the review. Finally, we chose to perform a scoping review but a systematic review of the literature could have been interesting and would have more comprehensibly assessed the quality of suggested models. However, the simultaneous evaluation of so many different models would have been overly complex, and a scoping review approach seems more appropriate.

We would particularly like to point out the lack of information provided to assess the predictability and explainability of models on SL data. First, very few articles evaluate the accuracy of their models in terms of explainability and predictability. This is however crucial, as SL is a phenomenon with many

potential determinants: accuracy criteria could help researchers focus on determinants that are more relevant than other. Second, many of the accuracy scores obtained are fairly poor, as mentioned earlier. A possible reason could be the fact that SL is likely related to many determinants, but most studies investigate them partially. A systematic review of the determinants of SL would be very valuable: the SL literature is abundant but nevertheless, the mechanisms of sick leave remain unclear. This said, such an exercise would be complex since, as explained above, definition of outcomes can be very different. Indeed, distinct reviews would be needed for frequency and duration analyses.

Another point to improve the predictability of sick leave is data collection. In particular, an article retrieved in this review uses neural networks to predict sick leave and its results seem encouraging [A439]. These models work well for massive data set, both in terms of number of observations and of covariates. In practice, these data are very rare and difficult to collect: a more systematic collection and analysis of absence data could lead to a better understanding of the phenomenon and a more complete evaluation of these machine learning or deep learning models.

Another way to improve the explainability of SL would be to investigate alternative statistical approaches. In addition to a systematic review on determinants, new methods for hierarchizing SL determinants could be helpful in providing new insights on them. By introducing a great number of heterogeneous variables into a model, methods such as random forests might more easily identify the most important variables[15]. Such methods may be appropriate for the study of sick leave because the possible determinants of SL are numerous and these methods, running without linearity hypotheses, may potentially better explain SL. Those models could also take easily into account the multilevel nature of SL data by implicitly testing all interactions. Moreover, only 19 articles investigated causal relationships between covariates and SL using structural equation modeling, despite this approach seeming rather informative for the study of SL. Indeed, since the most commonly used regression methods described in this scoping review give fairly poor explanatory results, structural equation modeling could be used to effectively and intuitively test the links between different variables. Bayesian network models, none of which were identified in this review, might be an appropriate and powerful

solution for causal inference[16,17]. Furthermore, models for bivariate outcomes, and in particular multi-state models, also appear to have been overlooked in previous research. Indeed, most of the retrieved articles deal with SL in a single dimension (duration or frequency), yet both are of interest and potentially correlated. A joint study of the two could make it possible to improve the explanatory accuracy of models, while also helping to better understand the causes and features of SL.

FUNDING

The Ph.D. works of T.D. are funded by a 3-year grant from the Association Nationale de la Recherche et de la Technologie (grant 2017/1517) and by Malakoff médéric humanis.

ELECTRONIC SUPPLEMENTARY MATERIAL

Supplementary references (indexed by A in the text) and data from the scoping review are available at 10.6084/m9.figshare.9741203.

REFERENCES

1. Koopmanschap, M. A., Rutten, F. F., van Ineveld, B. M. & van Roijen, L. The friction cost method for measuring indirect costs of disease. *J Health Econ* **14**, 171–189 (1995).
2. van den Berg, S., Burdorf, A. & Robroek, S. J. W. Associations between common diseases and work ability and sick leave among health care workers. *International Archives of Occupational and Environmental Health* **90**, 685–693 (2017).
3. Roskes, K., Donders, N. & van der Gulden, J. Health-related and work-related aspects associated with sick leave: a comparison of chronically ill and non-chronically ill workers. *International Archives of Occupational and Environmental Health* **78**, 270–278 (2005).
4. Beemsterboer, W., Stewart, R., Groothoff, J. & Nijhuis, F. A literature review on sick leave determinants (1984-2004). *Int J Occup Med Environ Health* **22**, 169–179 (2009).
5. Kröger, H. The stratifying role of job level for sickness absence and the moderating role of gender and occupational gender composition. *Social Science & Medicine* **186**, 1–9 (2017).

6. Kok, A. A. L., Plaisier, I., Smit, J. H. & Penninx, B. W. J. H. The impact of conscientiousness, mastery, and work circumstances on subsequent absenteeism in employees with and without affective disorders. *BMC Psychology* **5**, (2017).
7. Barmby, T. Worker absenteeism: a discrete hazard model with bivariate heterogeneity. *Labour Economics* **9**, 469–476 (2002).
8. van Drongelen, A., Boot, C. R. L., Hlobil, H., van der Beek, A. J. & Smid, T. Cumulative exposure to shift work and sickness absence: associations in a five-year historic cohort. *BMC Public Health* **17**, (2017).
9. Tricco, A. C. *et al.* PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Ann Intern Med* **169**, 467 (2018).
10. Alexanderson, K. Sickness absence: a review of performed studies with focused on levels of exposures and theories utilized. *Scandinavian Journal of Social Medicine* **26**, 241–249 (1998).
11. Bekker, M. H. J., Rutte, C. G. & Rijswijk, K. van. Sickness absence: A gender-focused review. *Psychology, Health & Medicine* **14**, 405–418 (2009).
12. de Vries, H., Fishta, A., Weikert, B., Rodriguez Sanchez, A. & Wegewitz, U. Determinants of Sickness Absence and Return to Work Among Employees with Common Mental Disorders: A Scoping Review. *J Occup Rehabil* **28**, 393–417 (2018).
13. Werner, E. L. & Cote, P. Low back pain and determinants of sickness absence. *Eur J Gen Pract* **15**, 74–79 (2009).
14. Segal Halperin, B. M. [Findings on common grippe as a factor in worker's absenteeism]. *Med Panam* **10**, 111–117 (1958).
15. Duchemin, T., Bar-Hen, A., Lounissi, R., Dab, W. & Hocine, M. N. Hierarchizing Determinants of Sick Leave: Insights From a Survey on Health and Well-being at the Workplace. *Journal of Occupational and Environmental Medicine* **61**, 1 (2019).
16. Mohammadfam, I., Ghasemi, F., Kalatpour, O. & Moghimbeigi, A. Constructing a Bayesian network model for improving safety behavior of employees at workplaces. *Appl Ergon* **58**, 35–47 (2017).

17. Chan, A. P. C., Wong, F. K. W., Hon, C. K. H. & Choi, T. N. Y. A Bayesian Network Model for Reducing Accident Rates of Electrical and Mechanical (E&M) Work. *Int J Environ Res Public Health* **15**, (2018).

2.2 Identifier les déterminants des arrêts maladie

Ce panorama des méthodes d'analyse nous a permis d'identifier plusieurs angles d'étude mais aussi d'identifier certains manques dans la littérature. Dans cette nouvelle section, nous allons nous aussi nous pencher sur l'analyse des déterminants des arrêts maladie.

Dans un premier temps, nous ferons un bref état de l'art des déterminants des arrêts de travail en raffinant le graphique en Figure 1.1. Dans un second temps, nous présenterons les données d'enquête utilisées pour les analyses : le Baromètre Bien-Être et Santé au Travail. Nous présenterons ensuite une analyse de ces données. Cette analyse a été publiée en 2019 au *Journal of Occupational and Environmental Medicine* [19] et a pour objectif de proposer une hiérarchisation des facteurs déterminants des arrêts de travail en fonction de leur durée. Cette analyse sera précédée de travaux préliminaires raffinant l'objectif et orientant les choix méthodologiques de l'article.

2.2.1 Etat de l'art : déterminants des arrêts maladie

Comme nous l'avons montré dans la section précédente, l'analyse des arrêts maladie a donné lieu à de nombreuses publications donc l'objectif est d'identifier les déterminants de ce phénomène. Avant de proposer notre propre analyse appliquée aux salariés du secteur privé français, effectuons une brève revue des déterminants identifiés par la littérature.

2.2.1.1 Préambule à l'état de l'art

Avant de présenter cette revue, il est primordial de présenter les difficultés et limites de cet exercice. Premièrement, comme nous l'avons discuté précédemment, les arrêts maladie sont un phénomène multidimensionnel et les déterminants peuvent même être différents en fonction de l'indicateur étudié : durée ou fréquence. Les déterminants peuvent ensuite varier en fonction de la population étudiée : les déterminants des arrêts de travail ne sont pas les mêmes pour des salariés qui sont cadres et travaillent derrière un bureau et pour des salariés qui sont ouvriers et ont un travail manuel. Beaucoup d'études sont ainsi présentées pour des populations particulières. Notamment, l'étude des arrêts maladie des infirmières a donné lieu à une littérature très vaste et à quelques revues systématiques [20, 21, 22]. De même, les arrêts maladie ont parfois des motifs explicites différents pouvant être identifiés et des analyses peuvent être effectués pour certains types d'arrêts qui ont soulevé plus d'intérêt que d'autres

dans la littérature comme les arrêts pour lombalgie [23] ou pour des problèmes de santé mentale [24]. Une dernière difficulté est que ces déterminants sont très nombreux.

Quelques revues de littérature (non-systématique) ont tenté d'identifier les déterminants clés des arrêts maladie en ne distinguant pas les indicateurs d'absentéisme (tout type d'absence, arrêt de travail compris) [25] ou en distinguant durée et fréquence [8]. Nous allons tenter, en nous inspirant de ces travaux et de notre *scoping review* de présenter une synthèse très succincte des déterminants de ces arrêts. Nous ne ferons la distinction entre fréquence et durée que quand cela sera nécessaire.

2.2.1.2 Déterminants individuels

Une première catégorie de déterminants sont les déterminants liés directement au salariés, les déterminants que nous appellerons *individuels*. L'hygiène de vie est un de ces déterminants : les fumeurs, les salariés en surpoids, qui font peu d'activité physique ou qui ont une consommation excessive d'alcool ont des arrêts plus longs et plus fréquents [7, 26]. Nous pouvons aussi y trouver la santé physique et mentale qui influencent aussi bien la durée que la fréquence des arrêts [6, 27, 28], tout comme comme la satisfaction vis-à-vis de sa vie personnelle [29]. Parmi les déterminants individuels, il y a aussi des déterminants sociodémographiques comme l'âge ou le sexe. Les femmes ont notamment un plus grand risque d'arrêts longs et fréquent [8]. Les salariés les plus âgés ont un risque d'arrêts plus longs mais ont moins de risque d'arrêts fréquents [8, 9]. Nous avons cependant montré dans l'article que nous présenterons plus tard [19] que, lorsque nous contrôlons sur la santé, l'âge ne semble plus être un véritable déterminant.

2.2.1.3 Déterminants professionnels

Une autre catégorie de déterminants sont les déterminants liés à l'entreprise et au contexte professionnel. Parmi ces déterminants se trouve l'exposition à des facteurs de risque psychosociaux qui peuvent être définis comme les facteurs de "stress professionnel" ou comme "les situations ou évènements, potentiellement traumatisant sur le plan psychique, survenus au cours du travail" [12]. Avoir de mauvais rapports avec ses supérieurs ou avec ses collègues, avoir des altercations tendues avec des clients ou faire face à des échéances trop strictes sont des facteurs de risques psychosociaux. Ces facteurs semblent agir principalement sur la durée des arrêts. Aagestad *et al* (2014) [30] a montré que 15% des arrêts de plus de de 40 jours peuvent être attribués à l'exposition à ces facteurs de risque ;

Aagestad *et al.* (2014) a montré dans une autre publication [31] que 13% des arrêts de plus de 21 jours des femmes travaillant dans la santé et le social peuvent être attribués à de la violence ou des menaces de violence. D'autres facteurs de déterminants professionnels sont les facteurs de pénibilité physique comme le fait d'avoir une mauvaise position ou d'effectuer des gestes répétitifs au travail [10, 11].

Plus généralement, les déterminants professionnelles proviennent des caractéristiques de l'emploi. Un salarié ayant un emploi manuel a plus de risques d'arrêts de travail qu'un salarié ayant un emploi de bureau : Melchior *et al.* (2005) [32] a montré qu'environ 20% des arrêts de travail peuvent être attribués aux caractéristiques de l'emploi.

Cette distinction cause individuelle/cause professionnelle est certainement artificielle et simpliste mais elle permet une catégorisation plus aisée des causes d'arrêts maladie. Par exemple, bien qui nous classions ce déterminant dans la catégorie "cause individuelle", une mauvaise santé physique ou mentale peut être causée par des difficultés professionnelles autant que par des difficultés personnelles. Autre exemple, la difficulté à concilier sa vie professionnelle et sa vie personnelle est causé par l'interaction entre un contexte professionnel tendu (horaires non adaptés par exemple) et un contexte personnel difficile (être aidant par exemple, c'est-à-dire devoir s'occuper d'un proche en situation de handicap).

2.2.1.4 Déterminants exogènes

Une dernière catégorie de déterminants des arrêts maladie est la catégorie des déterminants exogènes. Ces déterminants sont des évènements extérieurs qui peuvent influencer la prise d'arrêts maladie. Parmi ces déterminants, nous pouvons inclure les épidémies (comme les épidémies de grippe [13] ou de gastro-entérite [33]), les phénomènes météorologiques [34] ou la législation [14]. Notamment, Pollak (2017) a montré que l'introduction de trois jours de carence en France n'a pas diminué la fréquence des arrêts maladie mais a diminué la durée des arrêts maladie d'environ 3 jours.

Une synthèse des déterminants des arrêts de travail est présenté en Table 2.1.

2.2.2 Les données : le Baromètre Bien-Être et Santé au Travail

2.2.2.1 Présentation des données

Le Baromètre Bien-Être et Santé au Travail (BST) est une enquête annuelle dont l'objectif est d'évaluer année après année, comme son nom l'indique, le bien-être et la santé au travail des salariés.

Déterminants personnels	- Santé mentale et physique - Hygiène de vie - Equilibre de la vie personnelle - Critères sociodémographiques
Déterminants professionnels	- Exposition à des facteurs de risques psychosociaux - Pénibilité physique - Activité professionnelle
Déterminants exogènes	- Epidémies - Phénomènes météréologiques - Grèves

TABLE 2.1 – Synthèse des déterminants des arrêts maladie

Cette enquête est reconduite chaque année depuis 2009 et interroge tous les ans environ 3500 salariés représentatifs du secteur privé français : chaque année, la plupart des salariés interrogés sont différents de ceux de l'année passée (plus de 90%).

Cette enquête est mise en place par Malakoff Humanis et est distribuée par l'IFOP, un institut d'étude. Les personnes interrogées proviennent d'un *Access Panel*, c'est-à-dire qu'ils se sont préalablement inscrits volontairement dans une base de répondant sans savoir *a priori* les thèmes abordés par les questionnaires qui leur seront distribués. L'enquête est représentative en termes de genre, d'âge (moins de 30 ans/30-39 ans/40-49 ans/50 ans et plus), de statut (cadre/techniciens-agents de maîtrise/employés/ouvriers), de taille d'entreprise (moins de 20 salariés/20 à 49/60 à 499/499 et plus), de secteur (industrie/tertiaire) et de région.

Le BST contient, chaque année, une centaine de questions décrivant les caractéristiques sociodémographiques des salariés, leur état de santé, leurs conditions de travail et leurs arrêts maladie.

Les questions de santé évaluent l'état de santé général de santé ("Comment jugez-vous votre état de santé en général ?") et évaluent plus précisément la santé physique et la santé mentale des salariés. La santé physique est évaluée par une série de questions sur les douleurs physiques ("Au cours des 12 derniers mois avez-vous ressenti les difficultés suivantes (hors accident traumatique) ?" avec une déclinaison de différents endroits du corps comme le dos, la tête, *etc.*) et la santé mentale est évaluée par diverses questions sur la prise d'anxiolytiques, les troubles du sommeil ou un manque de tonus.

Les questions concernant les conditions de travail concernent en un premier lieu la pénibilité physique de l'emploi qui est évalué par divers items comme le fait de porter des charges lourdes, de rester en position pénible ou d'effectuer des gestes répétitifs. L'exposition aux facteurs de risques psychoso-

ciaux est aussi évaluée. Ces facteurs sont notamment évalués par des items interrogeant le salarié sur ses rapports aux supérieurs hiérarchiques ou avec ses collègues ou par des items évaluant la pression psychologique subie au travail.

Enfin, des questions évaluent les arrêts maladie des salariés : il est demandé aux salariés s'ils ont eu au moins un arrêt maladie dans l'année et, si oui, de préciser la durée des arrêts. Une liste de durée est proposée : arrêt de moins de 3 jours, arrêt de 3 à 5 jours, arrêt de plus d'une semaine, arrêt de plus d'un mois. A partir de la cinquième vague du questionnaire (2014), une question relative aux arrêts de plus de 3 mois a été ajoutée au questionnaire. La première vague de l'enquête est exclue de l'analyse car les questions relatives aux arrêts de travail n'étaient pas posées dans les mêmes termes.

L'objectif du questionnaire est d'avoir le rôle d'un baromètre, soit d'évaluer années après annése différents indicateurs et de permettre aux entreprises d'évaluer leur propre situation par comparaison. Le cœur du questionnaire reste ainsi inchangé vagues après vagues mais certaines questions plus conjoncturelles peuvent être ajoutées. Notamment, des questions relatives au télétravail ont été récemment ajoutée à l'étude afin d'évaluer le ressenti des salariés.

Le BST est une enquête représentative du secteur privé français mais il est aussi proposé aux entreprises qui peuvent évaluer ces différents items auprès de leurs salariés. Nous discuterons de l'utilisation de ce questionnaire en pratique dans des sections ultérieures.

2.2.2.2 Description des données

L'ensemble des questions retenues pour les analyses et les statistiques associées peuvent être lues en annexe A : les questions retenues sont les questions présentes dans l'ensemble des vagues du questionnaire et une pré-sélection de questions qui décrivent des facteurs d'ajustement et des facteurs qui pourraient donner lieu à des plans d'action.

Sociodémographie La Table 2.2 décrit les caractéristiques des individus interrogés dans les 8 vagues utilisées pour l'analyse qui sera présentée ensuite. L'analyse repose sur un échantillon de 32 327 salariés représentatifs de la population française.

Indicateurs d'arrêts de travail La Table 2.2 décrit aussi les indicateurs d'arrêts maladie recueillis dans le questionnaire. Chaque individu est classé dans une unique catégorie : la catégorie relative à

	Population générale	**Pas d'arrêt**	**Arrêt de moins de 3 jours**	**Arrêt de 3 jours à 1 mois**	**Arrêt de plus d'un mois**
n	32 327	20 665	4 128	5 187	2 347
Sexe					
Femme	46%	44%	50%	50%	51%
Homme	54%	56%	50%	50%	49%
Age (année)					
<30	21%	20%	29%	20%	14%
30-39	30%	29%	36%	31%	25%
40-44	15%	15%	14%	15%	16%
45-49	13%	14%	8%	13%	14%
50-55	14%	14%	9%	14%	20%
>55	8%	8%	5%	7%	12%
Taille de l'entreprise					
Plus de 1000 salariés	39%	37%	42%	40%	41%
50 à 999 salariés	24%	23%	25%	28%	28%
10 à 49 salariés	12%	13%	12%	12%	12%
Moins de 10	25%	27%	20%	19%	19%
Secteurs d'activité					
Bureau d'étude et ingéniérie	5%	5%	7%	3%	2%
BTP	8%	8%	8%	9%	9%
Commerce	16%	16%	17%	17%	17%
Industrie	25%	24%	24%	29%	28%
Santé et Social	9%	9%	8%	9%	11%
Services	23%	24%	23%	20%	20%
Transport, télécom et industrie	14%	14%	13%	13%	13%

TABLE 2.2 – Description des données sociodémographiques du Baromètre Santé et bien-être au Travail, stratifiées par arrêt de travail dans les douze mois précédant la réponse au questionnaire

l'arrêt de le plus long durant l'année précédant la réponse au questionnaire.

Environ 36,1% des salariés interrogés ont eu un arrêt maladie durant l'année. La plupart des arrêts sont des arrêts courts de moins d'un mois : 12,8% des salariés ont des arrêts de moins de 3 jours et 16,0% des arrêts de 4 jours à 1 mois. 7,3% des salariés ont un arrêt de plus d'un mois dans l'année, ce qui représente 20,2% des salariés ayant eu un arrêt de travail.

Les hommes ont moins d'arrêt maladie que les femmes et ceci indépendamment de la durée des arrêts. Les salariés de moins de 30 ans sont plus représentés parmi les salariés ayant eu un arrêt de moins de 3 jours (ils sont 29% dans cette catégorie et seulement 21% parmi la population générale) et les salariés de plus de 55 ans sont sur-représentés parmi les salariés ayant eu un arrêt de plus d'un mois (12% contre 8% dans la population générale).

Les salariés des entreprises de moins de 10 salariés sont sous-représentés parmi les salariés ayant des arrêts très courts (20% contre 25% dans la population générale) alors que les salariés d'entreprise de plus de 1000 salariés sont sur-représentés dans cette catégorie (42% contre 39%).

Qualité des données Du fait de son mode d'administration, les données du BST présentent un taux de non-réponse nul, ce qui facilite grandement les analyses. Le grand nombre de répondant et la représentativité des données est un autre grand avantage de ces données.

Certaines limites sont tout de même à noter. La première est liée au mode d'administration du questionnaire : l'utilisation d'un panel d'institut de sondage et la transmission électronique de l'enquête peuvent induire des biais. Les salariés ayant répondu doivent avoir accès à du matériel informatique et rien n'assure que les individus s'inscrivant à des panels de sondage soient représentatifs du reste de la population. Une deuxième limite est liée au caractère déclaratif des données. Cela n'est pas problématique pour la plupart des items du questionnaire puisque la déclaration est souvent la seule manière d'évaluer certaines variables. Cela est cependant problématique pour la déclaration des arrêts maladie : il est demandé aux salariés s'ils ont eu des arrêts dans l'année et, si oui, de préciser grossièrement leur durée. Il n'est pas évident qu'un salarié se rappelle, un an après, de l'occurrence d'un arrêt court et même de sa durée (a-t-il duré 3 jours ou 5 jours ?).

2.2.3 Hiérarchisation les déterminants des arrêts maladie

Les données présentées ci-dessus sont le support d'une analyse des déterminants des arrêts maladie que nous avons menée. Ces analyses ont donné lieu à une publication au *Journal of Occupational and Environmental Medicine* [19] que nous allons présenter. Nous décrirons tout d'abord l'objectif puis donnerons quelques éléments complémentaires pour justifier notre choix de méthode. Après avoir présenté l'article, nous apporterons quelques éléments de discussion supplémentaire.

2.2.3.1 Objectif

La littérature des déterminants des arrêts maladie est déjà très dense et notamment d'articles avec des niveaux de preuve assez élevés (grandes cohortes [35, 36] et même expériences randomisées [37, 38]). Notre objectif n'est pas d'ajouter un autre article à la littérature mais de tenter un tri parmi ces déterminants.

L'intérêt de notre enquête est qu'elle intègre un grand nombre de variables qui décrivent beaucoup de déterminants pouvant entrer dans chacune des catégories décrites dans une précédente section (à l'exception des déterminants exogènes). A notre connaissance, aucune étude n'analyse tous ces déterminants simultanément. Une telle analyse pourrait permettre de mettre en perspective toutes ces variables et de comprendre quelle est l'importance de ces différents déterminants dans l'occurrence d'arrêts maladie.

Cette grande quantité de variable pourrait aussi permettre d'évaluer la *prédictibilité* des arrêts de travail. Autrement dit, l'occurrence d'arrêt maladie est un événement très hétérogène d'un salarié à un autre [39] et un salarié malade pourra très bien avoir un arrêt maladie dans une situation mais pas dans une autre. Ainsi, la prédiction des arrêts de travail est un travail très complexe et, même si l'on contrôle sur les différentes catégories décrites ci-dessus, une part restera inexplicable. Comme nous pourrons ajuster sur la plupart des catégories décrites, la part d'inexpliqué de nos modèles peut s'interpréter comme cette hétérogénéité des arrêts maladie (et comme cette part provenant des causes exogènes, que nous n'avons pas dans le questionnaire).

Notre objectif est donc premièrement de hiérarchiser les déterminants des arrêts maladie et cette hiérarchisation permettra d'évaluer l'hétérogénéité dans l'occurrence de ces arrêts. La définition des indicateurs d'arrêts est contrainte par le questionnaire et nous allons hiérarchiser ces déterminants en

fonction de leur durée. Nous étudierons ainsi les arrêts de moins de 3 jours, les arrêts de 4 jours à 1 mois et les arrêts de plus d'un mois.

2.2.3.2 Méthode : pourquoi les forêts aléatoires ?

Avant de passer à l'article, il nous semble important de présenter quelques points relatifs à la méthode.

Afin d'expliquer l'occurrence d'arrêts de travail par nos divers déterminants, nous penchions au départ pour une méthode assez classique de régression logistique qui nous aurait permis de récupérer des coefficients faciles à interpréter. Cependant, cette méthode nous a paru très rapidement insuffisante. L'analyse est en effet réalisée sur un ensemble de 51 variables qui sont assez corrélées les unes aux autres : par exemple, de nombreux items se concentrent sur la santé des salariés et, souvent, une mauvaise santé physique va de paire avec une mauvaise santé mentale.

L'introduction d'autant de variables n'est pas souhaitable pour un tel modèle de régression et la réalisation d'une étape supplémentaire de sélection de variable n'est pas vraiment compatible avec notre objectif. L'étape de sélection de variable nous aurait forcé à sélectionner uniquement un groupe restreint de variables qui, *a priori*, ne décrit pas nécessairement l'ensemble des catégories de déterminants. De plus, la forte corrélation entre les variables et les interactions possibles entre ces dernières impliquent une forte instabilité dans la sélection des variables [40] que nous avons observé en pratique.

Afin de pallier ces problèmes, nous avons cherché une méthode permettant d'intégrer un grand nombre de variables, interagissant potentiellement entre elles, et permettant tout de même d'évaluer l'importance de chacune des variables dans le processus d'occurrence des arrêts de travail. Nous avons finalement décidé d'utiliser une méthode de forêts aléatoires conditionnelles [41, 42]. Les forêts aléatoires conditionnelles permettent d'attribuer à chaque salarié un score évaluent leur probabilité d'avoir un arrêt de travail en combinant les résultats de plusieurs arbres de décision. Cette méthode ne permet pas une interprétation précise de l'impact des différents coefficients dans la construction du score mais permet de prendre en compte les interactions (comme dans les arbres de décision) et permet aussi d'évaluer l'importance de chaque variable dans la construction du score.

2.2.3.3 Publication 2 : hiérarchiser les déterminants des arrêts maladie, JOEM 2019

Résumé de l'article en français : Nous avons hiérarchisé un ensemble de facteurs individuels et professionnels déterminant la prise d'arrêts maladie très courts (moins de trois jours), courts (de 4 jours à un mois) et longs (plus d'un mois).

Les données proviennent du BST, une enquête annuelle conduite auprès du secteur privé français entre 2011 et 2017. Nous avons hiérarchisé 51 facteurs déterminants des arrêts maladie en utilisant une approche de forêt aléatoire conditionnelle.

Nous avons identifié que les déterminants des arrêts de travail de longue durée étaient principalement des caractéristiques liées à la santé des salariés (une mauvaise santé ou une maladie chronique sont les déterminants principaux) mais il apparaît que l'exposition à des facteurs de risques psychosociaux, comme avoir un mauvais rapport avec ses supérieurs hiérarchiques, ou que la pénibilité physique de l'emploi, sont aussi des facteurs très importants. Au contraire, il est difficile d'évaluer les facteurs déterminants des arrêts de travail de courte durée et de très courte durée.

Cette étude a permis de proposer une hiérarchisation des facteurs déterminants des arrêts maladie. Elle démontre l'importance des conditions de travail, autant physiques que psychique. L'utilisation des forêts aléatoires semble de plus une approche méthodologique intéressante pour la compréhension de grandes enquêtes dont les items semblent tous très corrélés.

Hierarchizing determinants of sick leave: insights from a survey on health and wellbeing at the workplace

Running head: Hierarchizing sick leave determinants

Abstract (135 words)

Objective- We hierarchized a range of individual and occupational factors impacting the occurrence of very short (1-3 days), short (4 days to 1 month) or long-term (more than a month) sick leave spells.

Methods- Data were collected from a repeated cross-sectional survey conducted in the French private sector over the period 2011-2017. Fifty one sick leave determinants were ranked using a conditional random forest approach.

Results- The main determinants of long-term sick leaves were mainly health-related characteristics, such as perceived health, but also work-related covariates such as supervisor acknowledgement. On the contrary, very short-term spells were mainly defined by sociodemographic covariates.

Conclusion- These results could be useful for devising appropriate actions to prevent against sick leave at the workplace, particularly long-term spells. Random forest approach is a promising approach for ranking correlated covariates from large datasets.

Introduction

The study of sick leave (SL) suggests that it has many determinants, including working conditions, sociodemographic factors, health status, behavior, access to medical care, among many others [various refs]. The epidemiology of absence from work is all the more complex because these multiple causes are not independent. However, it would nevertheless be useful for decision-makers to have a ranking of all potential determinants of sick leave, in order to get an improved understanding of the importance of each, and better guide future prevention strategies, especially since SL data is recorded in all companies and, if properly analyzed, could be used to improve risk prevention.

Another difficulty related to sick leave is that it has a dual characteristic: the frequency of occurrence, and the duration of each spell. Thus, the determinants of SL may be different between long-term and short-term episodes [REF], and also different between rare and recurrent episodes [REF]. Understanding the underlying processes for these different types of sick leave is crucial for prevention policies. Better understanding, however, requires the use of appropriate statistical methods that take into account, in addition to frequency and duration, the large number of potential factors, and also possible collinearity and interaction between these factors.

The main objective of this study is to identify and prioritize the determinants of SL, separately for very short (up to 3 days), short (from 4 days to 1 month), and long-term (more than one month) sick leave episodes. The potential determinants are 51 variables describing the socio-demographic, health, and professional characteristics of 32,000 employees in the French private sector. The statistical analysis we employ involves a conditional random forest approach which can easily deal with large numbers of variables and potential interactions between them.

Methods

Data

Survey design

Data were collected from 2011 to 2017 by a self-administrated questionnaire completed by the participants during a repeated cross-sectional health and wellbeing survey at the workplace in French private sector. This nationwide survey has been conducted by Malakoff Médéric Humanis, a French health insurance company. The study questionnaire was emailed annually and, each year, approximately 4500 questionnaires were retrieved from a representative sample of the French population working in the private sector, in terms of age, gender, occupational sector, company size and area. A total of 32,327 responses were collected. To avoid missing data, a full answering of the questionnaire was required. The survey was constructed in order to describe the temporal evolution of the working population behaviors in terms of health, wellbeing, self-declared sick leave spells, workers' relationships with colleagues and hierarchy and workers' feelings towards changes at work. The questionnaire can be read in Supplemental Digital Content.

Sick leave assessment

SL behavior of workers was assessed with a question formulated as follows: "*How many times did you take sick leave (excluding maternity/parternity leave) during the last 12 months ?"*. The possible answers were "*Never"* and "*At least once"*. This question was used to document four types of sick leave: less than 3 days, between 4 and 5 days, more than a week, and more than a month. However, as the last two categories overlap, we focused only on the longest SL (which is the most severe and the most expansive [1], [2]) declared for the last 12 months to stratify the SL lengths into 3 classes: long-term SL (longer than a month), short-term SL (between 4 days and a month) and very short-term SL (less than 3 days). We chose to split the very short-term SL spells from the others because in France, 3 days corresponds to the "waiting period", i.e., an initial period in which workers are not compensated when taking SL.

Covariates

The questionnaire included 51 questions corresponding to:

- *Sociodemographic covariates:* at the individual level (for example, sex, age and family situation), company level (for example, size of the company and occupational sector), and at the geographic level (region of the country).
- *Health covariates:* for example, having a handicap, having a chronic disease, perceived health and pain at several locations.
- *Work-related covariates:* for example, level of job difficulty, exposure to risk factors, psychosocial risks, level of satisfaction.

All covariates are categorical, and the number of categories per covariate is between 2 and 13. An exhaustive description of the covariates (titles of all questions, possible answers and descriptive statistics) is available in **Supplemental Digital Content.**

Statistical methods

To hierarchize the covariates in terms of impact on the occurrence of SL, we performed random forests for each of the 3 SL categories: each model classifies workers in either the non-SL group or the SL class of interest. Although logistic modeling would traditionally be used in this setting, the high number of covariates and potential correlation between them - which can raise issues for logistic model [3]- led us to focus on the random forest approach, a non-parametric method that can manage such correlations more easily . In particular, the use of variable selection in logistic models to deal with the potential correlation could affect the ranking of the covariates as some important variables could be excluded as a consequence of a strong correlation. The random forest algorithm also tests implicitly interactions between covariates and it would be computationally very expensive to test those with a linear parametric model. Interactions are indeed a significant characteristic of SL [4]. We included all the 51 covariates in the analysis as the objective is to determine which factors are important in the occurrence of SL, with no *a priori* selection.

We ran the random forest algorithm using conditional inference trees [5]. This method avoids the bias in variable importance measures that arises from the different number of categories in different

covariates [6]. We split our data into a training set (80%) and a test set (20%) to evaluate the discriminative power of the trained forests and to avoid overfitting. The evaluation criterion we chose was the area under the ROC curve (AUC). We performed the analyses with R 3.4.3 and with the package *party (version 1.3.1)* [6]–[8].

We evaluated variable importance using an AUC-based permutation measure [9]. To check the consistency of our results, we trained 30 forests with 500 trees for each SL category versus no-SL workers, changing the random seed each time. We performed those replicates because Random Forest is an algorithm using random simulation, and we needed to check if the variable importance results obtained were true one and not simulation artefacts[10]. Choosing 30 forests and 500 trees is a compromise between the literature [10] which is recommending 50 forests and 2000 trees and our choice of algorithm which is more time consuming [11]. Moreover, we also found consistent results and a stable hierarchy, which does not call for other replications. We present in the *Results* section the mean variable importance measures calculated over the 30 random forests and discuss only variables that showed "consistent" results across the 30 forests. In fact, we chose the variables that had a mean importance value above 0.0025 and, to evaluate the variability of the results, we show the range of importance values among the 30 random forest replications. We chose 0.0025 as an *a priori* threshold; this could be interpreted as a 5% if we have an AUC of 0.5, which is the worst possible AUC [9].

Results

Descriptive statisticsErreur ! Source du renvoi introuvable.**Table 1** provides a description of the sociodemographic variables of the study. The full range of descriptive statistics are shown in **Supplemental Digital Content**. Women made up 46% of the overall population and were thus slightly overrepresented among workers with SL: 51% of the long-term SL group and 50% of the very short-term and short-term SL groups were women. We note also that 51% of the overall population was younger than 40, while 8% was older than 56. In this respect, the younger group as overrepresented in very-short term SL (65%), the older group in long-term SL (12%). Furthermore, 39% of the population

was employed in companies with more than 1000 workers, and this population was overrepresented in general in SL statistics, especially in the very short-term SL group (42%). Note also the most common occupational sectors in the general population were Manufacturing (25%), Services (23%), and Commerce (16%).

Insert Table 1 here

Table 1 - Sociodemographic description

One important conclusion from an in-depth descriptive analysis is that workers with no SL in the previous year declared being healthier and having better working conditions than those who took SL. Indeed, 77% of the population who took no SL considered their health as good or very good, versus 72% of workers in the very short-term SL group, 58% in the short-term SL group, and 38% in the long-term SL group. As for the variable: "worker considered their job nervously tiring", 66% of those who took no SL agreed, versus 71% of workers in the very short-term SL group, 74% in the short-term SL group, and 78% in the long-term SL group.

Importance of variables determining very short-, short-, and long-term sick leave

Figure 1 (a-c) shows an ordering of the covariates according to their mean importance as calculated from their 30 random forests for each model. The range of minimum to maximum values over the 30 forests is also shown.

The very-short term SL model had an AUC of 0.69, and the most important variable was age, leading to a mean decrease in AUC of 0.013. Age had a much higher impact than the subsequent variables (the next most important variable was "working in front of a screen", with an importance of 0.006). Other variables characterizing differences between the very short-term SL population and the population

with no SL in the previous year were, in descending order: lack of tonicity, region, company size, and head pain.

Second, the short-term SL model had an AUC of 0.70. The most important variable was perceived health of workers, corresponding to a mean decrease in AUC of 0.009. The subsequent covariates had slightly lower importance values: chronic disease, lack of tonicity, back pain, company size, anxiolytics, working in a noisy environment, and neck pain, each with importance values below 0.006.

Third, the long-term SL model had an AUC of 0.82, the highest of the three models. The two most important variables for defining workers with long-term SL versus those with no SL in the previous year were chronic disease, with a decrease in AUC of 0.024, and perceived health, with a decrease of 0.021. Two further covariates stand out from the others: "handicap", with a value of 0.012, and "anxiolytics" (taking anxiolytics, sleeping pills, or antidepressants), with an importance value of 0.010. The next 6 most important variables in characterizing long-term SL were lack of tonicity, back pain, leg pain, sleep troubles, acknowledgment from superiors, neck pain and carrying heavy loads at work.

Insert Figure 1 here

Figure 1- Covariates hierarchized according to their mean importance for the occurrence of (a) very short-term SL (≤3 days), (b) short-term SL (between 4 and 30 days) and (c) long-term SL (>30 days). The horizontal line around the mean value represents the range of importance values over 30 random forest replications. The plain vertical line represents the 0.0025 a priori threshold.

Discussion

When discussing SL prevention, a major challenge is to understand determinants of SL. Here. we have examined the importance of various covariates with respect to SL occurrence. Three major conclusions from our analyses are discussed below.

Workers taking long-term SL are defined by poor health but also poor working conditions

The model for long-term SL has a high AUC of 0.82, which means that workers with long-term SL can be clearly identified with respect to those that took no SL. The distinction between the former and the latter groups is mainly characterized by health-related covariates. The two most important

discriminatory variables were chronic disease and perceived health, which were more than twice as important as all other variables. The next eight ranked covariates were also closely related to health. The causes of the health issues involved are unknown, and may be partially linked to workers' jobs, but in any case, this overwhelming output of health-related variables tells us that long-term SL is mainly connected with health issues. We expected this result as other publications already pointed out the importance of health in long-term SL [12], [13] and our analysis underlines that this parameter is much more important than the others. This result is also very reassuring since we are trying to predict sick leave: it seems natural that health is the most important factor.

The last two selected covariates require further discussion. The first of them is acknowledgement by superiors; this suggests that psychosocial risk factors may have an impact in the occurrence of long-term SL. This result can be linked with the other variables selected that could be related to psychosocial health: the taking of anxiolytics, sleeping pills or antidepressant, lack of tonicity and sleep issues. This importance of psychosocial risk factors we see here is in accordance with results already published on psychosocial risks and long-term SL [14], [15]. The last selected covariate is carrying heavy loads at work, which is also related to working conditions. The appearance of these two variables suggests the potential for reducing the number of long-term SL spells through management intervention: workers with poor health are seriously affected by these long-term SL and poor working conditions worsen this vulnerability.

Lastly, we would like to highlight the absence of the variable "age" in the group of covariates selected. Indeed, age is usually pointed to as one of the main factors defining long-term absence [16], [17] and our descriptive statistics showed that older people are overrepresented among workers with long-term SL. Indeed, 8% percent of workers are more than 56 years old in our study, and their proportion is 12% in the long-term SL population. However, we only retrieved a mean importance value of 0.0016 for this variable. A likely explanation of this result is that age is not a direct predictor of long-term SL spells but instead a proxy for health, which is the main predictor for such SL. We found a Cramer's V

of 0.16 with chronic disease and of 0.08 with perceived health. Age is therefore a necessary covariate when health data are not available, but it is better to use health data when possible.

Differences between workers who took very short-term SL and those who did not take SL are mainly related to sociodemographic variables

The model for very short-term SL has weaker results than the long-term SL model, but still has a relatively decent discriminative power, with an AUC of 0.69. The leading six covariates are almost completely different to those found for the long-term SL model. Only two of them are related to health, the others are sociodemographic variables. Age is the most important variable and induces a decrease in AUC which is 2.5 times greater than the second most important variable. Furthermore, we saw from the descriptive statistics that workers below 40 years old are over-represented in the group of workers who took very short-term SL. This may be in part explained by the fact that this population potentially has young children, which leads to a more difficult work-life balance. However, there are doubts associated with this conclusion because the number of children was a covariate in the model but did not emerge as an important one.

Working in front of a screen was the second most important variable, and the descriptive statistics showed that employees working in front of a screen most of the time represented 59% of the workers with no SL spells the previous year, and 70% of those in the very short-term SL group. Very short SL spells therefore mainly affect office workers. This does not mean that workers with manual jobs take less SL, but rather that their job and health issues may lead to longer SL spells. Indeed, employees who never worked in front of a screen represented 16% of the employees without SL in the previous year (respectively 59% for workers working most of the time in front of a screen) and 23% of those taking at least one long-term SL (respectively 45%). The nature of the job therefore influences patterns of absence.

The other sociodemographic covariates were company size and administrative region. For the former, the descriptive statistics showed that workers in companies with more than 1000 employees were

over-represented in short-term SL compared to the overall population. One reason for this might be the nature of smaller companies: being absent in these may have more consequences on operations than in larger companies, which may be a brake on taking SL, leading instead to presenteeism. Indeed, workers with greater responsibilities are more prone to presenteeism [18]. The importance of the region people live in shows that differences in lifestyle and culture are related to SL. For instance, though they represent only 12% of the population without SL, the inhabitants of the Île de France region are 16% of the population that took very short-term spells. This overrepresentation in the SL group can be explained by the fact that this region (which contains Paris) is the most urban region of France, and in particular a region where transport times can be long. As such trips are more unpleasant, employees in this region may be more likely to take very short spells, thus avoid such inconveniences. Note finally that the two selected health-related covariates were lack of tonicity and head pain: these covariates may also be symptoms related to the exposure to psychosocial risk factors, but it shows that a poorer health also leads to very short term sick leave.

All workers are likely to have very short- or short-term SL spells at some point

Models for very short-term and short-term SL populations have relatively low AUCs (0.69 and 0.70 respectively) compared to the long-term SL model (0.82). The low AUC of the very-short term SL class may be due to several factors. First, workers taking very short SL spells may not be clearly categorizable, in part because the causes of very short-term spells may be essentially random and unpredictable. For instance, a very short SL spell might be the consequence of a seasonal illness [19]. The predictability of short-term spells seems then very poor. In contrast, the discriminating power for long-term SL is quite high because those spells occur mostly when workers have poor health and working conditions, implying the existence of variables that can be analyzed further.

The second reason behind these low AUCs for very short- and short-term SL could be related to the data structure. With more detailed information available, we could expect a better classification performance. Indeed, we can only distinguish workers who took at least one SL from those who did

not; workers who took repeated SL spells may have had different characteristics and been easier to identify [20]–[22]. Furthermore, if we knew the exact dates of each spell, we could have included extra information such as seasonal outbreaks, which are known as predictors of SL spells [23]. Lastly, the short-term SL group included spells that could have quite different natures, e.g., the causes of a 4-day spell may be completely different to those of a 30-day one, which, as an aside, could be why the variables selected here were a mixture of those selected by the other two models (with the exception of "Noisy Environment" and "Neck Pain"). The results for short-term SL seems irrelevant because the short-term SL spells are too heterogeneous, and the cut-off does not seem convincing, which is why we have focused our discussion on the other two models.

Limitations of the study

The use of declarative data may be the main limitation of this study. First, we have described perceived health and perceived working conditions, which are not objective measures. Furthermore, we described self-declared SL, not certified SL. This could lead to a selection bias in our data because workers may confuse very short-term with short-term spells, or short-term with long-term spells. Some workers may also remember more easily having a certain type of spell (in particular, very short-term ones) than other workers. Here we have modeled populations of workers who have remembered having spells, rather than workers that had spells, and the two may be slightly different. The SL prevalence observed was however consistent with results from other French studies. Indeed, in the French study "SUMER" on working conditions and occupational risks, 35% of the workers took SL in 2010 [24], and in the French study "PSCE" on companies' supplementary social protection, 30% of works took SL in 2009 [25]. In our study, 36% of the workers took at least SL the previous year.

Another limitation comes from the use of random forests. Random forests do not provide any information on the direction or strength of covariates' impact on the occurrence of SL. We nevertheless chose to work with random forests because parametric approaches were unsuccessful with our data

set and with ranking purposes, whereas random forests have the capacity to allow for interactions and high correlation between covariates [26], [27] and a high dimensional context.

Policy implications

In this analysis, we have shown that the population that took long-term SL is mainly defined by poor health characteristics. As these are the costliest and most serious spells, they require particular attention, and a decent health prevention policy at work could be provided to contain this phenomenon. Two work-related covariates were also shown to be important, which suggest a potential place for management intervention, and the importance of "acknowledgement by the hierarchy" calls especially for a greater attention to psychosocial risk factors.

Very short-term and short-term SL populations are more difficult to define: every worker seems likely to have these kinds of spell at any point. However, certain sociodemographic differences appear to exist, and workers in large companies, in certain regions, and with office jobs, are more likely to take short-term SL. These results suggest interventions that are differentiated according to socio-demographic differences in absence behavior: for instance, telework programmes could be proposed for urban office workers in order to reduce the occurrence of very short-term SL. Moreover, it should be noted that no covariates with any obvious intervention were related to very short-term SL, such as for instance poor management or poor posture.

Concluding remarks

In this paper, we have proposed an ordering of covariates characterizing three types of workers taking different SL spells: at least one lasting more than a month, at least one lasting 4 days to a and at least one of less than 3 days.

We have shown that workers that took long-term SL spells were characterized essentially by variables involving poor health, though also by poor job conditions involving physical hardship and exposure to psychosocial risk factors. There is therefore potential for improvement using management

intervention. We then showed that the most relevant covariate for very short-term SL was age, followed by working in front of a screen, lack of tonicity, region, company size, and head pain. Such results call for interventions that are differentiated according to socio-demographic profiles but do not immediately suggest any management interventions that could reduce very short-term SL. Finally, we showed that workers with short- and very short-term SL spells were harder to define. The main issue with very short-term SL spells is that they may affect a given population indiscriminately and could be essentially governed by unpredictable processes. The main issue with short-term SL is that it cannot be defined as a homogenous group; a more precise cut-off or another data structure could be needed to better understand this phenomenon.

References

[1] M. Henderson, N. Glozier, and K. H. Elliott, 'Long term sickness absence', *BMJ*, vol. 330, no. 7495, pp. 802–803, Apr. 2005.

[2] Caisse Nationale d'Assurance Maladie, 'Améliorer la qualité du système de santé et maîtriser les dépenses - Rapports Charges et produits pour les années 2018 et 2019', Jul. 2018.

[3] M. J. Mackinnon and M. L. Puterman, 'Collinearity in generalized linear models', *Communications in Statistics - Theory and Methods*, vol. 18, no. 9, pp. 3463–3472, Jan. 1989.

[4] J. P. Vistnes, 'Gender Differences in Days Lost from Work Due to Illness', *ILR Review*, vol. 50, no. 2, pp. 304–323, Jan. 1997.

[5] T. Hothorn, K. Hornik, and A. Zeileis, 'Unbiased Recursive Partitioning: A Conditional Inference Framework', *Journal of Computational and Graphical Statistics*, vol. 15, no. 3, pp. 651–674, Sep. 2006.

[6] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, 'Bias in random forest variable importance measures: Illustrations, sources and a solution', *BMC Bioinformatics*, vol. 8, no. 1, p. 25, Jan. 2007.

[7] T. Hothorn, P. Bühlmann, S. Dudoit, A. Molinaro, and M. J. Van Der Laan, 'Survival ensembles', *Biostatistics*, vol. 7, no. 3, pp. 355–373, Jul. 2006.

[8] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, 'Conditional variable importance for random forests', *BMC Bioinformatics*, vol. 9, no. 1, p. 307, Jul. 2008.

[9] S. Janitza, C. Strobl, and A.-L. Boulesteix, 'An AUC-based permutation variable importance measure for random forests', *BMC Bioinformatics*, vol. 14, no. 1, p. 119, Apr. 2013.

[10] R. Genuer, J.-M. Poggi, and C. Tuleau-Malot, 'Variable selection using Random Forests', p. 11.

[11] C. Strobl, T. Hothorn, and A. Zeileis, 'Party on!', *The R Journal*, vol. 1, no. 2, pp. 14–17, 2009.

[12] L. E. C. Wijnvoord, J. J. L. Van der Klink, M. R. De Boer, and S. Brouwer, 'Predictors of sickness absence in college and university educated self-employed: a historic register study', *BMC Public Health*, vol. 14, p. 420, May 2014.

[13] M. Laaksonen, P. Martikainen, O. Rahkonen, and E. Lahelma, 'Explanations for gender differences in sickness absence: evidence from middle-aged municipal employees from Finland', *Occup Environ Med*, vol. 65, no. 5, pp. 325–330, May 2008.

[14] T. Clausen, H. Burr, and V. Borg, 'Does Affective Organizational Commitment and Experience of Meaning at Work Predict Long-Term Sickness Absence? An Analysis of Register-Based

Outcomes Using Pooled Data on 61,302 Observations in Four Occupational Groups':, *Journal of Occupational and Environmental Medicine*, vol. 56, no. 2, pp. 129–135, Feb. 2014.
[15] T. Lund, M. Labriola, K. B. Christensen, U. B??ltmann, E. Villadsen, and H. Burr, 'Psychosocial Work Environment Exposures as Risk Factors for Long-Term Sickness Absence Among Danish Employees: Results From DWECS/DREAM':, *Journal of Occupational and Environmental Medicine*, vol. 47, no. 11, pp. 1141–1147, Nov. 2005.
[16] M. B. Nielsen, A.-M. R. Indregard, and S. Øverland, 'Workplace bullying and sickness absence: a systematic review and meta-analysis of the research literature', *Scand J Work Environ Health*, vol. 42, no. 5, pp. 359–370, Sep. 2016.
[17] J. Airaksinen *et al.*, 'Prediction of long-term absence due to sickness in employees: development and validation of a multifactorial risk score in two cohort studies', *Scandinavian Journal of Work, Environment & Health*, vol. 44, no. 3, pp. 274–282, May 2018.
[18] R. M. Merrill, S. G. Aldana, J. E. Pope, D. R. Anderson, C. R. Coberley, and and the H. R. S. S. Whitmer R. William, 'Presenteeism According to Healthy Behaviors, Physical Health, and Work Environment', *Population Health Management*, vol. 15, no. 5, pp. 293–301, Aug. 2012.
[19] E. B. Akyeampong, 'Trends and seasonality in absenteeism', *Statistics in Canada*, vol. 8, pp. 13–15, Jun. 2007.
[20] C. A. M. Roelen, P. C. Koopmans, J. H. de Graaf, J. W. van Zandbergen, and J. W. Groothoff, 'Job demands, health perception and sickness absence', *Occup Med (Lond)*, vol. 57, no. 7, pp. 499–504, Oct. 2007.
[21] C. M. Stapelfeldt *et al.*, 'Are environmental characteristics in the municipal eldercare, more closely associated with frequent short sick leave spells among employees than with total sick leave: a cross-sectional study', *BMC Public Health*, vol. 13, no. 1, Dec. 2013.
[22] A. Väänänen, S. Toppinen-Tanner, R. Kalimo, P. Mutanen, J. Vahtera, and J. M. Peiró, 'Job characteristics, physical and psychological symptoms, and social support as antecedents of sickness absence among men and women in the private industrial sector', *Social Science & Medicine*, vol. 57, no. 5, pp. 807–824, Sep. 2003.
[23] J. Ryan, Y. Zoellner, B. Gradl, B. Palache, and J. Medema, 'Establishing the health and economic impact of influenza vaccination within the European Union 25 countries', *Vaccine*, vol. 24, no. 47–48, pp. 6812–6822, Nov. 2006.
[24] T. Lesuffleur, J.-F. Chastang, N. Sandret, and I. Niedhammer, 'Psychosocial factors at work and sickness absence: Results from the French National SUMER Survey: Psychosocial Factors at Work and Sickness Absence', *American Journal of Industrial Medicine*, vol. 57, no. 6, pp. 695–708, Jun. 2014.
[25] C. Pollak, 'The impact of a sick pay waiting period on sick leave patterns', *The European Journal of Health Economics*, vol. 18, no. 1, pp. 13–31, Jan. 2017.
[26] A.-L. Boulesteix, S. Janitza, J. Kruppa, and I. R. König, 'Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics', *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 2, no. 6, pp. 493–507, 2012.
[27] L. Breiman, 'Random Forests', *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001.

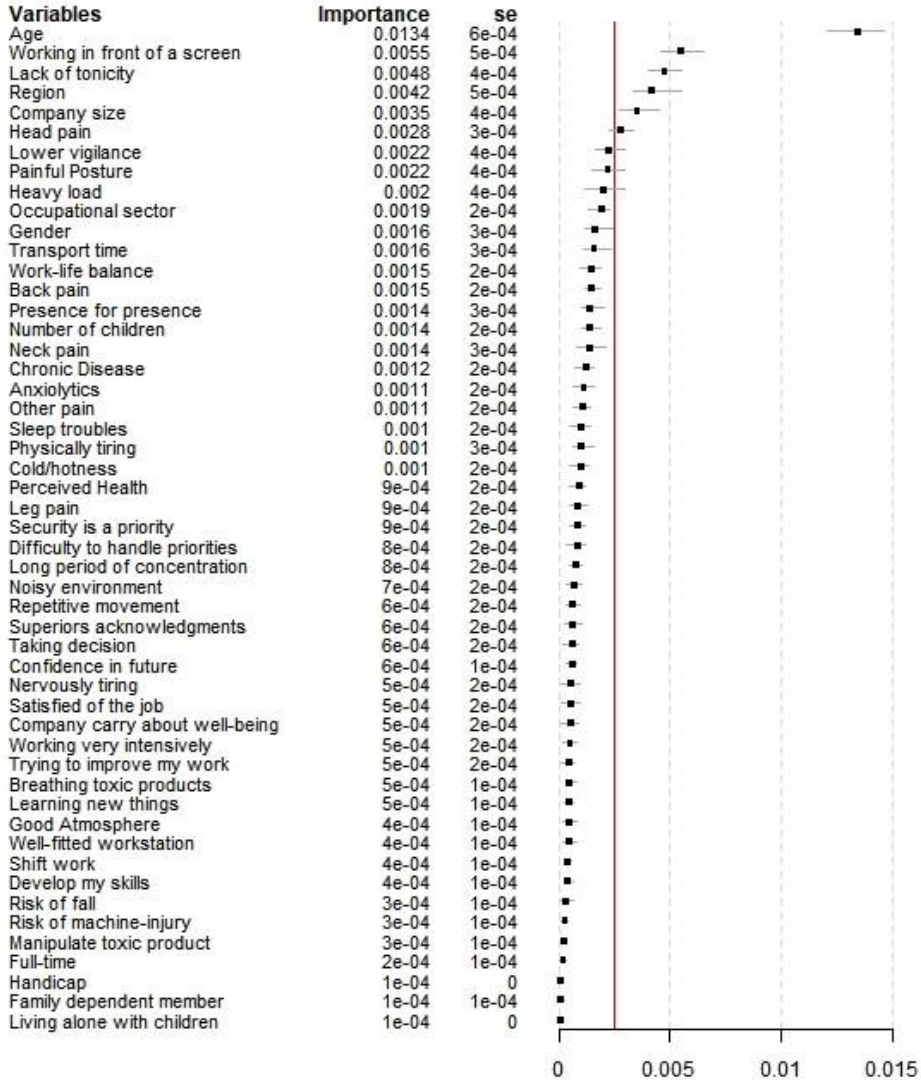
Variables Importance se
Age 0.0134 6e-04
Working in front of a screen 0.0055 5e-04
Lack of tonicity 0.0048 4e-04
Region 0.0042 5e-04
Company size 0.0035 4e-04
Head pain 0.0028 3e-04
Lower vigilance 0.0022 4e-04
Painful Posture 0.0022 4e-04
Heavy load 0.002 4e-04
Occupational sector 0.0019 2e-04
Gender 0.0016 3e-04
Transport time 0.0016 3e-04
Work-life balance 0.0015 2e-04
Back pain 0.0015 2e-04
Presence for presence 0.0014 3e-04
Number of children 0.0014 2e-04
Neck pain 0.0014 3e-04
Chronic Disease 0.0012 2e-04
Anxiolytics 0.0011 2e-04
Other pain 0.0011 2e-04
Sleep troubles 0.001 2e-04
Physically tiring 0.001 3e-04
Cold/hotness 0.001 2e-04
Perceived Health 9e-04 2e-04
Leg pain 9e-04 2e-04
Security is a priority 9e-04 2e-04
Difficulty to handle priorities 8e-04 2e-04
Long period of concentration 8e-04 2e-04
Noisy environment 7e-04 2e-04
Repetitive movement 6e-04 2e-04
Superiors acknowledgments 6e-04 2e-04
Taking decision 6e-04 2e-04
Confidence in future 6e-04 1e-04
Nervously tiring 5e-04 2e-04
Satisfied of the job 5e-04 2e-04
Company carry about well-being 5e-04 2e-04
Working very intensively 5e-04 2e-04
Trying to improve my work 5e-04 2e-04
Breathing toxic products 5e-04 1e-04
Learning new things 5e-04 1e-04
Good Atmosphere 4e-04 1e-04
Well-fitted workstation 4e-04 1e-04
Shift work 4e-04 1e-04
Develop my skills 4e-04 1e-04
Risk of fall 3e-04 1e-04
Risk of machine-injury 3e-04 1e-04
Manipulate toxic product 3e-04 1e-04
Full-time 2e-04 1e-04
Handicap 1e-04 0
Family dependent member 1e-04 1e-04
Living alone with children 1e-04 0
0
0.005
0.01
0.015

Figure 1a

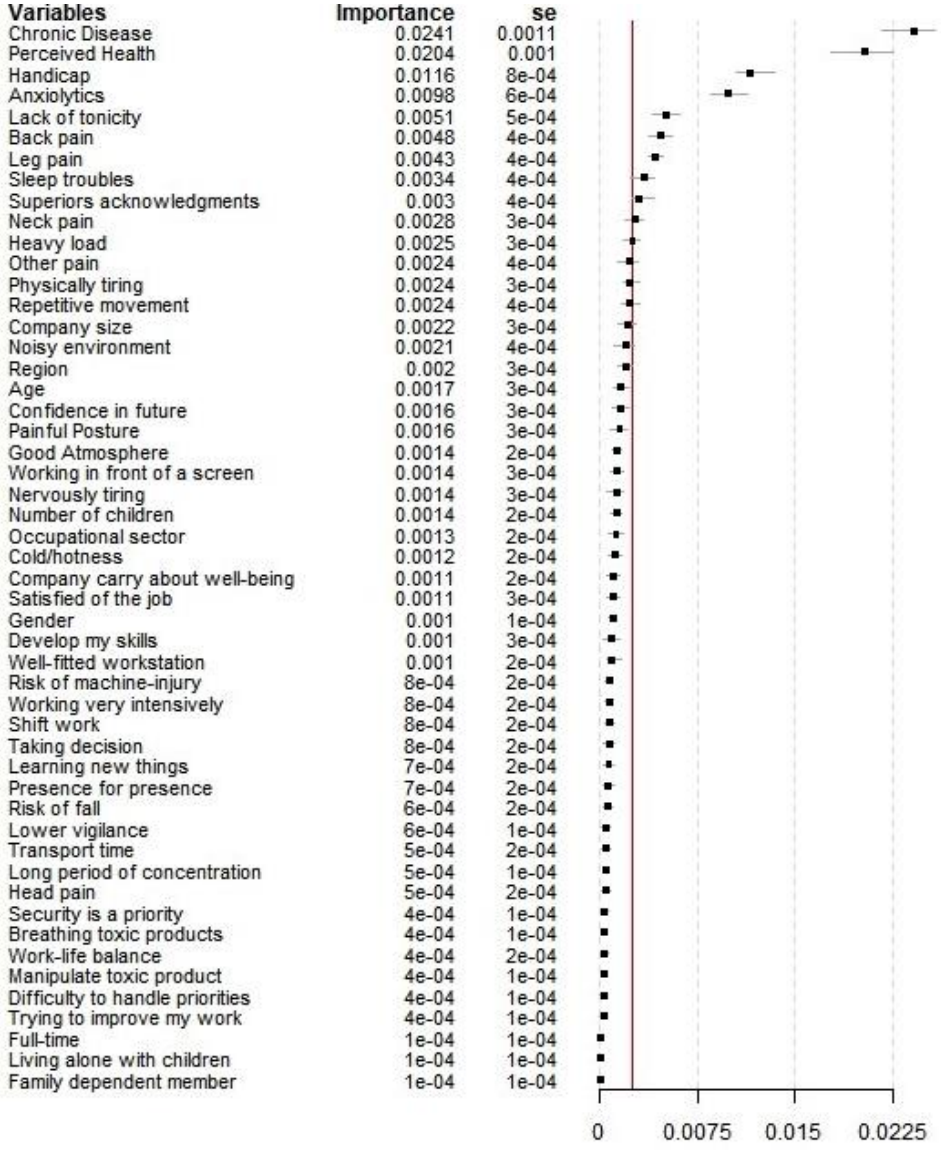
Variables Importance se
Chronic Disease 0.0241 0.0011
Perceived Health 0.0204 0.001
Handicap 0.0116 8e-04
Anxiolytics 0.0098 6e-04
Lack of tonicity 0.0051 5e-04
Back pain 0.0048 4e-04
Leg pain 0.0043 4e-04
Sleep troubles 0.0034 4e-04
Superiors acknowledgments 0.003 4e-04
Neck pain 0.0028 3e-04
Heavy load 0.0025 3e-04
Other pain 0.0024 4e-04
Physically tiring 0.0024 3e-04
Repetitive movement 0.0024 4e-04
Company size 0.0022 3e-04
Noisy environment 0.0021 4e-04
Region 0.002 3e-04
Age 0.0017 3e-04
Confidence in future 0.0016 3e-04
Painful Posture 0.0016 3e-04
Good Atmosphere 0.0014 2e-04
Working in front of a screen 0.0014 3e-04
Nervously tiring 0.0014 3e-04
Number of children 0.0014 2e-04
Occupational sector 0.0013 2e-04
Cold/hotness 0.0012 2e-04
Company carry about well-being 0.0011 2e-04
Satisfied of the job 0.0011 3e-04
Gender 0.001 1e-04
Develop my skills 0.001 3e-04
Well-fitted workstation 0.001 2e-04
Risk of machine-injury 8e-04 2e-04
Working very intensively 8e-04 2e-04
Shift work 8e-04 2e-04
Taking decision 8e-04 2e-04
Learning new things 7e-04 2e-04
Presence for presence 7e-04 2e-04
Risk of fall 6e-04 2e-04
Lower vigilance 6e-04 1e-04
Transport time 5e-04 2e-04
Long period of concentration 5e-04 1e-04
Head pain 5e-04 2e-04
Security is a priority 4e-04 1e-04
Breathing toxic products 4e-04 1e-04
Work-life balance 4e-04 2e-04
Manipulate toxic product 4e-04 1e-04
Difficulty to handle priorities 4e-04 1e-04
Trying to improve my work 4e-04 1e-04
Full-time 1e-04 1e-04
Living alone with children 1e-04 1e-04
Family dependent member 1e-04 1e-04
0
0.0075
0.015
0.0225

Figure 1b

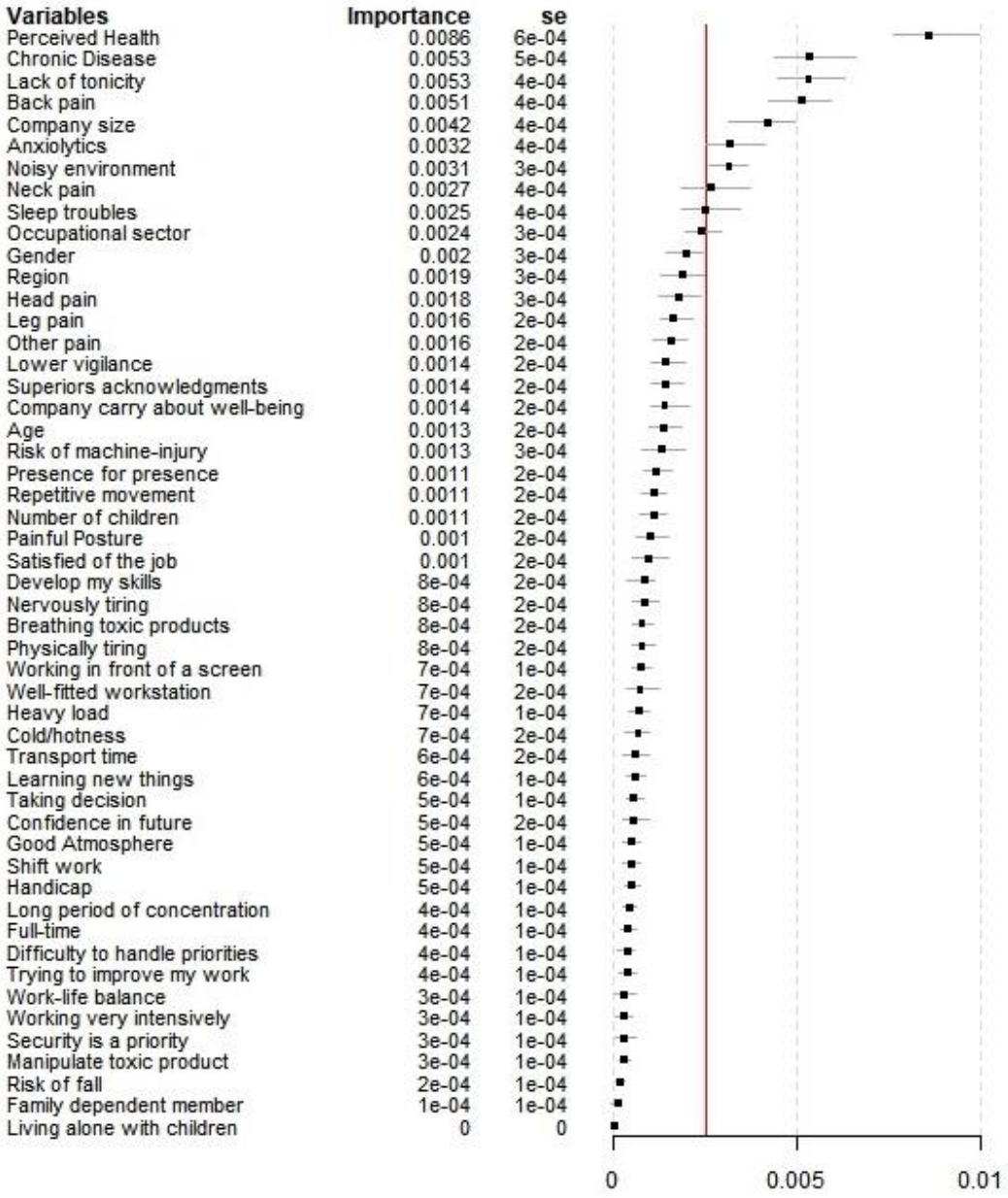
Variables Importance se
Perceived Health 0.0086 6e-04
Chronic Disease 0.0053 5e-04
Lack of tonicity 0.0053 4e-04
Back pain 0.0051 4e-04
Company size 0.0042 4e-04
Anxiolytics 0.0032 4e-04
Noisy environment 0.0031 3e-04
Neck pain 0.0027 4e-04
Sleep troubles 0.0025 4e-04
Occupational sector 0.0024 3e-04
Gender 0.002 3e-04
Region 0.0019 3e-04
Head pain 0.0018 3e-04
Leg pain 0.0016 2e-04
Other pain 0.0016 2e-04
Lower vigilance 0.0014 2e-04
Superiors acknowledgments 0.0014 2e-04
Company carry about well-being 0.0014 2e-04
Age 0.0013 2e-04
Risk of machine-injury 0.0013 3e-04
Presence for presence 0.0011 2e-04
Repetitive movement 0.0011 2e-04
Number of children 0.0011 2e-04
Painful Posture 0.001 2e-04
Satisfied of the job 0.001 2e-04
Develop my skills 8e-04 2e-04
Nervously tiring 8e-04 2e-04
Breathing toxic products 8e-04 2e-04
Physically tiring 8e-04 2e-04
Working in front of a screen 7e-04 1e-04
Well-fitted workstation 7e-04 2e-04
Heavy load 7e-04 1e-04
Cold/hotness 7e-04 2e-04
Transport time 6e-04 2e-04
Learning new things 6e-04 1e-04
Taking decision 5e-04 1e-04
Confidence in future 5e-04 2e-04
Good Atmosphere 5e-04 1e-04
Shift work 5e-04 1e-04
Handicap 5e-04 1e-04
Long period of concentration 4e-04 1e-04
Full-time 4e-04 1e-04
Difficulty to handle priorities 4e-04 1e-04
Trying to improve my work 4e-04 1e-04
Work-life balance 3e-04 1e-04
Working very intensively 3e-04 1e-04
Security is a priority 3e-04 1e-04
Manipulate toxic product 3e-04 1e-04
Risk of fall 2e-04 1e-04
Family dependent member 1e-04 1e-04
Living alone with children 0 0
0
0.005
0.01

Figure 1c

	General population	No SL	Very short-term SL	Short-term SL	Long-term SL
n	32,327	20,665	4,128	5,187	2,347
Gender					
Female	46%	44%	50%	50%	51%
Male	54%	56%	50%	50%	49%
Age (years)					
<30	21%	20%	29%	20%	14%
30-39	30%	29%	36%	31%	25%
40-44	15%	15%	14%	15%	16%
45-49	13%	14%	8%	13%	14%
50-55	14%	14%	9%	14%	20%
≥56	8%	8%	5%	7%	12%
Company size					
More than 1000	39%	37%	42%	40%	41%
49 to 999	24%	23%	25%	28%	28%
10 to 49	12%	13%	12%	12%	12%
Less than 10	25%	27%	20%	19%	19%
Occupational sector					
Consulting and engineering	5%	5%	7%	3%	2%
Construction industry	8%	8%	8%	9%	9%
Commerce	16%	16%	17%	17%	17%
Manufacturing	25%	24%	24%	29%	28%
Health	9%	9%	8%	9%	11%
Services	23%	24%	23%	20%	20%
Transport, telecom and energy	14%	14%	13%	13%	13%

Table 1

2.2.3.4 Discussion complémentaire à l'article

L'article précédent a permis de mettre en évidence une hiérarchie des facteurs déterminants des arrêts en fonction de leur durée. Nous aimerions ajouter quelques éclairages complémentaires avant de passer à de nouvelles analyses.

Un premier élément de discussion est l'introduction de la variable *santé perçue*. Nous pouvons nous demander si l'introduction de cette variable n'est pas problématique puisqu'un arrêt maladie est, il faut tout de même le rappeler, pris pour raison de maladie. Ainsi, un arrêt maladie grave est souvent synonyme d'une mauvaise santé et, en introduisant cette variable dans le modèle, nous risquons de prédire une variable en utilisant une approximation de cette dernière. Les résultats du modèle montrent tout de même que la santé perçue, en tant qu'indicateur, n'est pas suffisante pour prédire l'occurence d'arrêts graves même si elle en est le déterminant principal. Plutôt que comme un proxy des arrêts maladie, la santé perçue peut aussi être vue comme un facteur de médiation : les différentes variables déterminent la santé des salariés mais déterminent aussi la prise d'arrêt de travail. Ces arrêts ne dépendent en effet pas uniquement de la santé puisque, à santé perçue identique, un salarié peut décider de s'arrêter un autre pas. Nous pourrions ainsi mettre à jour le graphique introductif en Figure 1.1 en introduisant la santé en facteur de médiation, comme montré en Figure 2.1. L'introduction de la santé perçue dans le modèle est donc sujet à débat mais nous avions fait le choix de tout de même l'intégrer.

Un deuxième élément de discussion concerne la méthode de hiérarchisation des variables. Nous avons utilisé, dans l'article, une méthode de mesure d'importance basée sur une permutation aléatoire qui nous a permis d'évaluer le gain en performance associé à chaque variable. Cette méthode pose des problèmes puisqu'elle ne permet pas d'évaluer directement le sens de l'impact de ces variables et nous avons dû ruser pour l'évaluer en analysant notre base de données. D'autres méthodes d'analyse des algorithmes de *machine learning* ont récemment émergé et permettent d'évaluer plus précisément l'impact de chaque variable dans le processus de classification. C'est notamment le cas de la méthode SHAP [43, 44] qui permet d'expliquer l'importance de chaque variable dans la classification de chaque observation et donc, après une synthèse de l'ensemble des observations, du modèle en général. Des travaux engagés par un stagiaire ont permis d'appliquer cette méthode et de réfléchir à son usage en pratique pour expliquer les déterminants des arrêts dans les entreprises.

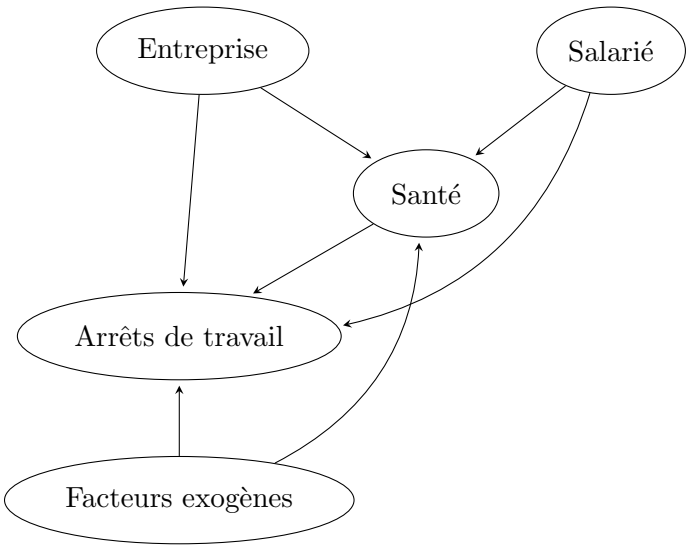

FIGURE 2.1 – Nouveau graphique synthétique des causes d'arrêt maladie introduisant l'effet médiateur de la santé

2.3 Trajectoires d'arrêts maladie

2.3.1 Objectif

La section précédente a balayé les différents déterminants des arrêts maladie en traitant des données d'enquête. Les résultats ont permis d'identifier les principaux facteurs qui peuvent mener à des arrêts maladie, et principalement de longue durée.

L'identification des déterminants est un exercice intéressant puisqu'elle permet ensuite de mettre en place des plans d'action mais elle est insuffisante : si l'on ne connait pas les déterminants spécifiques à notre entreprise, comment choisir son plan d'action ? Autre problème : les plans d'action peuvent être coûteux si on les propose à l'ensemble de la population de salariés, comment identifier les populations qui bénéficieraient le plus de ces plans d'action ?

Nous allons, dans cette section, mener une réflexion sur l'identification des profils de salariés qui ont un risque accru d'arrêts de longue durée, à partir de données facilement accessibles par les entreprises. Plus prosaïquement, nous allons chercher à construire, à partir de données administratives, un modèle de prédiction des arrêts de travail de longue durée. Les arrêts de longue durée nous intéresse le plus puisque, comme démontré dans la section précédente, ce sont des arrêts dont les causes semblent identifiables : une mauvaise santé, une exposition à des facteurs de risque psychosociaux, une pénibilité

physique de l'emploi. Les arrêts de travail plus courts sont plus difficiles à expliquer (ce qui peut s'observer lorsque l'on constate la faible capacité prédictive des modèles précédents) même s'ils peuvent être très importants pour les entreprises, surtout en termes d'organisation.

2.3.2 Données : les Déclarations Sociales Nominatives

Description générale des données Les analyses de cette section reposent sur des données immédiatement accessibles par l'entreprise : les *Déclarations Sociales Nominatives* (DSN). Ces données proviennent des systèmes des ressources humaines et doivent être obligatoirement déclarées mensuellement auprès de l'Etat et des organismes en charge de la prévoyance. Elles permettent d'identifier à un instant donné le nombre de salariés sous contrat, le nombre de salariés en arrêts maladie ainsi que leurs caractéristiques sociodémographiques telles que l'âge, le sexe, la catégorie socioprofessionnelles, le statut du contrat ou la rémunération. Ces données permettent ainsi de tracer la trajectoire d'absence de chacun des salariés.

Les DSN ont remplacé un autre système de données, les Déclarations Annuelles des Données Sociales (DADS), en 2017. Les DADS décrivaient les mêmes données, à quelques exceptions près, et étaient déclarées annuellement plutôt que mensuellement. Comme les DSN sont des données relativement récentes, quelques mois voire années ont été nécessaires pour obtenir des données propres robustes. Nous avons ainsi exclu des analyses toutes les déclarations antérieures au 1er janvier 2018 mais aussi les entreprises de moins de 50 salariés qui présentaient, dans l'ensemble, des données moins cohérentes.

Nous présenterons les échantillons et variables utilisées dans les analyses dans les sections suivantes.

Qualité des données Les DSN (et les DADS) sont des données relativement fiables puisqu'elles proviennent des systèmes de paie des entreprises : les entreprises sont contraintes légalement de les remplir et en ont elles-mêmes besoin pour leur bon fonctionnement. La qualité des données a cependant varié dans le temps : la première année de fonctionnement, en 2017, les données étaient très souvent incohérentes et nous avons donc choisir d'exclure cette année des analyses. De même, les entreprises de petite taille semblent avoir des problèmes de déclaration et nous avons décidé d'exclure toutes les entreprises de moins de 50 salarié. Ceci peut s'expliquer par le fait que, dans ces entreprises, il n'y ait pas de personnel dédié à la gestion de la paie.

Un autre défaut de ces données et l'absence d'informations relatives aux congés des salariés : en général, un salarié en congé n'a pas la possibilité d'être en arrêt maladie et il s'agit donc d'une variable très importante pour évaluer les arrêts maladie des entreprises. Plus généralement, ces données administratives sont très pauvres en termes de variables explicatives.

2.3.3 Typologie de trajectoires

L'objectif de cette section est d'identifier les salariés à risque d'arrêts maladie longue durée. Les données utilisées sont très pauvres en variables explicatives et, comme nous l'avons montré auparavant, les variables sociodémographiques sont très insuffisantes pour identifier des salariés à risque. Nous allons donc devoir identifier des signaux faibles en travaillant les données différemment. Nous avons décidé, dans un premier temps, d'analyser les trajectoires d'absence des salariés afin d'observer si certaines trajectoires mènent avec une probabilité plus grande vers des arrêts de longue durée.

2.3.3.1 Méthode : analyse de séquences

Afin d'évaluer si des trajectoires d'arrêts de travail pourraient mener plus certainement que d'autres vers des arrêts de longue durée, nous allons procéder en trois étapes :

Construction des trajectoires d'arrêt de travail Pour chaque salarié, nous construisons une trajectoire d'absence/présence d'un an. Une trajectoire est définie, dans notre analyse, comme une succession d'état *absence* et *présence*, chaque état représentant la situation hebdomadaire du salarié. La situation hebdomadaire est définie comme *absence* si le salarié est absent au moins un jour de la semaine ou comme *présent* sinon.

Pour chaque salarié, une fenêtre aléatoire d'un an est sélectionnée sur sa période de suivi dans la base de données.

Construction d'une typologie de trajectoires Pour l'ensemble des salariés ayant eu un arrêt de plus d'un mois dans les trois mois suivant la fin de l'année suivie (et donc de la trajectoire construite), nous allons effectuer une classification des trajectoires. Notre objectif est ainsi d'identifier des motifs récurrent de trajectoires d'absence qui pourraient mener vers un arrêt de longue durée.

Afin de construire cette typologie de trajectoire, une matrice de distance entre chaque trajectoire va être calculée.

La distance choisie est la distance *LCS* pour *Longest Common Subsequence* [45, 46]. Nous avons choisi cette méthode puisqu'elle permet de conserver la notion de séquence qui est importante dans nos données (deux semaines d'arrêts de travail consécutives n'ont pas le même sens que deux semaines étalées entre le début de l'année et la fin de l'année) mais aussi parce qu'elle est plus flexible en détectant des similarités même s'il n'y a pas un parfait alignement entre les deux séries.

Une classification ascendante hiérarchique par la méthode de Ward [47] sera ensuite effectuée afin de regrouper les différentes trajectoires dans une typologie que nous choisirons selon deux critères. Le premier critère est un critère de distance : la distance entre les deux dernières classes qui ont été fusionnées ne doit pas être trop grande. Le second critère est un critère d'interprétabiltié : le nombre de classe doit être raisonnable pour pouvoir les définir et le nombre de salariés par classe doit être assez élevé.

Les analyses sont effectuées en utilisant le package *TraMineR* sur R [45] pour la construction de la matrice de distance et la visualisation des données et en utilisant le package *cluster* [48] pour la construction de la typologie. Le calcul de la matrice de distance étant très long, nous avons décidé d'effectuer ces analyses sur un échantillon de 2400 salariés tirés aléatoirement.

Evaluation de la capacité prédictive de la typologie Afin d'évaluer la pertinence de cette typologie de trajectoire, nous allons évaluer la puissance de prédiction d'un modèle en introduisant cette variable. Nous allons attribuer à un ensemble de salariés, ayant des absences de longue durée ou non, ces catégories d'absence. Nous tenterons ensuite de prédire l'occurrence d'arrêt de travail de ces salariés en utilisant un modèle de forêt aléatoire, comme dans la section précédente, en ajustant sur ces catégories d'absence mais aussi sur des critères socioéconomiques identifiées dans les DSN (âge, sexe, rémunération, catégorie socioprofessionnel et secteur d'activité de l'entreprise). Nous évaluerons la capacité prédictive du modèle avec et sans les catégories de trajectoire grâce au critère de l'AUC, présenté précédemment.

Pour faciliter les calculs, nous travaillerons à nouveau sur des échantillons de population en utilisant les 2400 salariés avec un arrêt long sélectionné précédemment et un échantillon de 6000 salariés sans arrêt long.

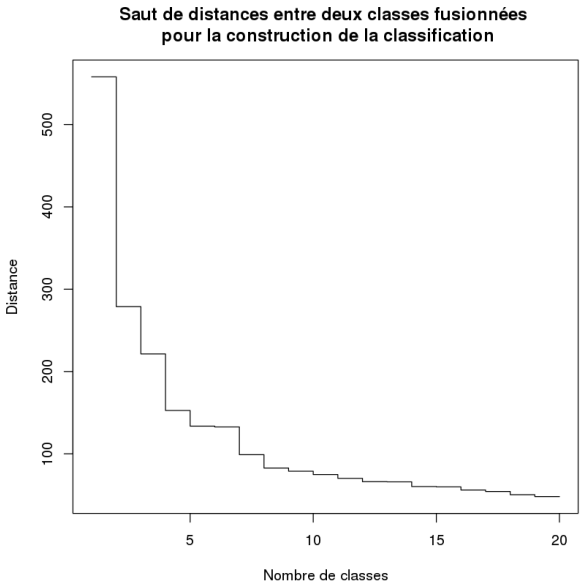

FIGURE 2.2 – Saut de distances entre deux classes fusionnées pour la construction de la classification.

2.3.3.2 Résultats

Typologie de trajectoire Nous avons construit une typologie de 4 catégories de trajectoire d'arrêt de travail à partir de cette classification. La distance induite par la fusion de deux de ces classes étaient en effet trop grandes comme cela peut être lu en Figure 2.2. Nous aurions aussi pu choisir une classification en 8 classes selon ce critère mais cela aurait rendu plus difficile l'interprétation et, comme nous le verrons plus tard, certaines catégories de trajectoires décrivent un nombre déjà faible de salariés.

Des exemples de salariés de chacune des quatre catégories sont présentées en Figure 2.3. Les quatre classes peuvent être décrites ainsi :

1. Le premier groupe rassemble 88% des salariés avec des arrêts de longue de durée et ce sont des salariés qui n'ont presque pas d'arrêt l'année précédent leur arrêt long. Leur arrêt de travail long pourrait ainsi être défini comme *accidentel* puisqu'il n'y a pas de signes précurseurs. Ils représentent 96,4% de la population générale.

2. Le second groupe rassemble 7,5% des salariés avec un arrêt long et décrit des salariés qui ont un arrêt maladie dans les quelques mois précédents l'arrêt long. On pourrait définir ces salariés comme les salariés ayant eu des *signes avant-coureurs* d'un arrêt long. Ils représentent 1,4% de

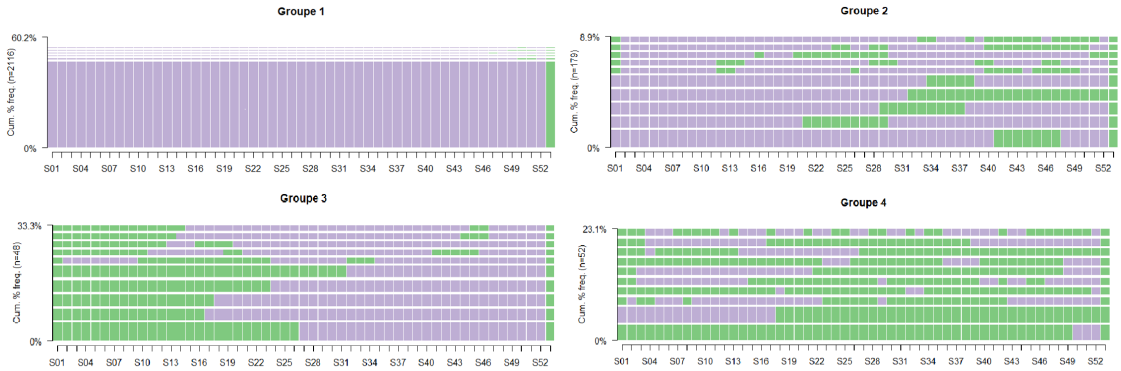

FIGURE 2.3 – Typologie de trajectoires

la population générale.

3. 2% des salariés sont présents dans la troisième catégorie qui ressemble des salariés qui ont déjà eu un arrêt de longue durée de plusieurs mois avant leur nouvel arrêt de longue durée. Ce sont des arrêts qui ont une rechute d'arrêt maladie de longue durée. Il représente 1,4% des salariés en population générale.
4. Enfin, 2,2% des salariés avec un arrêt long sont présents dans la dernière catégorie. Ce sont des salariés qui sont très souvenr en arrêt pendant l'année et l'on pourrait ainsi les définir comme des absents *chroniques*. Ils représentent 0,5% de la population générale.

Ces différentes trajectoires montrent des motifs récurrent d'absence parmi les salariés ayant des arrêts graves. Cependant, la première catégorie, qui représente 88% de la population générale, montre que même les arrêts de longue durée semblent arriver de manière accidentelle et sans signe avant-coureur.

Puissance prédictive de la typologie Nous avons construits deux forêts aléatoires : une forêt aléatoire avec et une forêt sans les trajectoires d'absence. L'introduction des catégories de trajectoires permettent d'augmenter l'AUC du modèle de 0.644 à 0.651 ce qui est une puissance prédictive assez pauvre. Pour rappel, le modèle basé sur les données d'enquête permettait d'obtenir un AUC de 0.82.

2.3.3.3 Discussion

La description des trajectoires d'absence a permis d'identifier diverses trajectoires typiques qui peuvent mener vers un arrêt grave. Ces travaux ont surtout permis de mettre en lumière que la

plupart des arrêts graves n'étaient pas précédés de trajectoires bien identifiables (ou, tout du moins, que cette méthode d'analyse ne permettait pas de les identifier) : la plupart des salariés ayant un arrêt de longue durée n'ont pas vraiment d'événement précurseur bien que, dans un certain nombre de cas, certaines trajectoires peuvent se dessiner.

La méthode utilisée apporte quelques éléments d'interprétations mais est peut-être trop restrictive pour notre analyse : en effet, la méthode nous contraint à uniquement regarder une année d'absence fixée alors que nos données sont parfois tronquées et que nous n'avons pas assez d'historique (ou bien parfois un historique trop grand dont on ne profite pas assez).

2.3.4 Modèle multi-états

Nous allons, dans une seconde analyse, explorer plus profondément l'hypothèse selon laquelle des signaux faibles pourraient être identifiées dans les trajectoires d'absence des salariés. Nous allons construire un modèle multi-état permettant d'évaluer le risque des salariés d'avoir un nouvel arrêt selon ses caractéristiques sociodémographiques et son historique d'absence.

Ces travaux ont été présentés en juillet 2018 à l'International Biometric Conference à Barcelone [49]. Bien qu'ils restent à l'état d'ébauche, il nous semble important de les présenter puisqu'ils nous donnent quelques informations complémentaires pour la compréhension des arrêts maladie.

2.3.4.1 Méthodes et données

Notre objectif est de pouvoir modéliser la trajectoire d'arrêts maladie des salariés pour estimer leur probabilité d'avoir un nouvel arrêt et le temps passé en arrêt dans une période donnée. Ce modèle pourrait ensuite servir, à l'échelle de l'entreprise, pour évaluer le nombre attendu de salarié en arrêt dans le futur.

Méthode Pour répondre à cet objectif, nous avons développé un modèle multi-état afin d'évaluer la distribution des probabilités transition entre un état "au travail" et un état "en arrêt maladie". Nous proposons de construire un modèle multi-état de type semi-Markov. Ces types de modèle permettent une estimation plus simple des probabilités de transition grâce à la propriété de Markov. Cette propriété pose que la distribution de probabilité des états futurs, sachant le passé et le présent, ne dépend que de l'état présent et pas des passés.

Cette hypothèse est très restrictive dans notre cas puisque la probabilité d'avoir un arrêt maladie est fortement augmentée si le salarié en a déjà eu un dans le passé [50]. Pour contourner ce problème tout en conservant la propriété markovienne, nous avons décidé de multiplier le nombre d'états "au travail" et "en arrêt maladie" afin de modéliser la probabilité d'avoir un nouvel arrêt (ou de revenir au travail) différente en fonction du passé.

Nous procéderons ainsi pour construire le modèle :

1. Pour choisir le nombre d'états dans le modèle, nous calculons l'estimateur de Kaplan-Meier de la durée de séjour dans les états *Travail* et *Arrêt Maladie* en fonction du nombre N d'arrêts qui ont eu lieu dans le passé. Pour $N > 0$, nous testons si la distribution est différente entre l'estimateur de Kaplan-Meier à N et N+1 selon un test log-rank. Tant que le test est significatif à 1%, nous continuons à augmenter N. La première valeur N qui n'est pas associée à un test significatif donnera le nombre d'états *Arrêt Maladie* et *Travail*. Un état absorbant est aussi introduit dans le modèle pour prendre en compte les départs de l'entreprise.
2. Nous choisissons un modèle paramétrique parmi certaines distributions (Weibull, Exponentielle et *Exponentiated Weibull*) pour modéliser le risque de base de chacune des transitions et nous la choisissons selon le critère d'Aikake ;
3. On ajuste les probabilité de transitions sur différentes covariables grâce à un modèle de Cox ;
4. On évalue ensuite les qualités d'ajustement du modèle. Notamment, notre objectif est de pouvoir estimer des indicateurs d'absence pour certaines entreprises et nous allons donc mesurer si le modèle permet de prédire des fréquences et durée d'absence d'une population donnée.

Analyser les données d'absence par des modèles multi-états n'est pas une nouveauté et quelques analyses ont été identifiées dans la littérature [51, 52, 53, 54]. Nous souhaitons cependant évaluer notre modèle différemment de ces études puisque notre objectif est certes d'expliquer les arrêts maladie mais aussi de pouvoir anticiper le nombre de salariés absents. Nos analyses sont effectuées sur *R* et principalement en utilisant le package *SemiMarkov* [55].

Données Comme expliqué précédemment, les travaux menés sur les trajectoires d'arrêt maladie se justifient par le faible volume de variables explicatives dans nos données et la nécessité d'exploiter les données le plus exhaustivement possibles pour y trouver de l'information. Nous allons utiliser dans ces analyses non pas les DSN mais les DADS (Déclarations Annuelles de Données Sociales) puisque

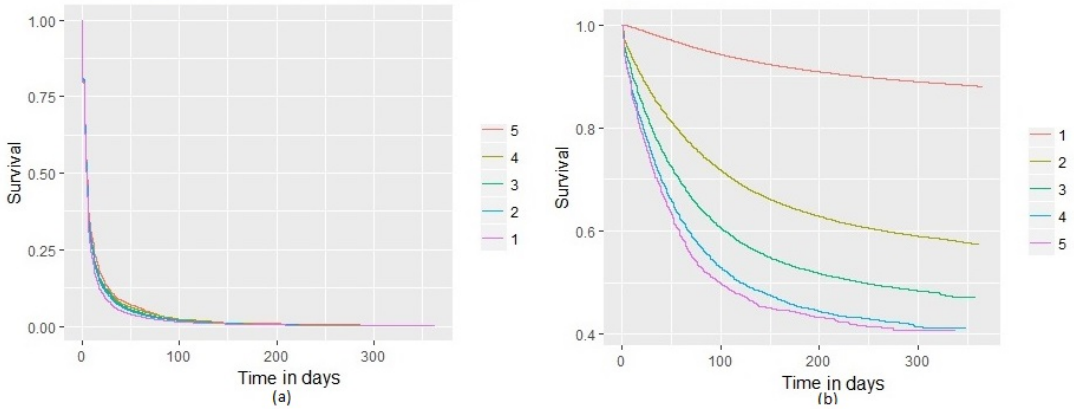

FIGURE 2.4 – Courbes de Kaplan Meier de la durée de séjour en état (a) arrêt maladie ou (b) travail selon le nombre de passage dans des états similaires

ces travaux ont été engagés en 2018, année où les DSN ont été mises en place. Les DADS sont les ancêtres des DSN et sont des données avec la même structure mais qui étaient déclarés annuellement (au lieu de mensuellement aujourd'hui). Comme elles sont anciennes, elles permettent d'atteindre un grand historique de données.

Nous utilisons ici un échantillon de salariés qui sont entrés dans la base entre 2010 et 2017 et qui ont au moins une année entière de suivi. Cela représente 842 527 salariés. Pour ajuster les modèles sur quelques covariables : le sexe des salariés, leurs âges et leurs catégories socioprofessionnelles.

2.3.4.2 Résultats

Etape 1 : choix des états La Figure 2.4 montre les courbes de Kaplan-Meier de la durée de séjour dans les états *Arrêts Maladie* ou *Travail* selon le nombre de passages dans ces états. Les courbes du temps passé en arrêt maladie semblent se superposer : les arrêts maladie sont en général relativement courts (le premier arrêt maladie dure en moyenne 9.8 jours). Au contraire, le temps passé dans l'état *Travail* diffère énormément en fonction du nombre d'arrêts passés du salarié : graphiquement, on peut constater que plus le salarié a eu d'arrêts maladie, plus la probabilité d'en avoir un nouveau augmente. La courbe de Kaplan-Meier qui décrit les salariés lorsqu'ils n'ont jamais eu d'arrêts maladie est intéressante puisqu'elle montre que très peu de salariés ont un arrêt maladie dans la première année.

La Table 2.3 présente les p-values des tests Log-Rank calculés pour déterminés le nombre optimal

Transition Travail (T) à Arrêt Maladie (AM)	
Transition	**Log-Rank**
T0>M1 vs T1>AM2	<E-12
T1>AM2 vs T2>AM3	<E-12
T2>AM3 vs T3>AM4	<E-12
Transition Arrêt Maladie (AM) à Travail (T)	
Transition	**Log-Rank**
AM1>T1 vs AM2>T2	<E-12
AM2>T2 vs AM3>T3	0,125
AM3>T3 vs AM4>T4	0,297

TABLE 2.3 – Tests Log-Rank entre les transitions afin d'identifier le nombre optimal d'états

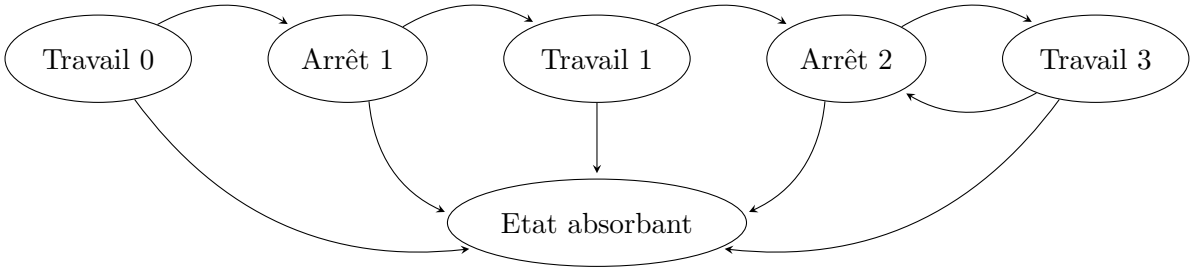

FIGURE 2.5 – Modèle multi-état sélectionné

d'état. Les distributions de la durée séjour dans l'état *Travail* sont toujours significativement différentes : le fait d'avoir eu un arrêt de travail semble augmenter la probabilité d'en avoir un nouveau. Les distributions de la durée de séjour dans l'état *Arrêt Maladie* est significativement différente entre le premier arrêt maladie et le deuxième arrêt maladie. Ensuite, la différence n'est plus significative.

Les résultats des tests Log-Rank laissent à penser que l'on pourrait multiplier à l'infini les états *Travail* pour pouvoir bien s'ajuster à nos données. Cependant, par soucis de simplicité et de lisibilité, nous allons réduire le nombre d'états Arrêt Maladie à deux états, comme cela nous est suggéré par les tests Log-Rank : la durée de séjour entre les deuxièmes et troisièmes arrêts semble similaire et les états ne méritent pas d'être divisés. Nous allons donc construire un modèle avec trois états *Travail* et deux états *Arrêts Maladie*. Le modèle est représenté graphiquement par la Figure 2.5.

Etape 2 - choix des modèles paramétriques : La Table 2.4 présente le critère d'Aikake pour chacune des distributions paramètriques ajustées à nos données. La distribution la plus pertinente est, pour l'ensemble des transitions, la distribution *Exponentiated Weibull* qui est la distribution associé

Transition	**Exponentiel**	**Weibull**	***Exponentiated Weibull***
T0>AM1	1856448	1844433	1837269
AM1>T1	688810	673787	646663
T1>AM2	522987	510322	507665
AM2>T2	424501	414938	400242
T2>AM2	338369	329048	326615

TABLE 2.4 – Critères d'Aikake pour les choix de distribution paramétrique pour l'ensemble des transitions du modèle

au critère d'Aikake le plus faible. Seules ces trois distributions ont été testées puisqu'elles étaient les seules proposées dans l'outil que nous avons utilisés [55].

Etape 3 - modèles de Cox : Le risque de base étant estimé par un modèle *Exponentiated Weibull*, nous allons maintenant explorer l'impact des covariables grâce à un modèle de Cox. Les Hazard Ratios (HR) des modèles de Cox peuvent être lus en Table 2.5. Premièrement, nous pouvons constater que certains facteurs augmentent fortement le risque d'arrêt maladie : être un femme (HR de 4.8 pour le premier arrêt), être âgé de moins de 35 ans ou être ouvrier. Ces différences s'estompent entre ces populations lorsque les salariés ont déjà eu des arrêts maladie : le HR d'être une femme n'est plus que de 2.88 lorsque les salariés ont déjà eu un arrêt et remonte à 4.13 lorsque les salariés ont déjà eu au moins deux arrêts. De même, les différences sont moins prononcés entre les CSP et les classes d'âges.

Le fait que les salariés plus âgés est un risque d'arrêt plus faible peut s'expliquer de deux façons. Premièrement, la construction de la base de données implique un biais dans les analyses : nous avons en effet uniquement conservé les salariés qui entraient dans la base de données, ce qui signifie qu'il commençait un nouveau contrat. Les salariés plus âgés qui commencent un nouvel emploi ne sont pas représentatifs de la population des salariés plus âgés : nous pouvons en effet imaginer que cette population est en meilleure santé que la population générale. La deuxième explication est, qu'en effet, les salariés plus âgés ont moins d'arrêt maladie que le reste de la population : nous avons en effet montré dans notre hiérarchisation des arrêts maladie [19] que les salariés les plus âgés avaient une probabilité plus faible d'avoir un arrêt de courte durée.

Les résultats des transitions *retour au travail* conforte cette intuition puisque l'on observe que les salariés les plus âgés ont un risque plus grand d'avoir des arrêts maladie plus long (HR 0.58 pour le premier arrêt puis HR de 0.73 pour les arrêts suivants). Les femmes ont des arrêts plus courts (HR 1.09

	Transition				
	T0>AM1	**AM1>T1**	**T1>AM2**	**AM2>T2**	**T2>AM2**
Sexe					
Homme	ref.	ref.	ref.	ref.	ref.
Femme	4,8***	1,09***	2,88**	1,06***	4,13***
Age					
<= 35 ans	ref.	ref.	ref.	ref.	ref.
35-50 ans	0,38***	0,72***	0,57***	0,80***	0,49***
> 50 ans	0,34***	0,58***	0,59***	0,73***	0,55***
CSP					
Ouvrier	ref.	ref.	ref.	ref.	ref.
Employé	0,10***	1,07***	0,21***	0,92***	0,12***
Profession intermédiaire	0,12***	1,23***	0,13***	1,04**	0,15***
Cadre	0,09***	1,11.	0,21***	0,93***	0,14***

*** p-value<0,00001, . 0,05<p-value<0,10

TABLE 2.5 – Hazard Ratio des modèles de Cox modélisant chaque transition

pour le premier arrêt maladie) tous comme les professions intermédiaires (HR 1.23 pour le premier arrêt maladie). Les différences s'estompent à nouveau lorsque les salariés ont déjà eu un arrêt maladie sauf pour la variable des catégories sociodémographiques qui présente un résultat intéressant. En effet, alors que les employés et cadres avaient des arrêts moins longs que les ouvriers lors du premier arrêt maladie (HR respectivement de 1.07 et 1.11), ce résultat s'inverse lorsque les salariés ont déjà eu au moins un arrêt (HR de 0.92 et de 0.93 respectivement). Nous pourrions interpréter ces résultats ainsi : les ouvriers ont des arrêts plus fréquents mais moins longs ; les employés et cadres ont des arrêts moins fréquents mais, lorsqu'ils sont récurrent, peuvent être des arrêts graves et donc plus longs.

En résumé, ces résultats montrent des différences de comportement selon des variables sociodémograpiques et montrent surtout que les trajectoires d'arrêts sont des déterminants des arrêts futurs.

Etape 4 - Evaluation du modèle Pour évaluer la pertinence du modèle en termes d'ajustement, nous avons décidé de comparer empiriquement les résultats obtenus par le modèle aux indicateurs directement évalués sur les données. Le premier critère est la fréquence moyenne d'arrêt maladie par salarié. Nous observons 1,60 arrêt par salarié absent et le modèle en prédit 1,55. Le deuxième critère est la durée moyenne du premier arrêt maladie : nous observons 9,84 jours sur nos données et le modèle en prédit 8,97.

2.3.4.3 Discussion

Nous sommes conscients des limites du modèle que nous avons tout de même décidé de présenter car il présente des avantages en termes d'interprétation mais aussi de prédiction et que le travail sur ce modèle nous a permis de formuler quelques problèmes qui devront être adressés dans les futurs travaux.

Les limites du modèle sont en effet nombreuses. De premières limites concernent la base de données. Premièrement, pour simplifier l'analyse, nous avons décidé d'utiliser uniquement les nouveaux salariés. Les données des autres salariés sont en effet tronquées à gauche : nous ne savons pas leur historique d'arrêts maladie et ne savons donc pas dans quel état commence ces salariés. Une deuxième limite est que nous n'utilisons qu'un historique d'un an et qu'un historique plus long est nécessaire pour estimer plus précisément les transitions et principalement les transitions entre états les plus avancées du modèle.

D'autres limites concernent plutôt les choix méthodologiques de modélisation et d'évaluation du modèle. Premièrement, l'utilisation de modèle paramétrique pour modéliser est potentiellement trop restrictive : le risque de base que nous introduisons dans le modèle de Cox est dérivé d'un modèle *Exponentiated Weibull* qui n'est peut-être pas très pertinent compte tenu de nos données. premièrement, nous observons un effet plateau et donc peut-être un biais de "survivant de long-terme" dans le premier état (*Travail 0*) [56]. Ensuite, pour les états *Travail* suivant, nous observons un risque de rechute très rapide d'arrêt maladie mais qui se tasse dans le temps et nous pouvons imaginer que, passer une période à définir, le risque d'arrêt maladie redevient similaire au cas où le salarié n'avait jamais eu d'arrêt maladie (ce qui n'est pas pris en compte dans le modèle). De même, le risque d'arrêt maladie est saisonnier et ceci n'est pas pris en compte dans notre modèle. Enfin, la dernière grande limite concerne l'évaluation du modèle : nous n'avons évalué le modèle que partiellement grâce à deux critères, importants pour les projections d'absence que nous pourrions proposer. Cependant, pour évaluer plus sérieusement ce modèle, il aurait fallu : **(1)** utiliser des données qui ne sont pas utilisées pour l'apprentissage du modèle, **(2)** évaluer la pertinence des résultats à l'échelle des des salariés dans une entreprise et non pas à l'ensemble des salariés de la base et **(3)** évaluer des indicateurs plus nombreux.

Notre modèle présente cependant de nombreux avantages. Premièrement, il permet de formuler des interprétations simples et très informatives. Il pourrait ainsi permettre d'évaluer la pertinence de

services dont l'objectif serait de réduire le volume d'arrêt maladie. Ensuite, il permet de faire des projections en termes des nombreux indicateurs d'absence : proportion d'absence, durée des absence, fréquence des absences. Enfin, bien que nous l'ayons peu abordé ici, les modèles multi-états peuvent fournir des graphiques claires sous forme de trace de Markov qui peuvent être pris en main par tous : ces graphiques représentent la proportion de salarié attendu par état. Le modèle multi-état est un modèle très flexible qui permet d'étudier les arrêts maladie dans l'ensemble de leurs dimensions.

2.4 Discussion générale

Les analyses développées dans cette première partie nous ont permis de développer deux niveaux de réflexion.

2.4.1 Identifier les déterminants des arrêts maladie

La première réflexion concerne les déterminants des arrêts de travail et les méthodes statistiques qui pourraient permettre de les identifier et de les hiérarchiser.

Nous avons dans un premier temps identifier les indicateurs d'absence d'intérêt et les méthodes associées. Nous avons principalement travaillé sur la durée des arrêts de travail et principalement sur les arrêts longs, qui sont les plus graves pour les salariés mais aussi parce que nous étions contraints par nos données.

Nous avons utilisé une méthode permettant de hiérarchiser de nombreux déterminants de différentes catégories : des déterminants individuels, professionnels et exogène. Les déterminants des arrêts de moins d'un mois sont assez flous et montrent que la cause de ces arrêts semblent principalement exogènes. Les arrêts longs, au contraire, semblent plus facilement explicables : les causes proviennent de la santé individuelle mais aussi du contexte professionnel lié à la pénibilité physique mais aussi psychologique du travail. Ces résultats ne sont pas révolutionnaires (un arrêt maladie s'explique parce que l'on est malade) mais ont permis de démontrer l'utilité de plan d'action sur ces trois points d'intérêt. Cette méthode pourrait aussi être réutilisé si des entreprises souhaitent mener des enquêtes au sein de leurs entreprises pour hiérarchiser leurs propres facteurs de risque.

2.4.2 Analyser les trajectoires d'absence

La seconde réflexion a permis d'étendre notre analyse à une vision plus générale des arrêts de travail : les trajectoires d'absence qui incluent aussi bien les indicateurs de durée que les indicateurs de fréquence.

Ces travaux ont aussi pour objectif d'identifier les salariés qui ont le plus de risque d'avoir des arrêts de travail grave et donc de proposer des plans d'action à ces salariés en priorité. Ces salariés doivent être simple à identifier et nous devons donc utiliser des données faciles d'accès, contrairement aux données d'enquête qui demandent du temps et de l'argent. Nous utilisons alors les données administratives d'arrêt de travail qui sont rendus disponibles par toutes les entreprises.

Les travaux menés ont permis d'identifier des signaux faibles qui risquent de mener vers des arrêts graves : le principal prédicteur des arrêts de travail s'avèrent être le fait d'avoir déjà eu des arrêts de travail. Une haute fréquence d'arrêt va donc mener à plus d'arrêt mais aussi à des arrêts plus graves que la moyenne.

2.4.3 Vers un modèle de prédiction des arrêts maladie ?

Une des demandes soulevée lors de l'élaboration du projet de thèse était de pouvoir développer un moteur de prédiction des arrêts de travail afin de pouvoir aider les entreprises à s'organiser mais aussi à identifier les salariés qui auraient le plus besoin de services pour les aider à lutter contre les causes de leur absentéisme.

Comme nous pouvons en avoir l'intuition maintenant, ce travail semble délicat lorsque nous étudions les données au niveau de l'individu. Même avec des données d'enquête très précises sur les nombreux déterminants d'absence auxquels peuvent être exposés les salariés, la prédiction des arrêts de travail est très complexe. Lorsque nous observons les trajectoires d'arrêts de travail, peu de motifs peuvent être identifiés et la plupart des arrêts semblent accidentels. La prédiction des arrêts de travail, au niveau individuel, semble donc très complexe, sinon impossible avec les données en notre possession.

Chapitre 3

Détecter les excès d'arrêt maladie

Contenu

3.1 Introduction : surveiller l'absentéisme des entreprises, comment et pourquoi ?

Nous venons de montrer que les arrêts de travail sont causés par des facteurs très divers et les comportements des salariés sont très hétérogènes. La compréhension des déterminants des arrêts de travail dans une entreprise et l'identification des salariés à risque s'avèrent être un travail très complexe et coûteux puisqu'il faut récolter des données relatives aux arrêts de travail mais aussi des informations plus précises sur l'activité et la santé des salariés.

Le monitoring des arrêts maladies dans les entreprises pourrait débuter par une étape préalable plus simple et moins coûteuse : l'analyse des données d'absence au sein de l'entreprise. Ces données sont disponibles directement puisqu'elles sont entrées nécessairement dans le système de gestion de la paie pour un transfert aux organismes d'intérêt (ministère du Travail et institut de prévoyance notamment) et sont porteuses d'informations précieuses.

Ces données pourraient permettre d'évaluer simplement le niveau d'absence de l'entreprise en se comparant à des valeurs repère et de suivre l'absentéisme dans le temps. La surveillance a donc ici un objectif double :

1. on cherche à voir si, structurellement, l'absentéisme de l'entreprise est anormalement haut ;
2. on cherche à identifier si, à un moment donné, le taux d'arrêt maladie (ou tout autre indicateur) est excessivement haut par rapport à la normal.

L'identification d'une anomalie peut, dans les deux cas, s'expliquer par un phénomène qui agit dans l'entreprise. Les causes d'arrêts maladie peuvent en effet provenir d'éléments présents dans l'entreprise et donc de problèmes organisationnelles comme de mauvaises pratiques managériales ou une mauvaise gestion de la pénibilité physique.

La premier niveau de surveillance (détecter si l'absentéisme structurel de l'entreprise est anormalement haut) est délicat puisque chaque entreprise à des comportements d'absentéisme différents qui peuvent provenir des caractéristiques sociodémographiques des salariés mais aussi de différentes réglementations en termes d'arrêts de travail (présence ou non de jours de carence, taux d'indemnisation des arrêts de travail) ou de culture d'entreprise (la prise d'arrêts de travail peut être parfois plus ou moins découragée par certaines entreprises). La comparaison de valeurs d'absentéisme à des

valeurs-étalon doit ainsi être pratiquée avec prudence.

Le second niveau de surveillance apporte des informations plus tangibles. En effet, si nous observons, dans le temps, un moment où l'absentéisme d'une entreprise est excessivement élevé, nous pouvons conclure à une anomalie dans l'entreprise. Nous pouvons ainsi procéder à des investigations rapides et identifier les problèmes. Ceci permettrait ainsi de concentrer les investigations sur les entreprises ou les entités de l'entreprise le nécessitant le plus.

Dans ce chapitre, nous présenterons dans une brève première partie quelques éléments de bonne pratique pour comparer son niveau d'absentéisme à des valeurs repère. Nous présenterons ensuite en détails un algorithme de surveillance des arrêts de travail inspiré de méthodes de surveillance épidémiologique. Une application sur des données d'arrêt de travail pendant la pandémie de la Covid-19 sera présentée. Nous présenterons ensuite une autre application de la surveillance des arrêts maladie pour la détection d'épidémie de grippe. Nous concluerons enfin ce chapitre par une discussion sur l'intérêt des méthodes élaborées.

3.2 Comparer l'absentéisme des entreprises : bonnes pratiques

Le premier niveau d'étude provient d'une question pratique très simple : comment l'entreprise doit-elle interpréter ses indicateurs d'arrêt de travail ? Les entreprises consultent en effet régulièrement des tableaux de bord présentant divers indicateurs d'absentéisme agrégés [57, 58] : taux d'absentéisme, fréquence des arrêts, durée des arrêts, *etc.*

L'interprétation de ces valeurs est délicate et peut mener à des contre-sens. L'étude de ces tableaux de bord et de cette problématique nous a permis d'approfondir le sujet et d'identifier quelques bonnes pratiques que nous allons présenter. Cette étape préalable nous permettra, de plus, de présenter en détails les données qui seront utilisés pour le second modèle de surveillance.

3.2.1 Que mettre dans un tableau de bord de l'absentéisme ?

Un tableau de bord d'absentéisme est un outil permettant à l'entreprise de comprendre l'absentéisme de son entreprise. Pour comprendre et gérer l'absentéisme, quelques éléments doivent être impérativement présenter :

1. des indicateurs permettant une description presque exhaustive de l'absentéisme : la proportion

de salariés ayant été en arrêt pendant l'année, la durée moyenne des arrêts et le nombre d'arrêts par salarié semblent les indicateurs les plus importants.

2. des indicateurs permettant d'interpréter ces valeurs d'absentéisme, autrement dit un élément de comparaison. Nous allons y revenir tout de suite.

3. des indicateurs permettant d'évaluer l'évolution de l'absentéisme dans le temps : évolution annuelle des indicateurs, voire hebdomadaire pour le taux d'absentéisme afin d'observer la saisonnalité du phénomène. L'outil de surveillance que nous proposerons dans la section précédente permet de surveiller automatiquement ce phénomène.

Nous n'entrerons pas dans le détail du choix des indicateurs qui est pourtant une question très importante : les indicateurs doivent être choisis en fonction de l'objectif du tableau de bord. Si l'objectif du tableau de bord est sanitaire, des indicateurs de durée doivent être présentés : un arrêt long est très souvent causé par des problèmes graves de santé. Si l'objectif du tableau de bord est managérial, des indicateurs de fréquence doivent être présentés : des arrêts très fréquents sont problématiques pour l'organisation de l'entreprise et témoignent très souvent d'un problème organisationnel.

Revenons maintenant sur le deuxième élément du tableau de bord, il s'agit d'indicateurs construits pour interpréter les valeurs d'absence de l'entreprise. Il s'agit donc de valeurs auxquelles il est possible de se comparer et qui proviennent de données externes. Le choix de ces données est délicat puisqu'il faut choisir des données pertinentes. Se comparer à des indicateurs moyens nationaux est une solution simple mais cet exercice peut être dangereux : chaque entreprise a en effet une activité, une structure sociodémographique et même une culture différente qui peuvent influencer le niveau d'absence de l'entreprise. Ainsi, comparer le niveau d'absence d'une entreprise avec des salariés assez âgés et avec une activité manuelle très physique à un indicateur national n'aura pas vraiment de sens. Le niveau d'absence sera vraisemblablement plus haut dans l'entreprise cité puisque les salariés âgés avec des métiers manuels sont plus fragiles. Il ne faut donc pas tirer le signal d'alarme dans cette situation.

Pour évaluer si le niveau d'absence est vraiment trop élevé, il faudrait se comparer à des entreprises similaires : une comparaison par secteur d'activité et taille d'entreprise semble déjà plus précise mais demeure insuffisante.

3.2.2 Typologie d'entreprises pour des comparaisons pertinentes

Nous présentons maintenant un exemple de typologie d'entreprises construite afin de proposer des valeurs repères. Ces valeurs sont intégrées à un tableau de bord proposé par Malakoff Humanis.

3.2.2.1 Méthodes : arbre de décision

L'objectif de ce travail est de construire, pour chaque entreprise, une valeur repère afin d'évaluer son niveau d'absentéisme pour chacun de ses indicateurs. Cette valeur doit être facilement compréhensible pour le lecteur.

Cette valeur repère est basée sur une typologie d'entreprises dont le profil d'absence est similaire au regard de leurs caractéristiques.

La typologie est construite par un arbre de décision de type *CART*. Ces arbres permettent de regrouper des observations homogènes selon un critère en utilisant diverses variables explicatives. Dans notre cas, nous construisons des groupes de variables homogènes en termes d'absence (l'indicateur sélectionné est le nombre annuel de jours d'absence par ETP) selon les variables sociodémographiques de l'entreprise définies ci-dessous. Nous avons sélectionné le nombre de jour d'absence par ETP pour représenter l'absentéisme globale puisqu'il s'agit d'un indicateur composite, décrivant aussi bien la proportion de salariés absents que la durée des arrêts, ce qui représente assez bien le niveau d'absentéisme global d'une entreprise.

Les critères statistiques choisis pour la construction de l'arbre, et donc le nombre de classe dans cet arbre, sont les suivants :

— un premier élagage est effectué de manière objective : les feuilles de l'arbres doivent contenir plus de 100 entreprises et le paramètre de complexité de l'arbre est sélectionné par une règle du coude : il doit y avoir un gain marginal à construire un nouveau nœud qui est supérieur à 1% d'erreur.

— Un deuxième élagage plus subjectif peut ensuite être effectué : si l'arbre sélectionné a trop de feuilles (ou pas assez), l'arbre peut être ajusté afin d'obtenir une typologie claire et facilement interprétable. Des statistiques descriptives sont enfin effectuées pour définir les différentes catégories.

Les analyses sont effectuées sur R.

3.2.2.2 Données

Le tableau de bord est proposé aux entreprises de plus de 50 salariés affiliés chez Malakoff Humanis en utilisant les données d'arrêt de travail qui sont automatiquement transmises à l'assureur : les *Déclarations Sociales Nominatives* que nous avons déjà présentées en section 2.3.2.

Dans le cadre de l'analyse suivante, nous utilisons un périmètre de 2681 entreprises sélectionnées pour la qualité de leur données. Nous avons agrégé les données au niveau annuel et au niveau de l'entreprise. Ainsi, les variables suivantes sont utilisées pour construire la typologie :

— Proportion de femmes parmi les Emplois Temps Plein (ETPs) en 2018 ;
— Proportion de salariés de moins de 35 ans, entre 35 et 44 ans, de 45 à 54 ans, de plus de 55 ans parmi les ETPs en 2018 ;
— Proportions de cadre, de professions intermédiaires, d'ouvrier et d'employé parmi les ETPs en 2018 ;
— Nombre d'ETP en 2018 ;
— Secteur d'activité de l'entreprise (Industrie et BTP, Commerce, Service ou Santé)
— Proportion de CDI parmi les ETP en 2018.

3.2.2.3 Résultats

La typologie sélectionnée peut être lue en Figure 3.1. L'arbre se lit ainsi : à chaque feuille, la valeur se situant tout en haut représente le nombre de jour moyen par ETP dans les entreprises de cette classe ; les valeurs se situant en bas présentent respectivement le nombre et le pourcentage d'entreprises dans cette classe. Lorsqu'une feuille se scinde en deux, la règle de décision de cette scission se lit sous la feuille. Par exemple, pour la première scission de l'arbre : nous pouvons trouver à droite l'ensemble des entreprises dont plus de 21% des salariés sont cadre et à droite les autres entreprises.

La typologie finale décrit 8 segments d'entreprise définies par leur absentéisme et leurs caractéristiques sociodémographiques. Les caractéristiques précises des entreprises de ces groupes sont décrites en 3.1.

Nous allons présenter chacun des segments de gauche à droite.

1. Le premier segment est le segment des entreprises dont plus de 54% des salariés sont des cadres. Cela représente 19% des entreprises et ces salariés ont un niveau d'absence très bas : seulement

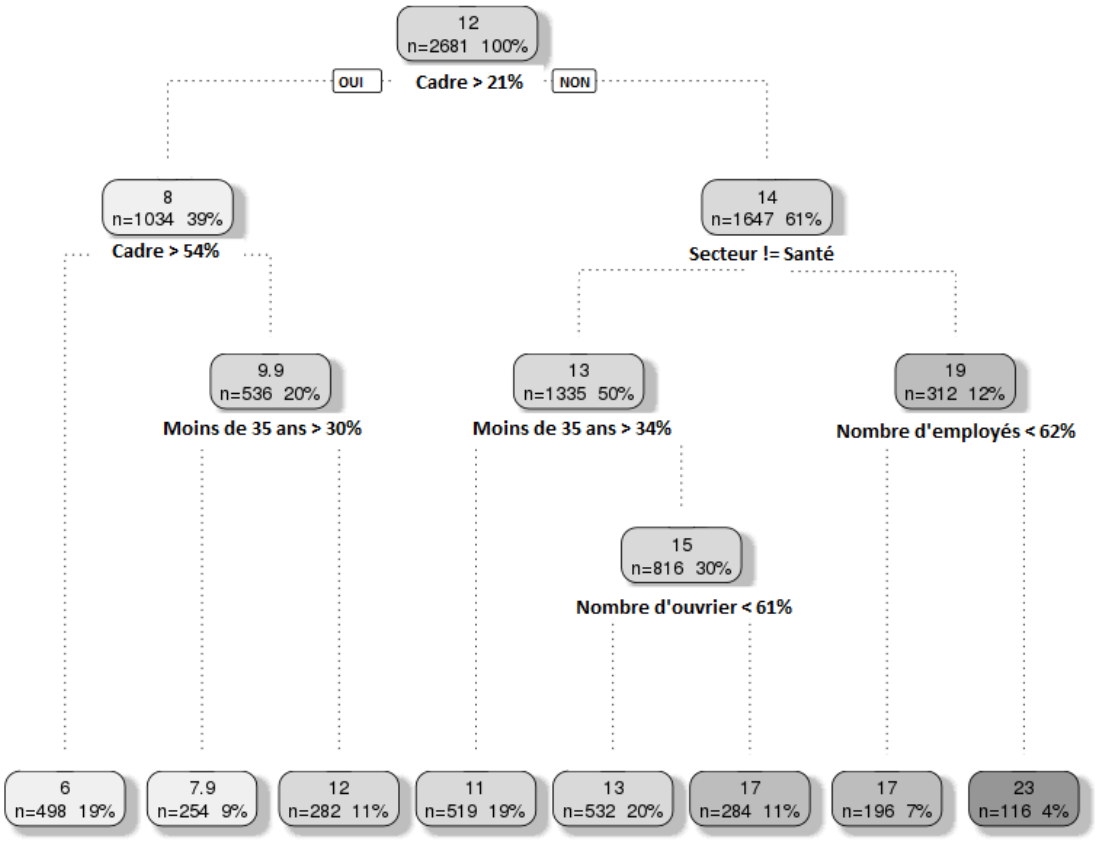

FIGURE 3.1 – Arbre *CART* présentant une typologie des entreprises en fonction de leurs caractéristiques d'arrêt de travail

6 jours d'absence par salarié en moyenne contre 12 jours pour l'ensemble des entreprises. Les salariés de ces entreprises sont jeunes (43% ont moins de 35 ans) et les entreprises viennent principalement du secteur des services. Il s'agit d'entreprise avec des salariés très diplômés comme certaines entreprises du conseil.

2. Le deuxième segment contient les entreprises qui ont entre 21% et 54% de salariés cadres et plus de 30% de salariés de moins de 35 ans. Les salariés de ces entreprises sont à nouveau moins absent que la moyenne mais plus que les entreprises du premier groupe (7,9 jours d'absence en moyenne par ETP). Ces entreprises emploient beaucoup de professions intermédiaires et beaucoup de salariés jeunes.
3. Le troisième segment contient à nouveau uniquement des entreprises qui ont entre 21% et 54% de salariés cadres smais moins de 30% de salariés de moins de 35 ans. Leur profil d'absentéisme est identique à celui de la moyenne (12 jours d'absence par ETP). A nouveau, on y retrouve beaucoup de professions intermédiaires parmi les salariés et ces entreprises proviennent beaucoup des secteurs du service et de l'industrie.
4. Le quatrième segment d'entreprise regroupe des entreprises qui ne proviennent pas du secteur de la santé qui emploient moins de 21% de cadres et plus de 34% des salariés de moins de 35 ans. Le niveau d'absence est légèrement plus bas que la moyenne (11 jours d'absence par ETP) et il s'agit un des segments qui emploient le plus de salariés en CDD. Beaucoup d'entreprises proviennent des secteurs du commerce et des services. Il peut s'agir d'entreprises recrutant beaucoup de saisonniers.
5. Le cinquième segment d'entreprise regroupe des entreprises qui ne proviennent pas du secteur de la santé et qui emploient moins de 21% de cadres, moins de 61% d'ouvriers et plus de 34% de salariés de moins de 35 ans. Le niveau d'absence est légèrement plus élevé que la moyenne (13 jours d'absence par ETP). Contrairement au segment précédent, les salariés en CDD sont moins nombreux et le secteur de l'industrie y est plus fortement représenté. Il s'agit de plus d'entreprises assez grandes (médiane de 110 salariés).
6. Le sixième segment regroupe des entreprises qui ne proviennent pas du secteur de la santé avec moins de 21% de cadres, plus de 61% d'ouvriers et plus de 34% des salariés de moins de 35 ans. ces entreprises sont très pénalisées en termes d'absentéisme puisque le nombre moyen de jours d'absence par salarié est de 17 jours. Les salariés de ce segment sont plutôt âgés et proviennent

Segment	1	2	3	4	5	6	7	8
n	498	254	282	519	532	284	196	116
Nombre de jour d'absence par ETP	6	7,9	12	11	13	17	17	23
Structure d'absence								
Prévalence d'absence	24,20%	27,70%	32,00%	28,90%	33,40%	36,40%	36,40%	36,90%
Nombre de jour d'absence par ETP absent	24,9	26,7	35,9	37,4	39,9	44,3	45,8	60,3
Nombre d'épisode par ETP absent	1,67	1,73	1,7	1,86	1,75	1,74	1,85	1,97
Sociodémographie								
Âge (% moyen)								
Moins de 35 ans	43,30%	47,10%	20,30%	52,00%	22,20%	21,70%	29,40%	35,20%
35 – 44 ans	28,10%	27,50%	28,70%	23,80%	28,40%	25,90%	25,70%	21,40%
45 – 54 ans	19,20%	18,10%	32,00%	17,20%	32,10%	34,40%	27,10%	27,00%
54 ans et plus	9,30%	7,20%	18,90%	7,00%	17,20%	17,90%	17,60%	16,30%
Hommes (% moyen)	60,60%	60,70%	57,10%	59,80%	56,60%	72,20%	27,10%	15,30%
CSP (% moyen)								
Cadres	77,40%	35,30%	32,90%	7,10%	10,30%	6,00%	6,00%	4,90%
Prof. Int.	11,20%	31,20%	31,20%	28,70%	35,50%	12,80%	50,30%	16,40%
Employés	8,20%	19,40%	17,00%	34,20%	27,00%	6,10%	29,60%	73,70%
Ouvriers	0,90%	9,40%	16,30%	23,90%	25,90%	74,40%	9,60%	4,10%
CDI (% moyen)	94%	91%	94%	84%	91,60%	94,30%	83,80%	78,50%
Secteur (% moyen)								
Commerce	5,00%	11,70%	15,00%	32,00%	22,00%	6,10%	0%	0%
Industrie et BTP	6,40%	22,30%	35,60%	25,90%	36,10%	88,90%	0%	0%
Service	86,40%	64,80%	39,90%	41,40%	41,90%	5,00%	0%	0%
Santé	2,20%	1,20%	9,60%	0%	0%	0%	100%	100%
Taille d'entreprise (médiane)	89,1	85,3	132,9	80,4	110,6	98	110,4	69

TABLE 3.1 – Description des entreprises selon leur segment d'absence. Les numéros des segments suivent l'ordre des segments de l'arbre en Figure 3.1

majoritairement de l'industrie et du BTP. Il s'agit majoritairement d'entreprise du bâtiment ou de l'industrie avec des emplois très manuels.

7. Le septième segment regroupe des entreprises qui proviennent du secteur de la santé avec moins de 21% de cadres et moins de 62% d'employés. Ces entreprises sont très pénalisées en termes d'absentéisme puisque le nombre moyen de jours d'absence par salarié est de 17 jours. Ces entreprises emploient beaucoup de professions intermédiaires et beaucoup de CDD. Il s'agit principalement d'entreprises du social ou de laboratoires ou cliniques employant des infirmières.
8. Le huitième segment regroupe des entreprises qui proviennent du secteur de la santé avec moins de 21% de cadres et plus de 62% d'employés. Ces entreprises sont les plus touchées puisque leurs salariés ont en moyenne 23 jours d'absence par an. Il s'agit d'entreprises où les salariés sont plutôt âgés et très nombreux en CDD. Il s'agit majoritairement d'EHPAD.

3.2.2.4 Discussion

La typologie construite permet donc de construire des groupes d'entreprise au profil d'activité assez similaire que l'on ne peut pas identfier en utilisant uniquement le secteur d'activité. Notamment,

cette typologie permet de scinder les entreprises de la santé très touchées par l'absentéisme (comme les EHPAD) d'entreprises de la santé à l'activité très différente (comme les entreprises de l'industrie pharmaceutique). Malgré le fait que le critère de choix soit le nombre de jours d'absence par ETP, les segments sont aussi assez homogènes en termes d'autres critères comme le nombre d'épisodes par ETP absent ou le nombre de jours d'absence par ETP absent. Cet outil permet ainsi de donner des valeurs de comparaison plus précises que des valeurs nationales ou uniquement segmentées par secteur. Cette typologie est aujourd'hui implémenté dans les tableaux de bord d'absentéisme proposé par Malakoff Humanis.

Cette typologie a des limites puisque ces catégories ne représente pas le comportement des entreprises françaises mais le comportement des entreprises assurées par Malakoff Humanis qui ne sont pas représentatives des entreprises françaises. Il faut de plus rester vigilant puisque de nombreux paramètres ne sont pas pris en compte dans cette typologie comme les politiques d'indemnisation des arrêts maladie ou tout simplement la culture de l'entreprise en termes d'arrêts de travail. Ainsi, le taux d'absence habituel des entreprises peut dévier de la valeur repère sans pourtant être le signe d'un mal-être dans l'entreprise. Pour identifier ces dérives, d'autres outils doivent être développés et c'est ce que nous allons faire dans la prochaine partie.

3.3 Surveillance des arrêts maladie

Nous présentons dans cette section un algorithme de surveillance développé pour le cas spécifique des arrêts de travail. Nous décrirons dans un premier temps les spécificités des données d'arrêts maladie avant d'effectuer, dans un deuxième temps, une revue des méthodes de surveillance statistique. Dans un troisième temps, nous présenterons l'algorithme développé qui a donné lieu à une publication scientifique.

3.3.1 Les arrêts maladie : des données spécifiques

Les arrêts maladie des entreprises ont une structure très proches des données utilisées en surveillance des maladies infectieuses. Les données d'absence sont en effet des données que l'on peut décrire sous forme de données de comptage (incidence ou prévalence d'absence hebdomadaire par exemple) que l'on va suivre dans le temps. Comme beaucoup de maladies infectieuses, les arrêts ma-

ladie sont un phénomène saisonnier [13] : le taux d'absence est élevé en hiver, lors des épidémies saisonnières, et faible en été, notamment à cause des vacances des salariés (qui ne prennent pas d'arrêt pendant ces saisons). Le nombre d'absents, au travail ou à l'école, a par ailleurs été étudié comme une donnée syndromique dans quelques modèles de surveillance afin de détecter les épidémies de grippe [59, 60, 61].

Les données d'arrêts de travail ont cependant des particularités qui ne peuvent convenir à l'ensemble des modèles qui s'offrent à nous. Les arrêts de travail sont, comme nous l'avons longuement expliqués précédemment, corrélés à un grand nombre de facteurs comme l'âge des salariés ou leur catégorie socioprofessionnelle. Un modèle identifiant des excès d'absence doit ainsi prendre en compte ces différentes caractéristiques. Par exemple, une entreprise employant principalement des salariés âgés avec des emplois manuels doit *a priori* avoir un niveau d'absence plus haut qu'une entreprise employant principalement des salariés jeunes travaillant derrière leur écran. Une situation d'excès d'absence devra donc être définie différemment dans ces deux cas.

De même, chaque entreprise peut avoir des *habitudes* différentes en termes d'arrêt maladie qui peuvent mener à des taux d'absence de base différents. Ceci peut s'expliquer par des différences culturelles (les managers peuvent plus ou moins dissuader aux salariés de s'arrêter) ou par des différences plus administratives (les salariés peuvent être indemnisés différemment selon leur entreprise ou leur secteur d'activité). L'objectif du modèle de surveillance souhaité n'est pas de détecter des différences structurelles entre les entreprises mais de détecter des sursauts d'absence ponctuels afin de permettre des actions rapides. La détection de différences structurelles pourrait être détectée grâce à la typologie présenté dans la partie précédente qui donne des niveaux d'absence moyens pour les entreprises en fonction de leur activité et de leur structure sociodémographique.

3.3.2 Etat de l'art des méthodes de surveillance statistique

La surveillance épidémiologique et statistique est un sujet vaste et de nombreux états de l'art ou analyses comparatives ont déjà été publiés [62, 63, 64, 65, 66, 67, 68, 69, 70, 71].

Notre référence principale est ici la publication de Unkel *et al.* [62] dont nous utilisons la classification des méthodes de surveillance. La publication de Unkel *et al.* n'est pas la seule revue de littérature sur le sujet mais il s'agit d'une revue relativement récente et, à notre connaissance, de la revue la plus complète.

Pour simplifier la revue, nous nous concentrons uniquement sur les modèles de détection 'épidémie (*outbreak detection* en anglais) appliquées à des données de comptage. Ces modèles ont pour objectif d'identifier des valeurs anormalement hautes ; il pourrait aussi être intéressant d'identifier les valeurs trop basses, mais ce n'est pas notre sujet dans ce mémoire. Nous n'allons pas entrer dans le détail de chacune de ces méthodes mais nous allons tenter d'en extraire les quelques spécificités afin d'en évaluer l'intérêt vis-à-vis du cas des arrêts maladie. Nous expliquerons tout de même dans le détail le modèle de Farrington [18] qui est le modèle que nous avons choisi d'adapter à notre contexte.

3.3.2.1 Méthodes de régression

Méthodes paramétriques

— *Modèle de Stroup et de Serfling*

Deux approches historiques et toujours utilisées sont les modèles de Stroup [72] et de Serfling [73].

Le premier modèle n'est pas à proprement parler une régression mais il inspirera les modèles suivants. Stroup observe en effet un comptage en semaines t-1, t et t+1 sur plusieurs années et en calcule la moyenne et l'écart-type. L'observation de ce comptage sur cette période de temps réduite permet d'ajuster sur la saisonnalité. Les épidémies sont ensuite identifiées graphiquement en faisant une hypothèse d'erreur gaussienne.

Le second modèle est basé sur une régression linéaire incorporant un intercept, une tendance et une saisonnalité grâce à des coefficients de Fourier :

$$\beta_0 + \eta t + \sum_{s=1}^{S} \{\gamma_s cos(\frac{2\pi s}{T}) + \alpha_s sin(\frac{2\pi s}{T}\}.$$

avec y_t le comptage à l'instant $t > 0$, β_0 l'intercept, η le coefficient associé à la tendance et $\forall s \in 1, ..., S$ et $S > 0$, γ_s et α_s les coefficients associés aux coefficients de Fourier. T correspond à l'échelle temporelle étudiée. Un seuil d'alerte est à nouveau construit grâce à une hypothèse de normalité des erreurs

Le modèle est simple et efficace et c'est pour cette raison qu'il est parfois toujours utilisé en pratique. Cette simplicité apporte cependant certaines limites puisque le modèle n'est pas parfaitement adapté aux données. Premièrement, la régression linéaire est adaptée à des données continues et non pas à des données de comptage. Pour palier ce problème, un modèle de Poisson peut par exemple être préféré [74].

Ensuite, les données étudiées incluent des épidémies qui se caractérisent par des comptages anor-

malement hauts : les inclure dans l'entraînement du modèle va donc mécaniquement augmenter le seuil d'alerte et donc dégrader la sensibilité du modèle. Ceci peut-être résolu de diverses façons et une solution a été apportée par le modèle que nous allons présenter ensuite.

— *Modèle de Farrington (et raffinements)*

Une extension du modèle de Serfling est le modèle de Farrington [18]. Ce modèle utilise une équation de régression similaire au modèle précédent mais lui préfère une régression Quasi-Poisson pour mieux s'adapter aux données de comptages qui présentent souvent une surdispersion. Farrington s'inspire de plus de Stroup pour l'introduction de la saisonnalité et introduit une procédure de pondération pour donner moins d'importance aux potentielles alertes du passé. Le modèle de Farrington considère l'équation suivante :

$$log(\mu_t) = \beta_0 + \eta t.$$

avec $\mu_t = E(y_t)$. Le modèle prend en compte la saisonnalité en s'entraînant uniquement sur les semaines similaires des années passées : il est préconisé d'utiliser les semaines t-3,t-2,...,t,...,t+3. Le paramètre de surdispersion du modèle ϕ est estimé comme suit :

$$\hat{\phi} = max\left\{\frac{1}{n-2}\sum_{i=1}^{n}\omega_i\frac{(y_i-\hat{\mu}_i)^2}{\hat{\mu}_i}, 1\right\}$$

ω_i est une fonction de poids défini ainsi :

$$\omega_i = \begin{cases}\gamma s_i^{-2} \text{ if } s_i > 1,\\ \gamma \text{ otherwise.}\end{cases}$$

γ est une constante telle $\sum_{i=1}^{n}\omega_i = n$ et $forall i \in 1, ...n$ sont des résidus d'Anscombe standardisés que l'on définit ainsi :

$$s_i = \frac{3}{2\hat{\phi}^{1/2}}\frac{y_i^{2/3}-\hat{\mu}_i^{2/3}}{\hat{\mu}_i^{1/6}(1-h_{ii})^{1/2}}$$

où les h_{ii} sont les éléments diagonaux de la matrice chapeau (en anglais, *hat matrix*). La fonction de poids donne donc un poids moindre aux observations avec un haut résidu : avoir un résidu haut signifie que l'observation est aberrante et donc qu'il s'agit potentiellement d'une situation d'épidémie. L'objectif de la fonction de poids de sous-pondérer les alertes du passé est donc satisfait.

Enfin, une alerte est levée par le modèle si le comptage y_0 dépasse un seuil qui représente la borne supérieur de l'intervalle de prédiction de y_0 avec un niveau de confiance α. Ce seuil est construit ainsi :

$$U = \hat{\mu}_0 \left\{ 1 + \frac{2}{3} z_\alpha \hat{\mu}_0^{-1} (\hat{\phi}\hat{\mu}_0 + var(\hat{\mu}_0))^{1/2} \right\}$$

Nous ne détaillerons pas le calcul de ce seuil mais l'idée générale est que l'on suppose la normalité de $\hat{\mu}_0$ et, afin d'assurer cette hypothèse, une transformation à la puissance 2/3 est effectuée afin corriger l'asymétrie.

Le modèle de Farrington est ainsi mieux adapté aux données en choisissant une régression Quasi-Poisson et en intégrant les alertes passés dans la construction du seuil d'alerte. Cependant, des limites demeurent et des adaptations du modèle ont déjà été développées. Notamment, dans un article de 2008, Noufaily *et al.* [75] propose deux développements.

Le premier développement concerne l'introduction de la saisonnalité : le modèle de Farrington est très restrictif dans le sens où il ne prend en compte que les semaines similaires des années précédentes. Cela implique que le nombre d'années utilisées pour entraîner le modèle doit être grand (au minimum 5 ans pour Farrington) mais aussi que l'on n'utilise pas l'information disponible les autres semaines. Il a ainsi été proposé d'intégrer toutes les semaines à l'estimation du modèle en introduisant explicitement la saisonnalité dans la régression afin de donner plus d'importances aux semaines d'intérêt. La saisonnalité est introduite par une variable catégorielle à 10 niveau avec une période de référence de sept semaines (correspondant à la semaine de référence $t_0 \pm 3$ semaines) et neuf autres niveaux de cinq semaines chaque année. L'équation de régression devient donc :

$$log(\mu_t) = \beta_0 + \eta t + \delta_{j(t)},$$

avec $\delta_{j(t)}$ le facteur correspondant à la semaine t.

Le second développement concerne la fonction de poids : Noufaily *et al.* (2012) [75] constate en effet que la fonction de poids est trop restrictive en imposant de sous-pondérer toutes les observations pour lesquelles $s_i > 1$. L'article propose un autre seuillage pour cette fonction de poids et propose, suite à de multiples simulation, de sous-pondérer toutes les observations pour lesquelles $s_i > 2.58$. Cet article propose en effet une évaluation du modèle de Farrington par de nombreuses simulations. Ce nouveau modèle est souvent nommé *Farrington flexible* [69].

Le modèle de Farrington présente des caractéristiques assez adaptées au cadre des arrêts maladie. Le cadre de la régression est premièrement un cadre avantageux : la mise en place de régression est assez simple et l'introduction de covariables (qui est un de nos prérequis pour l'analyse des arrêt de travail) est chose aisée. Manitz & Höhle (2013) [76] ont notamment introduit des covariables dans une adaptation bayésienne du modèle de Farrington afin de prendre en compte l'influence des phénomènes météorologiques sur les maladies infectieuses. Enfin, le modèle a été longuement évalué et présente une robustesse intéressante. Un manque de ce modèle pour l'analyse des arrêts maladie est que ce modèle s'entraîne sur une seule série de donnée alors qu'une surveillance adaptée des arrêts de travail devrait prendre en compte plusieurs sites simultanément.

— *Modèle linéaire mixte généralisé*

Dans le cadre des modèles de régression appliqués à la surveillance, quelques références utilisent des modèles linéaires mixtes généralisés afin d'introduire des effets aléatoires pour prendre en compte les différences locales. Notamment, Kleinman, Lazarus & Platt (2004) [77] ont introduit des effets aléatoires pour la détection de clusters de bioterrorisme dans des zones géographiques très restreintes. Plus récemment et plus proches de nos sujets d'intérêt, Morbey *et al.* (2015) [78] ont introduit des modèles mixtes généralisés dans le système de surveillance syndromique anglais. Ce système utilise l'algorithme de Farrington pour l'analyse des données nationales et des régressions avec effets aléatoires pour les données locales. Le modèle local n'introduit cependant pas les spécificités de Farrington comme le système de sous-pondération des alertes passées.

Le développement d'un modèle de surveillance à effet mixte pour s'ajuster aux particularités locales suscite donc un intérêt et a déjà mené à quelques développements méthodologiques. Nos travaux s'inscrivent dans cette lignée et notre objectif est d'associer ces modèles mixtes aux intuitions du modèle de Farrington.

Méthodes semi et non-paramétriques Les modèles de régression précédents ont été développés dans un cadre paramétrique. Des régressions non-paramétriques ont aussi été utilisées dans le cadre de la surveillance.

Un premier exemple est le modèle de Stern et Lightfoot (1999) [79] qui utilise un lissage de la médiane sur 5 ans. L'utilisation de la médiane plutôt que de la moyenne permet de donner moins d'importance aux alertes passées. Cependant, ce modèle est peu adapté à nos usages puisque l'intro-

duction de covariable ne semble pas aisée.

Un autre exemple est le modèle de Wieland *et al.* (2007) [80] qui modélise la moyenne μ_t grâce à des modèles additifs généralisés en utilisant des noyaux gaussiens. Le modèle présente des avantages similaires aux modèles de régression paramétrique mais ne prend pas vraiment en compte les alertes du passé et pourrait donc être peu robuste.

Un dernier exemple, proposé par Zhang *et al.* (2003) [81], utilise des transformés en ondelette afin de se débarrasser de la tendance et des variations saisonnières mais aussi afin d'être plus robustes aux alertes passés et à certains artefacts liés à la structure des données. Notamment, l'article souligne les variations liées aux vacances scolaires qui sont très visibles sur les données d'arrêt de travail. Pour cette raison, les transformés en ondelette aurait pu être une approche efficace pour les traitement des arrêts maladie. Cependant, cette méthode implique des complexités en termes d'interprétation qui sont difficiles à concilier avec notre objectif de surveillance.

3.3.2.2 Méthodes de série temporelle

La littérature de la surveillance épidémiologique s'est aussi inspirée des méthodes de série temporelle. Ces méthodes peuvent en effet être utiles pour un tel objectif puisqu'elles permettent d'explorer assez finement la structure temporelle des données et notamment l'autocorrélation et la saisonnalité.

Des algorithmes de surveillance ont été développés à partir de modèles *ARIMA* (Auto-Regressive Integrated Moving Average) [82] qui s'adaptent très bien aux données avec un fréquence élevée (ce qui est le cas des données d'arrêts de travail). Ces modèles réclament cependant une stationnarité de la série temporelle qui est complexe dans notre cas puisque la série d'arrêts maladie est impactée par des covariables qui ne peuvent être intégrées. Le modèle est aussi complexe à automatiser puisqu'il y a de nombreux paramètres à estimer qui sont en général validés grâce à des critères graphiques.

Les séries temporelles ont parfois été associées à des modèles de Markov cachés [83]. L'idée derrière ces modèles est que la série temporelle est dirigée par deux distributions : une distribution pour l'état *normal* et une distribution pour l'état *épidémique*. On détecte ainsi une alerte lorsque le modèle passe en phase *épidémique*. Ces modèles ont prouvé leur efficacité mais l'introduction de covariables peut être complexe et peut surtout être complexe à interpréter pour un non-initié (contrairement à la simple présentation d'une valeur de seuil d'alerte).

Les méthodes de séries temporelles, à cause de la difficulté à introduire des covariables (mais aussi un effet propre à chaque entreprise), semble peu adaptées au contexte des données d'arrêt maladie maladie.

3.3.2.3 Méthodes inspirées des processus de contrôle statistique

La deuxième grande famille des méthodes de surveillance statistique provient d'outils inspirés des processus de contrôle.

Une des méthodes les plus populaires en surveillance statistique est la méthode *CUSUM* (*CUmulative SUM*) [84], elle est notamment utilisée dans le système de surveillance américain du *Centers for Disease Control and Prevention* [85]. La méthode CUSUM originale consiste à suivre une variable dépendante de la somme cumulée de l'incidence du phénomène d'intérêt et de déclarer, lorsque cette valeur devient anormalement haute, une alerte. La variable suivie est définie comme suit à l'instant t :

$$C_t = max\left(0, C_{t-1} + \frac{y_t - \mu_t}{\sigma_t} - k\right)$$

, avec k une constante qui dépend en la taille de l'épidémie que l'on souhaite détecter, C_0 et μ_t un volume de bas que l'on peut mesurer comme on le souhaite. L'alerte est levée lorsque $C_t > h$ avec h une constante que l'on décide en fonction de la taille des alertes attendues k et du temps moyen entre deux alertes (qui est un pendant du taux de faux positif).

La méthode CUSUM est très efficace et peut s'adapter très simplement. μ_t peut en effet être estimé de diverses manières et notamment par des modèles de régression avec des effets aléatoires. Une difficulté serait ici le choix du seuil d'alerte h qui dépend d'un critère (le temps moyen entre deux alertes) qui est difficilement quantifiable. Un autre argument pour préféré les modèles de régression plutôt que le modèle CUSUM est que, quitte à utiliser une régression complexe comme une régression avec effet mixte, autant construire le seuil à partir des informations de cette régression plutôt qu'à partir d'une autre variable comme C_t. Le modèle est aussi adapté pour des données d'incidence et la construction de la variable C_t semble moins adapté à des données de prévalence comme les arrêts maladie.

Une autre méthode similaire nommée *EWMA* (Exponentiated Weighted Moving Average) [86] propose un système similaire mais plus adaptée aux événements rares en donnant moins d'importance aux données historiques. Ce modèle présente les mêmes problématiques que le modèle CUSUM mais

pourrait convenir potentiellement à des problématiques qui sont hors de l'objectif de notre modèle. Par exemple, les arrêts maladie dans des entreprises avec peu de salarié sont des événements rares. De même, les accidents de travail peuvent aussi être définis comme des événements rares [87].

3.3.2.4 Autres méthodes de surveillance statistique

Nous n'avons présenté que deux grandes familles de méthodes de surveillance statistique : ces deux familles sont parmi les méthodes les plus utilisées en pratique et semblent, au premier regard, les plus adaptées à notre problématique. De nombreuses autres méthodes ont été développés, notamment basés sur les statistiques de scan [88] ou la théorie des valeurs extrêmes [89]. Nous allons rapidement décrire deux groupes de méthodes pouvant peut-être répondre aux problématiques de la surveillance des arrêts de travail.

Méthodes introduisant une dimension spatiale La littérature scientifique s'est beaucoup penchée sur l'élaboration de modèles pour la détection de clusters épidémiques et la détection locale d'épidémie. Ces méthodes s'inspirent souvent des méthodes précédemment présentées en y introduisant une composante géographique.

Il existe par exemple des méthodes CUSUM spatiales [90] calculant des C_t à l'échelle locale en introduisant dans le calcul les C_t des zones géographiques proches. Il existe aussi des méthodes de régression spatiotemporelles : Lawson *et al.* [91] introduit par exemple à chaque endroit k et instant t un paramètre de surrisque géographique lié au niveau local mais aussi au niveau des régions géographiques proches.

La dimension spatiale pour la surveillance des arrêts de travail pourrait être intéressante puisque, si l'intérêt d'une telle surveillance est d'étudier des maladies infectieuses comme la grippe, on pourrait imaginer un phénomène de contagion d'une entreprise proche à une autre. Il ne s'agit cependant pas de notre problématique ici puisque l'on cherche à étudier les phénomènes "intra-entreprise" et que l'on étudie chaque entreprise comme un lieu distinct des autres.

Méthode de détections multivariées Un autre pan de la surveillance épidémiologique concerne les méthodes de détection multivariées. Ces méthodes concernent l'étude de plusieurs séries de données simultanément pour la surveillance d'un même phénomène. Par exemple, la surveillance distincte

de plusieurs sous-populations (distinguées par âge ou sexe notamment) pourraient permettre une surveillance plus sensible de la grippe [92]. Ces méthodes peuvent fonctionner par des méthodes de réduction de dimension [93] ou par l'agrégation simultanée de plusieurs modèles [94].

A nouveau, cette famille de modèle pourrait s'adapter à nos données mais ne répond pas à la problématique soulevée.

3.3.3 Méthode de surveillance adapté aux données d'arrêts maladie

3.3.3.1 Introduction à la publication 3

Parmi l'ensemble de ces méthodes déjà développées, nous avons décidé d'adapter les algorithmes de Farrington [18] et Farrington-flexible [75]. Ces algorithmes sont basés sur des modèles de régression et ce type de modèle peut facilement être adapté à nos objectifs : introduire un effet spécifique à chaque entreprise, ajuster sur des covariables diverses et, surtout, construire un seuil d'alerte dépendant de ces paramètres.

Notre modèle est un modèle de régression similaire, basé sur une régression Binomiale Négative plutôt que Quasi-Poisson, introduisant tendance, temporalité (de façon identique à *Farrington Flexible*), des covariables et un effet aléatoire pour chaque entreprise. En gardant les notations présentées dans la section précédente, l'équation de régression est la suivante. $\forall i, t > 0$,

$$log(\mu_{i,t}) = \beta_0 + \eta t + \delta_{j(t)} + \beta \mathbf{X_{i,t}} + u_i,$$

avec $\mathbf{X_{i,t}}$ une matrice de covariables et β les coefficients associés et $u_i \sim \mathcal{N}(0, \sigma^2)$ l'effet aléatoire associé à l'entreprise i .

Outre l'équation de régression et la distribution sélectionnée, nous proposons aussi une nouvelle manière de calculer la borne supérieure de l'intervalle de prédiction de niveau α de $Y_{i,t}$ qui sert de borne d'alerte à l'algorithme. L'algorithme de *Farrington* propose une borne estimée par une approximation gaussienne et cette borne n'est plus valide dans le cadre d'un modèle à effet aléatoire. Nous proposons donc de calculer la borne en utilisant la fonction quantile de la distribution Binomiale Négative et les paramètres estimés par le modèle (espérance et variance).

Le modèle est présenté en détails dans la publication présentée dans la prochaine section. Nous l'avons évalue par une série de simulations et l'avons appliqué sur des données d'arrêts maladie pendant

la pandémie de la COVID-19.

3.3.3.2 Publication 3 : modèle de surveillance pour données multi-site avec une application aux données d'arrêts maladie

La publication présentée ci-dessous est aujourd'hui soumise à une revue scientifique internationale.

Résumé de l'article en français : L'article suivant propose un algorithme de surveillance statistique pour la détection d'alerte dans un cadre où plusieurs sites doivent être surveillés simultanément.

Le cas d'usage motivant ces travaux est la surveillance du taux d'arrêts maladie dans les entreprises pour détecter des dérives par entreprise. Cet algorithme peut aussi s'appliquer à des données diverses nécessitant l'analyses simultanées de plusieurs lieux et donc l'usage d'effets aléatoires comme cela a parfois été utilisée en surveillance syndromique [78].

L'algorithme est une adaptation du modèle de Farrington dans le cadre de modèle de régression avec effets mixtes. L'usage d'effets mixtes a nécessité des ajustements du modèle, notamment l'introduction d'une nouvelle fonction de poids et la définition d'un nouveau seuil d'alerte.

L'article propose une série de simulations inspirées des simulations de Noufaily *et al.* (2013) [75] afin de valider le modèle. L'algorithme présente des résultats similaires avec un taux de faux positifs et une probabilité de détection des alertes similaires au modèle de Farrington dans son cadre original.

A statistical algorithm for outbreak detection in a multi-site setting: the case of sick leave monitoring

Tom Duchemin, Mounia N. Hocine, Angela Noufaily

August 4, 2020

Abstract

Surveillance for infectious disease outbreak of for other processes should sometimes be implemented simultaneously on multiple sites to detect local events. Sick leave can be monitored accross companies to detect local outbreaks and identify companies-related issues as local spreading of infectious diseases or bad management practice. In this context, we proposed an adaptation of the Farrington algorithm for multi-site surveillance. The proposed algorithm is a Negative-Binomial regression with a new re-weighting procedure to account for past outbreak and increase sensitivity of the model. We perform several simulations to assess the performance of the model in terms of False Positive Rate and Probability of Detection. We propose an application to sick leave rate in the context of Covid-19. The proposed algorithm provides good overall performance and opens up new opportunities for multi-site data surveillance.

1 Introduction

The increasing flow of data and the recent epidemic threats have increased the need for the development of robust epidemiological surveillance. Epidemiological surveillance is not limited to the study of new cases of a disease, but can be used for a wide variety of data as drug consumption, concentration of a virus in waste water or Google queries [1, 2, 3]. The diversity of these data calls for adaptative methods that could fit to different issues.

Many reviews of epidemiological surveillance methods have already been performed [4, 5, 6, 7] and many statistical techniques have been used for this purpose: regression, time series, statistical process control, spatio-temporal methods for instance. Within the framework of regression models, a widely-used algorithm developed by Farrington [8] uses a Quasi-Poisson regression adjusted on trend and seasonality and reweighted to account for past outbreaks. The Farrington algorithm was extensively validated with simulations [9]. Another well-known algorithm is RAMMIE [10] and used mixed-model Poisson regression. In this second algorithm, a mixed effect was included to monitor infectious

diseases at local levels. Both algorithm are used in routine by health authorities [**?**, 10].

Our methodological development is motivated by the analysis of sick leave data. Companies are places where lots of diseases, as infectious disease or stress [11, 12], can spread. Those diseases could lead to sick leave and then to sick leaves outbreaks if actions are not taken in time. The monitoring could help companies to identify ongoing issues and to provide a better environment to workers.

Epidemiological surveillance was already used to monitor occupational health issues as work-related injury to identify ongoing issue [13], to monitor school absenteeism or agregated sick leave data at the regional level to identify influenza outbreaks [14, 15]. To our knowledge, no method was developed on the specific case of the surveillance of sick leave rate in multiple companies.

Monitoring sick leave data raised specific methodological issues: sick leave rate of companies shows a strong seasonal pattern since it is highly correlated to seasonal infectious diseases as influenza [16] and should then be adjusted on trend and seasonality. Moreover, it should also be adjusted on covariates as sick leave rate is associated to exogenous events as school holidays or to the population of workers in each company as age is for instance associated to sick leave rate [17]. To adjust on those covariates, the model should also be fitted on many companies and a mixed effect should be included in the regression. To solve these issues linked to sick leave data, we propose a model inspired by Farrington and RAMMIE: a Negative-Binomial mixed regression algorithm with reweighting procedure to account for past outbreaks. We propose to assess it with extensive simulation and present an application to sick leave data.

In this paper, we adapt the Farrington algorithm to the case of a multi-site surveillance. Section 2 describes the new algorithm. Section 3 sets out the design of the simulation study and Section 4 describes the results in terms of False Positive Rate (FPR) and Probability of Detection (POD). In Section 5, we present an application to sick leave data. Finally, we discuss our findings and their implications in Section 6.

2 A mixed model for outbreak detection

To determine if a count outcome is unusually high and to detect outbreak, we use the same ideas as in the Farrington algorithm [8]. The Farrington algorithm output is an outbreak threshold based on a Quasi-Poisson regression and reweighted to downweight previous outbreaks. Our algorithm uses these ideas and adapts them to a mixed model context.

2.1 The algorithm

To determine if $Y_{i,T}$, the count outcome in a site $i \in \{1, ..., N\}$ with $N > 0$ at week $T > 0$ is an outbreak, we proceed in three steps.

First step First, we fit a Negative-Binomial regression on the past counts $Y_{i,t}$ with $t \in \{0, ..., T\}$ adjusting for trend, seasonality, covariates and a random effect $u_i \sim N(0, \sigma^2)$:

$$Y_{i,t} \sim NB(\mu_{i,t}, \theta) \text{ with}$$
$$\mu_{i,t} = exp(\beta_0 + \beta \mathbf{X_{i,t}} + \gamma \delta_{t,T} + \eta t + u_i).$$

$\mu_{i,t}$ is the mean of the Negative Binomial distribution and θ is the dispersion parameter such that $Var(Y_{i,t}) = \mu_{i,t} + \frac{\mu^2}{\theta}$. $\mathbf{X_{i,t}}$ is a matrix of covariates and β are the associated coefficients.

$\delta_{T,t}$ is a matrix of indicators to adjust for seasonality that is included in the flexible Farrington algorithm [9] to give more weights to the comparable periods in past years. Each column of $\delta_{T,t}$ describes a period: a first reference 7-week period (corresponding to week $T \pm 3$ weeks) and nine 5-week periods in each year. γ are the associated parameters.

To avoid adaptation of the model to emerging outbreaks, we exclude the 26 last weeks from the baseline data and we only fit the regression on the previous weeks.

Second step In a second step, we reweight the outliers of the training dataset to underweight past alerts and to fit a more robust outbreak threshold. We use the following weight function:

$$\forall i, t > 0,\ w_{i,t} = \begin{cases} \gamma \frac{S}{r_{i,t}}, & \text{if } r_{i,t} > s, \\ \gamma, & \text{otherwise.} \end{cases} \text{ with } \gamma \text{ s.t. } \sum_{i,t} w_{i,t} = NT$$

$S > 0$ is a constant controlling for the strictness of the reweighting and r_i are the Pearson residuals of the model and are defined as:

$$\forall i, t > 0,\ r_i = \frac{Y_{i,t} - \hat{\mu}_{i,t}}{\hat{\mu}_{i,t} + \hat{\mu}^2/\hat{\theta}}.$$

Third step The third step is the computation of the outbreak threshold. We fit a new Negative-Binomial regression with the previous reweighting to give less importance to past outbreaks. The outbreak threshold of company $i \in \{1, ...N\}$ is defined as $U_{it} = Q_{\hat{\mu}_{i,T},\hat{\theta}}(\frac{1+\alpha}{2})$ with $\hat{\mu}_{i,T}$ and $\hat{\theta}$ the estimation of $\mu_{i,T}$ and θ retrieved from the second regression, $Q_{\hat{\mu}_{i,T},\hat{\theta}}$, the quantile function of a Negative Binomial distribution with parameters $\hat{\mu}_{i,T}$ and $\hat{\theta}$ and α the chosen confidence level.

As in the Farrington algorithm, all the sites could be hiearchized and specific site can be flagged for further investigation with an exceedance score defined as:

$$U_{it} = \frac{Y_{it} - \hat{\mu}_{it}}{Y_{it} - U_{it}}$$

All the analyses are performed in R and with the help of the package *glmmTMB* [18]. Codes used in this article can be read from an online deposit referenced in **Supplementary Materials**.

2.2 Comparaison with Farrington and Farrington flexible

Our model provides an adaptation of the Farrington algorithm in the context of multi-site data. The main change is the inclusion of a random effect and it resulted in the modification of several formulas in the algorithm.

First, the formula for the threshold in the original Farrington algorithm used a normal approximation thank to Taylor expansion and to power transformation. Our threshold is more straightforward as it only used the quantiles of the estimated distribution.

Second, we used a Negative-Binomial distribution instead of a Quasi-Poisson: this was required to use the quantile function.

Finally, we chose a different weight function. The weight function used in the Farrington algorithm actually includes a hat matrix that cannot be computed in a mixed model framework. Our function is however similar since it used standardized Pearson residuals instead of standardized Anscombe residuals.

3 Simulation study

We will investigate the validity of the model with extensive simulation study. We will first describe the simulated datasets and their associated scenarios and we will then propose some exploratory analyses. the procedure is similar to Noufailly *et al.* [9] in order to allow for a comparison.

3.1 Simulated datasets

Baseline data We generate data using a negative binomial model of mean μ and variance $\mu + \frac{\mu^2}{\theta}$ with $\eta > 0$ a dispersion parameter. To be consistent with Noufaily *et al.*, we use in our simulatiothn an overdispersion parameter ϕ to have a variance equal to $\phi\mu$. $\mu_{i,t}$ is defined for $i \in \{1, ..., N\}$ and $t \in \{1, ..., T\}$ as:

$$\mu_{i,t} = exp\left(\beta_0 + \beta_X X_{i,t} + \beta_Z Z_{i,t} + \eta_t + \sum_{i=1}^{2} \gamma \left(cos\left(\frac{2\pi st}{52}\right) + sin\left(\frac{2\pi st}{52}\right)\right) + u_i\right)$$

$\mu_{i,t}$ is then defined by an intercept, two covariates $X_{i,t}$ and $Z_{i,t}$ we will define later, a trend, a seasonality we defined with Fourier terms and a random effect $u_i \sim \mathcal{N}(0, \sigma^2)$ with $\sigma > 0$.

In practice, we expect to have continuous and discrete covariates. The covariates should be quite stable in each site but can be very different from one site to another. We then simulate X and Z as follow:

$$X_{i,t} \sim \mathcal{N}(m_i, 1) \text{ with } m_1 \sim \mathcal{U}(30, 50),$$

$$Z_{i,t} \sim Bernoulli(p_i) \text{ with } p_i \sim \mathcal{U}(0, 1).$$

In all of the simulations, we use $N = 50$ sites and $T = 312$ weeks which correspond to 6 years of data. The last 52 weeks are used as the current data

set we use to evaluate de model and the previous 260 weeks are used as the baseline data we will use to fit the model. We will also fix the weight threshold s to $s = 2.5$ and try for $\alpha = 0.95$.

Outbreaks We simulated outbreaks as follows:

1. We randomly selected four weeks for the baseline data and one week for the current data;
2. for each week $t_0 > 0$, we randomly generated the outbreak size with a Poisson random variable of mean equal to $k > 0$ times the standard deviation of the baseline count at t_0 ;
3. we finally randomly distributed these cases in time with to a lognormal distribution of mean 0 and standard deviation 0.5.

In the baseline scenarios, we will use $k = 3$ to simulate medium outbreaks.

Simulated scenarios To evaluate the robustness of the model to a wide range of data sets we can meet in real life, we generate our simulations from 32 parameter combinations described in Table 1. We try different baseline volume (given by β_0), different trends and covariates (given by η, β_X and β_Z), different overdispersion (given by θ) and different standard deviation for random effect (given by σ^2).

For each scenario, we perform 5 replications of $N = 50$ companies for $T = 312$ weeks.

3.2 Evaluation of the model

We evaluate the performance of the model using criteria already calculated in Noufaily *et al.* to ensure a comparison between the two models. The two criteria evaluate the performance of the model when there is an outbreak and where there is no outbreak.

For each replicate of the simulations, we first calculated the False Positive Rate (FPR) as the proportion of observations where the observed value exceeded the threshold in the absence of any current outbreak. The FPR is a rate per week.

The second criterion we calculated is the Probability of Detection (POD): it describes the probability that an outbreak is detected and, in other words, the power of the model. The POD is defined as the proportion of outbreaks detected among the 50 companies in each replicate.

For both of these criteria, we will show an average value among the 5 replicates of the model and also the minimum and maximum values across those replicate, to briefly assess the variability of these criteria between simulations.

Scenario	β_0	η	β_X	β_Z	ϕ	σ^2
1	1	0	0	0	1,5	0,5
2	1	0	0	0	1,5	1,5
3	1	0,0025	0	0	1,5	0,5
4	1	0,0025	0	0	1,5	1,5
5	1	0	-0,5	1	1,5	0,5
6	1	0	0,5	1	1,5	1,5
7	1	0,0075	-0,5	0,5	1,5	0,5
8	1	0,0075	-0,5	0,5	1,5	1,8
9	3	0	0	0	1,5	0,5
10	3	0	0	0	1,5	2
11	3	0,0025	0	0	1,5	0,5
12	3	0,0025	0	0	1,5	2
13	2	0,0025	-1	1	1,5	0,5
14	2	0,0025	-1	1	1,5	1
15	2	0,0075	-0,5	0,5	1,5	0,5
16	2	0,0075	-0,5	0,5	1,5	1,8
17	1,5	0	0	0	3	0,5
18	1,5	0	0	0	3	1,5
19	1,5	0,0025	0	0	3	0,5
20	1,5	0,0025	0	0	3	1,5
21	0,5	0,0025	-1,5	1,5	3	0,5
22	0,5	0,0025	-1,2	1,2	3	1,5
23	0,5	0,0075	-0,5	0,5	3	0,5
24	0,5	0,0075	-0,5	0,5	3	1,5
25	3	0	0	0	3	0,5
26	3	0	0	0	3	1,5
27	3	0,0025	0	0	3	0,5
28	3	0,0025	0	0	3	1,5
29	3	0,0025	-1,2	1,2	3	0,5
30	2	0,0025	-1,2	1,2	3	1,5
31	3	0,0075	-0,5	0,5	3	0,5
32	2	0,0075	-0,5	0,5	3	1,5

Table 1: Parameters used to generate the 32 scenarios.

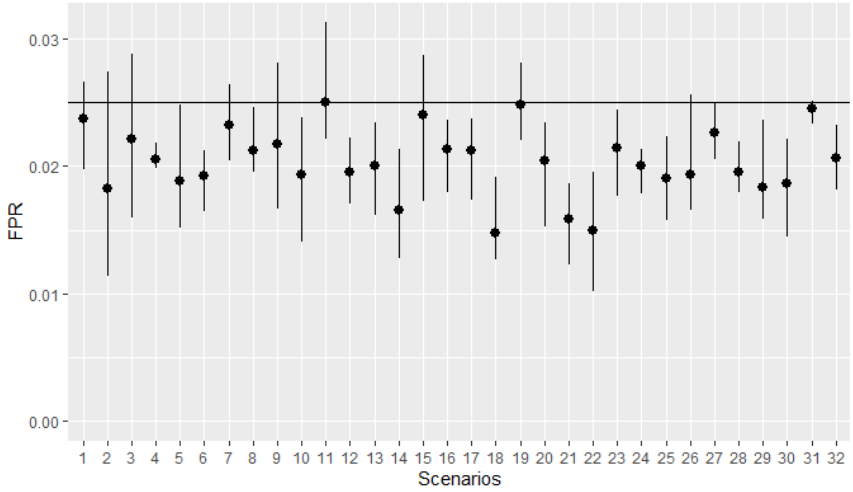

Figure 1: False positive rates obtained $\alpha = 0.95$, $s = 2.5$ and $k = 3$. The horizontal line represents the nominal value 0.025 and numbers refer to scenario.

3.3 Exploratory analyses

The previous simulations fixed some parameters that can have an impact on the results of the model as s or k. To check the impact of those parameters, we performed some quick exploratory analyses.

To evaluate the impact of s and k, we will perform on only one dataset simulated with the parameters of the 7th scenario which includes every parameter with a medium value.

4 Results of the simulation

4.1 False positive rates

Figure 1 shows the FPRs we obtained for $\alpha = 0.95$, $k = 3$ and $s = 2.5$. The point is the median of the five simulations and we also show the range of all those simulations. The nominal FPR is 0.025 and we see that the actual FPRs are a bit lower. Scenarios with no trend and with low random effect presents higher FPR but still lower than 0.025 in median.

4.2 Probability of detection

Figure 2 shows the PODs obtained for $\alpha = 0.95$, $s = 2.5$ and $k = 3$. The point is the median POD across the 5 iterations and the vertical line represent the range of PODs for those 5 iterations. PODs are varying around 0.5 with higher value for scenarios with covariates.

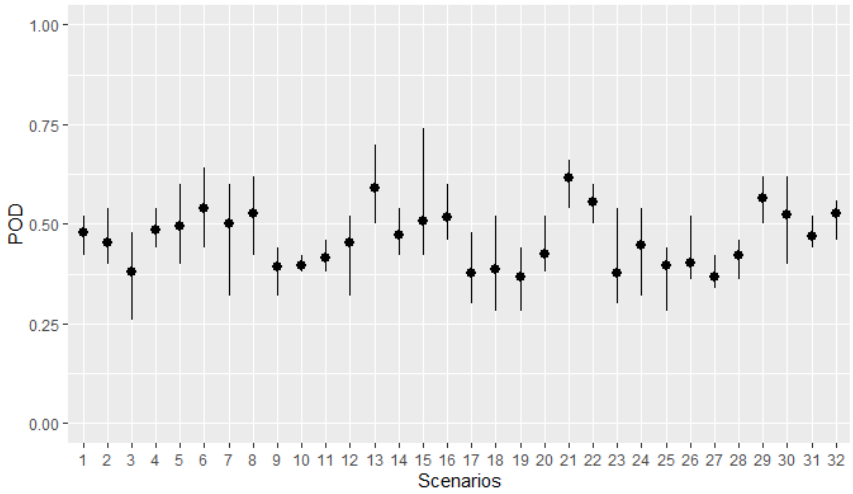

Figure 2: Probability Of Detection obtained $\alpha = 0.95$, $s = 2.5$ and $k = 3$. Numbers refer to scenario.

4.3 Exploratory analyses

Weight threshold Figure 3 shows the FPRs and the PODs obtained for different threshold s for $k = 3$ and $\alpha = 0.95$. For each k, the same dataset is used for the different s to isolate the impact of this weight threshold. We only use the 7th scenario, which includes all parameters of the simulation with medium values and that can be seen as an "average scenario".

As expected, we see that a higher threshold s leads to lower FPR: the underweighting is less strict so the alert threshold is higher. The optimal s, which is the s that gives a FPR around the nominal value, is different according to the size of the outbreaks of our dataset. For large outbreak (k=8), we should choose a low s around 1. For medium outbreak (k=3), we should choose a s aroung 2.5 (which is our baseline value for our previous simulations). For low outbreak (k=1), a higher s should be chosen ($s \pm 3$). Those results are quite consistent with the results from Noufaily *et al.* (2013) [9] that found that the optimal value was $s = 2.58$.

The POD remains almost constant for all values of s: the choice of the threshold s will mostly influence the FPR and is used to monitor the specificity of the model.

Outbreak size Figure 4 shows the FPRs and the PODs obtained for different threshold k for $k = 2.5$ and $\alpha = 0.95$. The same baseline dataset is used for the different k (only the outbreaks are modified) to isolate the impact of the outbreak size. As previously, we only used the 7th scenario.

The FPRs is lower when the outbreak size is higher: it underlines yet again

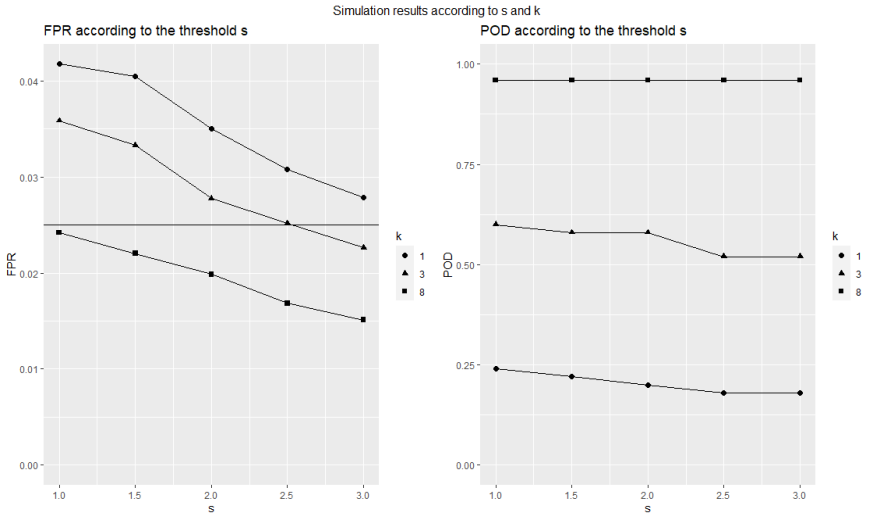

Figure 3: False Positive Rate and Probability Of Detection according to the threshold s

that the s should be adjusted according to the expected size of the outbreaks.

On the other hand, the POD increases greatly with the value of k to approach almost 100% after $k = 8$.

5 Case study: sick leave monitoring and the example of Covid-19

5.1 Data

We ran our algorithm on sick leave data with $s = 2$ and $\alpha = 0.95$. Our dataset describes 1376 French companies of more than 50 employees followed since January 2018.

The outcome of the model is the weekly sick leave rate of each company and is defined as the number of days of sick leave per week on the number of theoretical work days per week. The number of theoretical work days per week is included in the model as an offset.

The dataset also describes some characteristics of the companies we include in the model as covariates: the number of employees per category of age (35 years old and less, 36-45 years old, 46-55 years old, 56 years old and mmore), the number of workers with a temporary contract and the number of workers per occupational categories. We also add an indicator of week with high numbers of vacation days that correspond to low level of sick leaves. A report of the statistics departement of the French Ministry of Labour (DARES) shows that peaks in annual leave occur during the Christmas school holidays (last week of

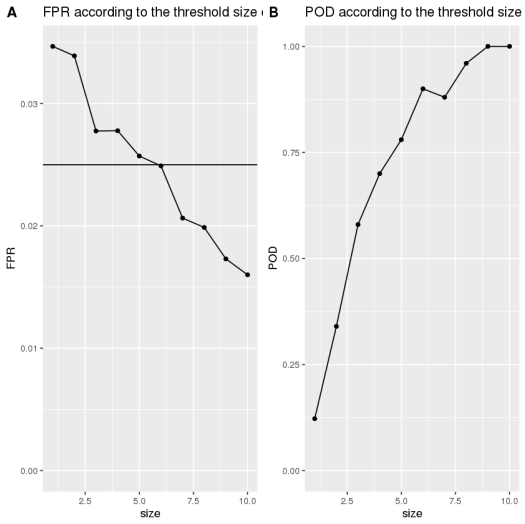

Figure 4: False Positive Rate and Probability Of Detection according to the outbreak size k for $\alpha = 0.95$ and $s = 2.5$

December and first week of January) and during summer (second week of July to third week of August) [19].

The dataset comes from digital files from companies insured by Malakoff Humanis, a French health insurer. The insured companies have to send monthly an update of the socio-demographic characteristics of their employees, their administrative status and their sick leaves. The dataset is named *Déclaration Sociale Nominative* and describes all employees of the companies.

We train the model on 2018 and 2019 and evaluate it on data from January to May 2020. We define an outbreak when we observe during two consecutive weeks a sick leave rate above the alert threshold.

5.2 Results

Figure 5 shows the evolution of the mean sick leave rate from 2018 to May 2020. 2020 is a special year because of the Covid-19 pandemics. Before the third week of March, the sick leave rate follows a distribution similar to the previous years. A peak is observed at the third week of March and corresponds to the first week of lockdown in France. This high sick leave rate should not be interpreted as a high incidence of Covid-19 patients but as an implication of regulatory change: employees who had to stay home for their children were provided sick leaves. We will then run the algorithm on 2020 data to identify companies which were impacted by Covid-19 after the lockdown and companies which had alerts non-related to Covid-19 before lockdown.

Table 2 gives the results of the algorithm run on the dataset in 2020. We

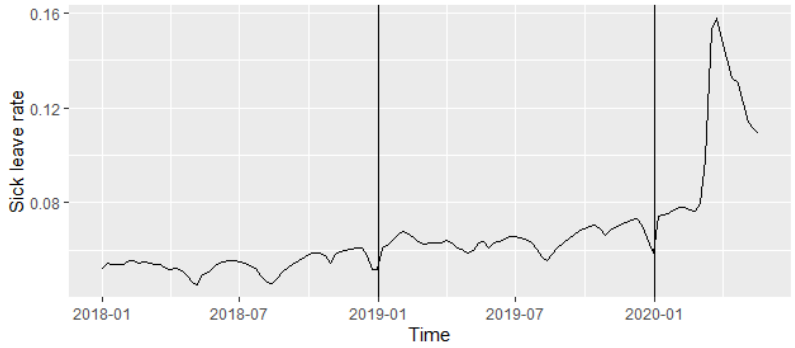

Figure 5: Weekly mean sick leave rate among all companies

	January	February	March	April	May
Number of companies with an outbreak	5.9%	7.9%	56.8%	58.9%	42.1%
Sick leave rate among companies with outbreak	7.5%	7.6%	9.6%	8.0%	6.5%
Sick leave rate among companies with no outbreak	8.3%	8.3%	14.4%	17.7%	17.7%

Table 2: Proportion of companies with a declared outbreak in the five first months of 2020

observe that, before lockdown, 5.9% and 7.9% of the companies have an outbreak in January and February. In March, the number of companies in outbreak rises to 56.8% in March, 58.9% in April and 42.1% in May. More than half of the companies therefore seem to have been affected by Covid-19 in terms of sick-leave. The companies in outbreak have higher sick leave rate after lockdown (17.7% in April and May) than before (8.3% in January and February).

Figure 6 shows four examples of companies and how the results. The first company represents a case where an outbreak occurs just after the lockdown and then the sick leave level goes back to the baseline level. The second company presents no outbreak. The third company has a large outbreak just after the lockdown and the sick leave level stays really high. The fourth company has an outbreak at the beginning of the year. We can observe that the alert threshold is consistent with the baseline level of absence of each company.

6 Discussion

We proposed an adaptation of the Farrinton algorithm for surveillance of multi-site data and we proposed an application to the case of sick leave. The inclusion of a random effect resulted in change in the choice of the weight function and of the alert threshold. Extensive simulation proved that the model provides

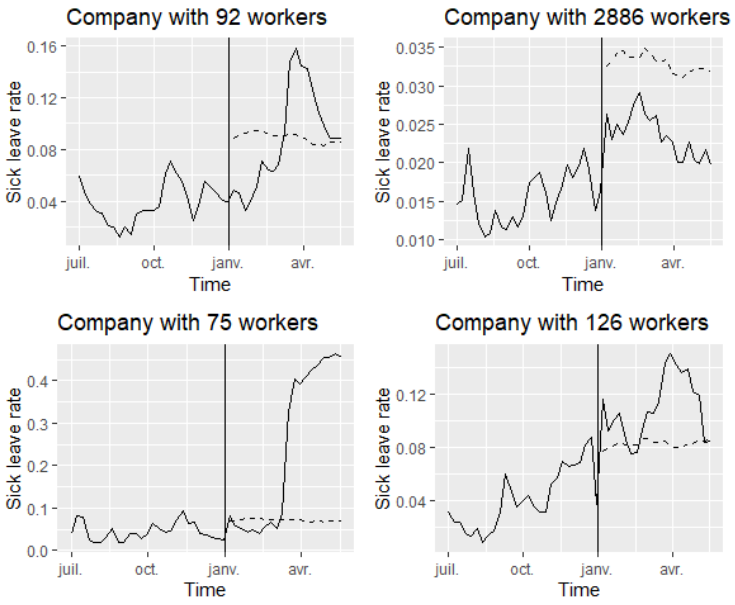

Figure 6: Four examples of companies. Solid line is the. The vertical line represents the first day of 2020, the solid line represents the weekly sick leave rate of the company and the dashed line represent the alert threshold.

correct FPR thank to the reweighting procedure and the results are consistent with the results of the flexible Farrington algorithm.

Computation time could be an obstacle for mixed model as it was sometimes mentionned [10]. It is still an issue with this algorithm as approximately an hour and a half was needed to run an iteration of the model for 1376 companies on our computer. To improve this computational weight, the model could have been stratified by group of companies (by company size for instance) and the new models could have been run in parallel. These difficulties lead us to think carefully about the need to use a mixed model: this model is not necessarily the most appropriate one for a each situation.

The application to sick leave provides interesting results in the case of Covid-19 and helped to indentify companies that was impacted by the pandemics. In more usual situations, this surveillance system could help to identify and alert companies that has unusual level of sick leave in a timely manner to monitor potential issues as bad management practices. Mixed model surveillance is already used in practice to monitor some syndromic data [10] and this study provides a validated algorithm including reweighting procedure.

Supplementary Material

R codes used for the analyses and the simulations can be found at: `https://github.com/TomDuchemin/mixed_surveillance`.

References

[1] Bounoure F, Beaudeau P, Mouly D, Skiba M, Lahiani-Skiba M. Syndromic surveillance of acute gastroenteritis based on drug consumption. Epidemiology & Infection. 2011 Sep;139(9):1388–1395. Publisher: Cambridge University Press.

[2] Ahmed W, Angel N, Edson J, Bibby K, Bivins A, O'Brien JW, et al. First confirmed detection of SARS-CoV-2 in untreated wastewater in Australia: A proof of concept for the wastewater surveillance of COVID-19 in the community. The Science of the Total Environment. 2020 Aug;728:138764.

[3] Kang M, Zhong H, He J, Rutherford S, Yang F. Using Google Trends for Influenza Surveillance in South China. PloS one. 2013 Jan;8:e55205.

[4] Unkel S, Farrington CP, Garthwaite PH, Robertson C, Andrews N. Statistical methods for the prospective detection of infectious disease outbreaks: a review. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2012;175(1):49–82.

[5] Salmon M, Schumacher D, Höhle M. Monitoring Count Time Series in R: Aberration Detection in Public Health Surveillance. Journal of Statistical Software. 2016 May;70(1):1–35.

[6] Buckeridge DL, Burkom H, Campbell M, Hogan WR, Moore AW. Algorithms for rapid outbreak detection: a research synthesis. Journal of Biomedical Informatics. 2005 Apr;38(2):99–113.

[7] Noufaily A, Morbey RA, Colón-González FJ, Elliot AJ, Smith GE, Lake IR, et al. Comparison of statistical algorithms for daily syndromic surveillance aberration detection. Bioinformatics (Oxford, England). 2019;35(17):3110–3118.

[8] Farrington CP, Andrews NJ, Beale AD, Catchpole MA. A Statistical Algorithm for the Early Detection of Outbreaks of Infectious Disease. Journal of the Royal Statistical Society Series A (Statistics in Society). 1996;159(3):547–563.

[9] Noufaily A, Enki DG, Farrington P, Garthwaite P, Andrews N, Charlett A. An improved algorithm for outbreak detection in multiple surveillance systems. Statistics in Medicine. 2013 Mar;32(7):1206–1222.

[10] Morbey RA, Elliot AJ, Charlett A, Verlander NQ, Andrews N, Smith GE. The application of a novel 'rising activity, multi-level mixed effects, indicator emphasis' (RAMMIE) method for syndromic surveillance in England. Bioinformatics (Oxford, England). 2015 Nov;31(22):3660–3665.

[11] Duchemin T, Bar-Hen A, Lounissi R, Dab W, Hocine MN. Hierarchizing Determinants of Sick Leave: Insights From a Survey on Health and Well-being at the Workplace. Journal of Occupational and Environmental Medicine. 2019 Jun;61:1.

[12] Labriola M, Lund T, Burr H. Prospective study of physical and psychosocial risk factors for sickness absence. Occupational Medicine. 2006 Oct;56(7):469–474.

[13] Schuh A, Camelio JA, Woodall WH. Control charts for accident frequency: a motivation for real-time occupational safety monitoring. International Journal of Injury Control and Safety Promotion. 2014 Apr;21(2):154–162. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/17457300.2013.792285.

[14] Cheng CKY, Cowling BJ, Lau EHY, Ho LM, Leung GM, Ip DKM. Electronic School Absenteeism Monitoring and Influenza Surveillance, Hong Kong. Emerging Infectious Diseases. 2012 May;18(5):885–887.

[15] Duchemin T, Bastard J, Ante-Testard PA, Assab R, Daouda OS, Duval A, et al. Monitoring sick leave data for early detection of influenza outbreaks. medRxiv. 2020 May:2020.05.28.20115782. Publisher: Cold Spring Harbor Laboratory Press.

[16] O'Reilly FW, Stevens AB. Sickness absence due to influenza. Occupational Medicine. 2002 Aug;52(5):265–269.

[17] Airaksinen J, Jokela M, Virtanen M, Oksanen T, Koskenvuo M, Pentti J, et al. Prediction of long-term absence due to sickness in employees: development and validation of a multifactorial risk score in two cohort studies. Scandinavian Journal of Work, Environment & Health. 2018;44(3):274–282.

[18] Brooks ME, Kristensen K, Benthem KJv, Magnusson A, Berg CW, Nielsen A, et al. glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling. The R Journal. 2017;9(2):378–400.

[19] DARES. Les congés payés et jours de RTT : quel lien avec l'organisation du travail ? DARES Analyse. 2017 Sep;55.

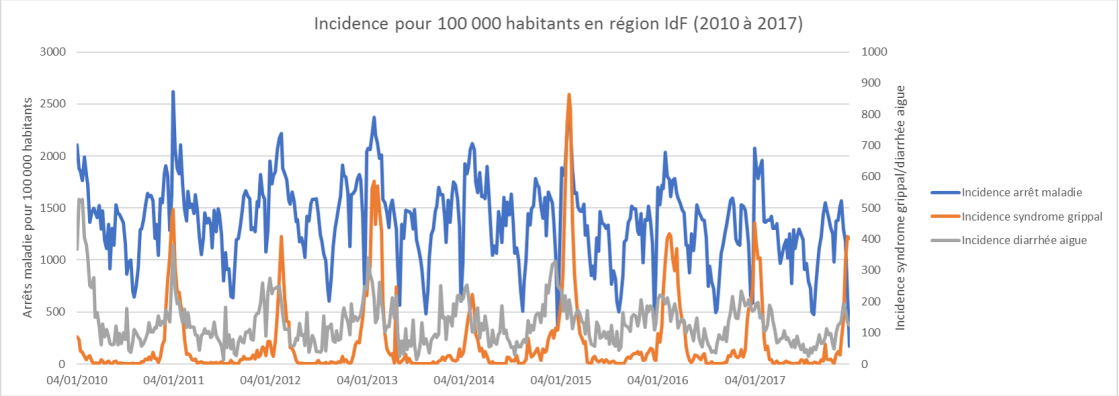

FIGURE 3.2 – Taux d'incidence des arrêts de travail, de syndromes grippaux et de diarrhée aiguë en région Île de France

3.4 Surveiller les arrêts de maladie pour identifier des causes exogènes à l'entreprise

3.4.1 Arrêts maladie et pathologies saisonnières

L'objectif du système de surveillance des arrêts maladie est de pouvoir identifier les entreprises dont le taux d'absence semble hors de contrôle pour des raisons internes à l'entreprise. Les arrêts maladie peuvent aussi être surveillés pour des raisons différentes.

Les travaux de cette thèse se sont pour l'instant consacrés aux facteurs individuels et aux facteurs d'entreprise déterminants les arrêts maladie. Ces arrêts sont aussi déterminés par des facteurs exogènes comme rappelé dans le graphique initiale en Figure 1.1.

La Figure 3.2 représente le taux d'incidence des arrêts maladie, de syndromes grippaux et de diarrhée aiguë en région Île de France de janvier 2010 à décembre 2017. Les données d'incidence des arrêts de travail proviennent des données historiques d'arrêt de travail des entreprises affiliés à Malakoff Humanis, les DADS (Déclaration Administratives de Données Sociales) qui présentent les mêmes informations que les DSN mais dont l'historique s'arrête en 2017. Les données de syndromes grippaux et de diarrhée aiguë proviennent du réseau Sentinelles [95].

Cette figure montre une très grande corrélation entre les données d'arrêt maladie et ces deux pathologies saisonnières. Les données d'absence pourraient ainsi être un bon moyen de surveillance de la grippe voire de la gastro-entérite. Des travaux ont déjà été menés pour intégrer des données d'absence dans des systèmes de surveillance mais principalement des données d'absence à l'école [60, 96] ou bien

des données d'absence au travail mais uniquement comme indicateur complémentaire à un système de surveillance [97].

3.4.2 Publication 4 : Surveiller les données d'arrêt de travail pour la détection de épidémies de grippe

Des travaux parallèles à ces travaux de thèse ont été menés pour évaluer la pertinence des données d'absence pour la surveillance de la grippe. Ces travaux ont mené à un article présenté en Annexe B qui est publié en ligne sur *MedRxiV* et actuellement en revue dans un journal scientifique.

L'article sus-cité évalue les performances et la faisabilité d'un système de détection des épidémies de grippe basé sur le taux d'incidence hebdomadaire des arrêts de travail. Ces travaux reposent sur les DADS présentés brièvement précédemment. Ces données décrivent environ 209 932 entreprises de diverses tailles et de divers secteurs d'activité chaque année. Les données ont été agrégées à l'échelle de la région et un modèle de Serfling a été ajusté sur les données. Une régression linéaire a été estimée pour évaluer le nombre moyen de nouveaux arrêts de travail de 2010 à 2017 dans les 12 régions françaises (Corse exclue) en ajustant sur la tendance, la saisonnalité et les vacances scolaires. Des alertes ont été détectées utilisant un intervalle de prédiction à 95%. Les résultats du modèle ont été comparés aux résultats du réseau Sentinelles, un *gold-standard* en France basé sur des déclarations de médecins volontaires.

Nous avons détecté 92% des épidémies de grippe déclarés entre 2016 et 2017 avec en moyenne 5,88 semaines d'avance par rapport au pic de l'épidémie. En comparaison avec le modèle Sentinelles, notre modèle à une forte sensibilité (89%) et une forte spécificité (86%) et a détecté les épidémies avec en moyenne 2,5 semaines d'avance.

La surveillance des arrêts de travail pourrait ainsi être un moyen sensible, spécifique et réactif pour détecter les épidémies de grippe. Les données d'arrêt de travail peuvent de plus être disponible en temps quasi-réel car les entreprises doivent alerter les autorités et les organismes assurant la prévoyance de l'occurrence de nouveaux arrêts maladie dans les 5 jours.

3.5 Discussion générale

3.5.1 Un outil de surveillance des arrêts maladie

Le système de surveillance des arrêts maladie est en cours d'implémentation dans l'outil de suivi de l'absentéisme proposé aux entreprises assurés par Malakoff Humanis. Le système de surveillance est apposé sur les données d'arrêts maladie présenté dans l'article, les Déclaration Sociales Nominatives. Seules les entreprises de plus de cinquante salariés sont inclus dans ce modèle : dans les entreprises de taille plus modeste, les arrêts de travail pourraient presque être considérés comme des évènements rares (on estime à 34% le pourcentage de salarié ayant un arrêt de travail par an) et une méthodologie différente pourrait être utilisée.

Les Déclarations Sociales Nominatives sont des données mises à jour mensuellement et permettent ainsi une surveillance réactive des arrêts de travail. L'utilisation en pratique des alertes est au jour de la rédaction de ce mémoire encore en cours de discussion : faut-il présenter l'alerte à l'entreprise dans l'interface de l'outil avec une représentation graphique et une explication brève de la situation ? faut-il présenter l'alerte à l'entreprise par le biais d'un consultant en prévention pour l'aider à interpréter ? Dans tous les cas, l'alerte permettra à l'entreprise d'interpréter plus finement son absentéisme et de pousser l'investigation si besoin : si la cause de cette alerte est connue de l'entreprise, des services de prévention peuvent être proposées ; si la cause de cette alerte est inconnue et que l'entreprise veut en savoir plus, des investigations plus poussées peuvent être menées via des enquêtes auprès des salariés.

3.5.2 Perspectives et développement

3.5.2.1 Outils à l'usage des entreprises

L'ensemble des outils présentés dans cette section pour l'identification des situations d'absence anormales offrent un éventail assez large de moyens d'interprétation de l'absentéisme des entreprises. Il reste à voir, si en pratique, ces outils sont véritablement utilisées. Pour une véritable réussite, l'identification d'alerte doit être proposé en complément des moyens de comprendre ces alertes. Une synthèse des facteurs déterminants des arrêts de travail présentés dans la section précédente peut être affichée à l'entreprise. Si la cause de cette alerte demeure difficile à comprendre, des enquêtes internes peuvent être proposées.

3.5.2.2 Modèle de surveillance appliqué à d'autres problématique

L'algorithme développé ouvre aussi des perspectives dans l'utilisation des modèles de surveillance à effet mixte. Aujourd'hui, le principal système de surveillance utilisant des effets mixtes est RAMMIE [78]. Les modèles utilisés dans ce système sont le modèle de Farrington sans effet mixte à l'échelle nationale et des modèles de régression simple avec effet mixte à l'échelle locale. Notre modèle permet d'introduire dans les modèles à effet mixte la pondération du modèle de Farrington pour améliorer la sensibilité du modèle.

L'article présentant RAMMIE [78] soulève un problème lié aux modèles mixtes et aux régressions binomial-négatives qui n'est pas résolu ici : le temps de calcul de ces modèles peut être rédhibitoire si le système nécessite une mise à jour très récurrent des modèles. Dans notre cas, l'entraînement du modèle pour évaluer le seuil d'alerte de 1376 entreprises pour une semaine durait en moyenne 1h30. Ce temps de calcul est peu problématique dans notre cas puisque la mise à jour se fait mensuellement et uniquement pour une série de données ; dans un contexte de systèmes de surveillance où des centaines de séries de données peuvent être évalués à des rythmes plus rapides, ces temps de calculs sont plus difficiles à tolérer et des modèles moins sensibles mais plus simples à entraîner peuvent être favoriser. Une solution pour résoudre ce problème pourrait être cependant de stratifier les entreprises en plusieurs groupes (par taille d'entreprise, par région ou par convention collective ; la convention collective régissant les règles liées aux arrêts de travail) et d'entraîner ces modèles parallèlement.

Chapitre 4

Discussion et perspectives

Contenu

4.1 Synthèse des résultats

L'objectif de nos travaux est de développer des outils statistiques pour la prévention des arrêts maladie. Nous y avons répondu de la manière suivante.

Identifier les déterminants d'arrêts maladie et les populations les plus à risque La première étape a été d'étudier les arrêts maladie à la source en identifiant leurs déterminants. Les arrêts maladie sont des données complexes mais nous n'arrivons bien sûr pas sur un terrain scientifique vierge : nous avons donc, dans un premier temps, rédigé une revue exploratoire de la littérature (*scoping review*) [98] afin d'identifier les méthodologies adéquates pour analyser ces données. Cette revue a permis de constater quelques problématiques pour l'explication ou la prédiction de ces arrêts. En effet, les arrêts maladie sont un phénomène multidimensionnel qu'on peut difficilement expliquer avec un seul indicateur et leurs déterminants sont très nombreux.

Nous avons ensuite analysé des données d'enquête afin d'identifier et de hiérarchiser nous-même les déterminants des arrêts en fonction de leur durée. Suite à la rédaction de notre *scoping review*, nous avons décidé d'utiliser des forêts aléatoires, un modèle non-paramétrique et intégrant très bien les interactions entre nos nombreuses variables. Nous avons pu identifier des déterminants marquants pour les arrêts de longue durée : outre une santé fragile (comme nous pouvions nous y attendre), les arrêts maladie semblent fortement déterminés par des facteurs liés à l'activité dans l'entreprise comme la pénibilité physique ou l'exposition à des facteurs de risques psychosociaux. Pour les arrêts de plus courte durée, les déterminants sont moins marquants : les arrêts de 3 jours à 1 mois sont des arrêts trop hétérogènes pour être analysés ; les arrêts de moins de 3 jours sont principalement associés à des facteurs sociodémographiques. De plus, nous avons pu constater que la performance explicative du modèle est très faible pour ces arrêts de courte durée, contrairement aux arrêts longs : les variables inclues dans le modèle sont donc sûrement insuffisantes et la maille d'étude, annuelle, est peut-être trop grossière. Ces analyses ont permis d'exhiber des leviers d'action possibles pour prévenir les arrêts de longue durée mais ont aussi permis de développer une méthode de hiérarchisation des facteurs déterminants pour les entreprises qui souhaiteraient enquêter dans leur population de salarié.

Afin de mettre en place plus efficacement ces plans d'actions, nous avons tenté d'identifier les salariés pour lesquelles ces interventions seraient les plus profitables à partir de données facilement

accessibles par les entreprises et les organismes de prévoyance. Nous avons donc chercher à identifier les salariés les plus à risque d'arrêts maladie de longue durée en explorant des données assez pauvres en variables explicatives mais très riches en trajectoires d'absence. Ces analyses nous ont permis d'identifier des trajectoires d'absence qui ont un fort risque d'aboutir à des arrêts de longue durée. Nous avons aussi commencé à développer un modèle multi-états permettant d'évaluer dynamiquement le risque d'arrêt maladie des salariés et d'évaluer, pour une entreprise, le volume attendu d'absence.

Les travaux de cette première étape ont permis de fournir quelques outils pour l'identification des déterminants et des populations les plus à risque d'arrêt maladie pour pouvoir fournir les services les plus adaptées aux salariés les plus adaptés.

Identifier les excès d'absence La seconde étape de nos travaux a été d'analyser l'absence à l'échelle de l'entreprise. L'objectif de la prévention des arrêts maladie n'est pas d'empêcher les salariés de prendre des arrêts maladie, qui sont très souvent inévitables, mais d'identifier quand ces arrêts pourraient, justement, être évités.

Si le niveau d'absentéisme est trop élevé, c'est peut-être parce que les salariés sont trop exposés à certains facteurs de risque et certains arrêts pourraient donc être évités par des plans d'action bien choisis. Nous avons donc proposé, dans un premier temps, un moyen d'évaluer le niveau d'absentéisme des entreprises en proposant une typologie construite à partir des données d'absence de plus d'un millier d'entreprise. Cette typologie prend en compte l'activité et la structure sociodémographique des entreprises et permet donc de se comparer à des entreprises relativement similaires.

Cet outil de comparaison est relativement simple et doit tout de même être utilisé avec précaution : même si la structure sociodémographique des entreprises est similaire, les comportements d'absence peuvent être différents sans que ceci s'explique par une exposition plus forte à des facteurs de risque d'arrêt. Notamment, certaines entreprises ont des politiques d'indemnisation plus généreuses et la culture des entreprises est parfois différente. Pour identifier plus certainement les dérives d'absence, nous avons développé un outil de surveillance multi-site permettant de détecter des excès ponctuels d'absence. Ce modèle est un développement d'algorithmes déjà utilisés en routine pour la surveillance épidémiologiques et permet, dans le cadre des arrêts maladie, de détecter des dérives d'absence à la hausse en ajustant sur les caractéristiques sociodémographiques des entreprises. Nous avons aussi montré que la surveillance des arrêts maladie pouvait être un moyen efficace de détecter les épidémies

de grippe.

4.2 Retour d'expérience : mise en place du système de monitoring

Les travaux menés ont ainsi permis de développer un certains nombres d'outils complémentaires pour prévenir les arrêts maladie des salariés. Ces travaux ont été en partie intégrés à un système de monitoring proposé par Malakoff Humanis.

Cet outil porte le nom de *Diagnostic et Protection du Capital Humain en entreprise* et comporte deux volets : un volet *Absentéisme* et un volet *Capital Humain*.

Nos travaux se portent plus logiquement vers le premier volet de l'outil. L'outil propose à l'entreprise de suivre divers indicateurs d'absentéisme (fréquence, durée, proportion) et de les comparer à des valeurs *benchmark*. La typologie proposée dans cette thèse y est inclus et permet aux entreprises de se situer par rapport aux autres et d'évaluer si son absentéisme est problématique ou non. Un outil non présenté dans cette thèse permet à l'entreprise de synthétiser automatiquement les informations clé de son tableau de bord : l'entreprise va premièrement voir si son entreprise a un niveau d'absentéisme problématique ou non et va ensuite voir si certaines populations de salariés sont plus sinistrées que d'autre. Le modèle de surveillance est aussi en cours d'intégration et permettra de signaler aux entreprises que leur taux d'absentéisme est anormalement haut et pour leur proposer des investigations supplémentaires.

Le premier volet des travaux présentés dans cette thèse se situe à l'intersection des deux volets de cet outil : le *Capital Humain* de l'entreprise est évalué à partir d'un questionnaire similaire à celui utilisé dans nos analyses et permet ainsi d'identifier des déterminants d'arrêts de travail. Ce volet de l'outil permet de communiquer sur des résultats d'étude nationale pour informer les entreprises, notamment sur les causes d'arrêts maladie, mais aussi de proposer aux entreprises de mettre en place leurs propres enquêtes parmi leurs salariés. Nos travaux ont ainsi permis de fournir des éléments à communiquer aux entreprises, de sélectionner des questions essentielles pour l'élaboration des questionnaires à destination des entreprises mais aussi un moyen de hiérarchiser facilement les facteurs de risque propres à chaque entreprise.

La complexité pour l'élaboration d'un tel outil est principalement lié à la complexité des données et des résultats. Nous avons tenté de répondre à ce problème en construisant une typologie afin d'effectuer

des comparaisons efficaces et en développant un modèle de surveillance afin d'identifier clairement les entreprises qui peuvent avoir besoin d'investigations quand à leur niveau d'absentéisme. La méthode d'identification et de hiérarchisation des déterminants des arrêts maladie a permis de produire des résultats synthétiques pour les communiquer aux entreprises mais permet aussi de reproduire ces analyses sur d'autres données pour les entreprises le souhaitant.

Une attente de Malakoff Humanis et des personnes ayant participé à la mise en place de cet outil était de développer un moteur de prédiction des arrêts de travail pouvant permettre aux entreprises d'anticiper les mouvements d'absence pouvant impacter leur organisation. Nous avons mené des travaux utilisant diverses méthodes des séries temporelles ou de *machine learning* afin d'analyser les données agrégées au niveau de l'entreprise et ces travaux ont toujours mené à des résultats décevants. Récemment, des modèles ont aussi émergés dans la littérature scientifique et utilisent des méthodes assez récentes d'apprentissage profond sur des données invididuelles mais surtout plus riches en variables explicatives que les nôtres et ont produit des résultats assez moyens [99]. Les arrêts maladie sont un phénomène très hétérogènes qui semblent mal se prêter à la prédiction. Nous avons tout de même proposé quelques travaux pour estimer le risque d'arrêt maladie à partir de trajectoires d'arrêts et quelques pistes d'exploration basés sur des modèles multi-états pouvant permettre d'évaluer la proportion espérée de salarié absent dans le temps.

4.3 Limites

Après avoir synthétisé nos travaux et avoir décrit leurs applications en pratique, penchons nous plus sérieusement sur les méthodes proposées dans cette thèse pour en identifier les limites.

Hiérarchisation des facteurs déterminants La hiérarchisation des arrêts maladie a permis de synthétiser facilement et avec peu d'hypothèses les déterminants de ces arrêts. La méthode choisie soulève tout de même quelques limites.

Nous avons déjà soulevé la problématique de l'introduction de la santé perçue dans le modèle. Les arrêts maladie étant pris pour raison de santé, il est logique que cette variable soit un déterminant important, voire même le plus important. Dans notre modèle, il pourrait plutôt s'agir d'une variable médiatrice que d'une variable vraiment déterminante des arrêts maladie. Cependant, des salariés gé-

néralement en bonne santé peuvent avoir des arrêts maladie et des salariés en mauvaise santé peuvent ne pas avoir d'arrêts maladie : en contrôlant sur la santé, nous pouvons donc aussi expliquer ce qui pousse ces salariés à prendre ou ne pas prendre des arrêts, indépendamment de leur condition de santé initiale. Même si sa présence dans le modèle peut être discuté, l'importance de cette variable dans la prédiction des arrêts maladie démontre tout de même que l'indicateur de santé perçue est un indicateur important pour comprendre l'évolution des arrêts maladie : il s'agit de plus d'un indicateur pertinent pour évaluer la qualité des vie des personnes [100] et suivre la santé perçue dans le temps présente ainsi de nombreux intérêts. Nous pouvons formuler une dernière remarque vis-à-vis de cet indicateur : la santé perçue, telle que recueillie dans le questionnaire, ne décrit pas la santé perçue juste avant l'occurrence de l'arrêt maladie mais plutôt une santé moyenne l'année où a pu avoir lieu cet arrêt. Le modèle ne montre donc pas juste que les arrêts maladies sont pris lorsque que les gens sont malades mais plutôt qu'une qualité de vie dégradée va causer plus d'arrêts maladie.

Un deuxième problème est méthodologique : les forêts aléatoires sont des méthodes puissantes pour la prédiction mais ne sont pas des outils très pratiques pour l'explication. Les résultats fournis par notre étude permettent de hiérarchiser les déterminants selon leur importance dans la classification absent/non-absent mais ne donnent pas d'information sur le sens de l'impact. Le sens de l'impact est souvent évident (une mauvaise santé détermine la prise d'arrêts maladie) mais ce n'est pas toujours le cas et, parfois, l'impact d'une variable peut interagir avec d'autres variables. Des méthodes différentes d'évaluation de l'importance de variables pourraient être utilisées, comme la méthode SHAP [43, 44] qui permet d'évaluer le sens de l'impact mais aussi les interactions entre les différentes variables. Cette méthode est relativement récente et nous n'en avions pas connaissance lorsque nous avons développés nos travaux sur les forêts aléatoires. Nous avons cependant essayé d'explorer ces méthodes dans le cadre d'un stage qui ont validé notre hiérarchisation des déterminants et qui a confirmé nos intuitions sur le sens des impacts.

Prédiction des arrêts maladie Nous avons brièvement traité du sujet de la prédiction des arrêts maladie dans notre mémoire en tentant deux approches basées sur les trajectoires d'absence :

1. Une première approche a consisté à se concentrer sur les variables explicatives du modèle en tentant de construire de nouvelles variables basées sur la trajectoire des salariés puis en utilisant des méthodes classiques de prédiction en classifiant des salariés selon leur statut d'absence ;

2. Une seconde approche a consisté à se concentrer sur une modélisation multi-états pour pouvoir évaluer les probabilités de transition qui mènent les salariés à l'absence. Ce modèle n'est pas à proprement parler un modèle de prédiction des arrêts maladie d'un salarié mais il peut servir à estimer le volume d'absence attendu pour une entreprise.

Ces deux méthodes ne permettent pas de prédire convenablement, pour un salarié, l'occurrence de nouveaux arrêts maladie dans un futur proche. Cet objectif semble de toute façon illusoire avec des données aussi pauvres : des modèles sur des données plus complètes ont permis d'atteindre des résultats légèrement meilleurs [99] mais toujours pas satisfaisants.

Le modèle multi-état, bien que rudimentaire tel que présenté dans ce mémoire, semble tout de même bien s'ajuster sur les données d'absence et pourrait permettre d'anticiper, pour l'entreprise et pas pour le salarié, des volumes d'absence futur. De nombreuses améliorations devraient cependant être prise en en compte et principalement concernant la modélisation probabilités de transition. Les transitions sont en effet modélisés par des modèles de Cox qui reposent sur une hypothèse de risques proportionnelles très contraignantes. Nos données présentent en effet quelques caractéristiques qui peuvent aller à l'encontre de cette hypothèse. Premièrement, beaucoup de salariés n'auront jamais d'arrêts maladie ce qui pose le problème des survivants de longue durée [56]. De plus, le risque d'arrêt maladie n'est pas constant dans le temps mais est saisonnier, ce qui ne peut pas être capté dans le modèle de Cox. Nous sommes donc bien conscients que le modèle présenté est assez faible mais il permet tout de même de proposer un premier brouillon pour anticiper les volumes d'absence.

Surveillance des arrêts maladie Le mémoire a présenté deux études basées sur la surveillance des arrêts maladie.

La première étude a consisté au développement d'un algorithme de surveillance pour identifier les entreprises en dérive d'arrêts maladie. L'algorithme proposé est une adaptation des modèles de Farrington et Farrington-flexible [18, 75] au cadre de données multi-sites grâce à l'introduction d'effet aléatoire. Le modèle présente des résultats similaires au modèle Farrington-flexible et permet même un calcul plus simple des bornes d'alerte. Ce modèle pose cependant des problèmes de calcul puisque le temps de calcul nécessaire d'une borne d'alerte semble relativement prohibitif lorsque le nombre de sites à évaluer est grand : nous aurions dû, d'ailleurs, évaluer plus précisément le temps de calcul nécessaire en fonction du nombre de sites, de périodes d'observation et de covariables.

Le calcul de la borne d'alerte présente aussi ses limites. Pour rappel, la borne d'alerte calculé repose uniquement sur le quantile d'une Négative Binomiale dont les paramètres sont estimés par le modèle. La borne ne prend donc pas en compte l'erreur d'estimation possible du modèle (et donc la variance des estimateurs). Nous avions proposé une construction alternative de cette borne supérieure par des simulations selon l'algorithme de Metropolis-Hastings : l'idée générale était de simuler l'ensemble des paramètres N fois selon leur distribution estimée par le modèle et de récupérer les quantiles empiriques de ces nouvelles observation simulées. Cette méthode permettait de bien prendre en compte l'incertitude autour de l'estimation mais le temps de calcul était encore plus prohibitif.

La deuxième étude concerne l'évaluation d'un modèle de surveillance des arrêts maladie à l'échelle régionale pour la détection des épidémies de grippe. Les résultats sont satisfaisants et ont montré que le modèle était très sensible et relativement spécifique par rapport aux modèles actuellement utilisés en routine. Cette étude évalue cependant ce modèle dans un cadre idéal où les données d'absence seraient disponibles presque immédiatement. Ces données sont en effet renseignées quotidiennement par les ressources humaines des entreprises et sont ensuite signalées aux organismes (assurance prévoyance et ministère du travail). Les certificats médicaux justifiant les arrêts maladie doivent de plus être envoyés dans les deux jours à l'Assurance Maladie. La construction d'indicateurs d'arrêts maladie d'incidence hebdomadaire semblent donc faisables en traitant les données de gestion des entreprises qui sont contraintes de signaler les arrêts de travail ou en calculant le nombre de certificats médicaux reçus par jour.

4.4 Perspectives

Hiérarchisation des déterminants des arrêts maladie Les travaux autour de la hiérarchisation des déterminants des arrêts maladie a permis d'identifier des facteurs déterminants mais a aussi permis d'évaluer l'intérêt d'une nouvelle méthode pour expliquer des résultats complexes d'enquête. Notre modèle ne permet certes pas d'évaluer quantitativement le sens de l'impact des variables mais permet de prendre en compte des relations plus complexes entre les variables. Récemment, d'autres méthodes d'explications de modèles ont vu le jour, comme *SHAP* qui a déjà été cité et permettent d'expliquer les résultats d'un modèle de prédiction [43, 44] en évaluant le sens et l'intensité de l'impact d'une variable sur une autre. Ces nouvelles méthodes sont parfois nommées méthodes *agnostiques* car elles peuvent s'appliquer à n'importe quel modèle prédictif (que ce soit une régression logique ou un réseau neuronal

complexe). Ces manières d'expliquer des modèles sont prometteuses pour l'analyse de causalité puisque l'on va pouvoir évaluer des interactions plus complexes entre les variables grâce à l'utilisation de modèles qui n'était auparavant utilisé que pour faire des prédictions. Nous avons poursuivi ces travaux en encadrant deux stages dont l'objectif était **(1)** d'évaluer l'intérêt de *SHAP* et la cohérence des nouveaux résultats avec ceux de notre précédent étude et **(2)** de faire une revue de littérature des différentes méthodes d'explication de variables. Nous avons aussi appliqué ces méthodes sur des données biologiques dans des travaux à paraître en collaboration avec Fanie Shedleur-Bourguignon et Philippe Fravalo de l'Université Vétérinaire de Montréal.

Dans le cadre du système de monitoring des arrêts maladie, les analyses de l'article ont permis d'informer les entreprises de quelques déterminants clé (notamment les déterminants liés à la pénibilité physique et aux facteurs de risque psychosociaux) à partir d'une enquête représentative de la population nationale. L'objectif serait maintenant d'appliquer ce modèle aux enquêtes lancées dans chacun des entreprises pour fournir une hiérarchisation des déterminants propres à chacune.

Prédiction des arrêts maladie Ces travaux de thèse ne propose qu'une esquisse autour de la prédiction des arrêts maladie. La littérature, comme nous l'avons montré, commence à s'intéresser à cette question et propose quelques modèles pour prédire l'absence au niveau individuel. Cependant, les données utilisées sont souvent très riches par rapport aux nôtres et la question de prédiction individuel n'est finalement pas ce qui nous intéresse : à quoi bon savoir qu'un salarié va s'arrêter dans les prochains jours ? Avoir la réponse à cette question pourrait même mener à des actions douteuses et notre objectif est plus large : nous souhaiterions anticiper le volume d'absence pour l'ensemble de l'entreprise plutôt que l'absence d'un salarié unique.

L'ensemble des travaux autour de ce sujet reste à faire : les modèles multi-états pourraient permettre, grâce à la représentation graphique de la trace de Markov, d'évaluer dans le temps l'évolution attendu du volume d'arrêt maladie. Il faudrait cependant résoudre les limites que nous avons soulevé ci-dessus et évaluer les résultats du modèle sur plusieurs entreprises et non uniquement sur l'agrégation de l'ensemble d'entre elles. De plus, ce modèle pourrait potentiellement permettre d'évaluer l'efficacité de diverses interventions et notamment le service de monitoring que nous présentons dans ce mémoire. Les modèles multi-états sont d'ailleurs souvent utilisés pour de tels objectifs et notamment dans le cadre des arrêts maladie [101, 102].

Surveillance des arrêts maladie Le modèle développé dans le contexte de cette thèse pourrait être appliqué dans d'autres contextes. En surveillance syndromique, on peut notamment utiliser des modèles mixtes pour identifier les épidémies à des échelles locales. L'exemple le plus concret est celui de l'algorithme *RAMMIE* [78] qui est un algorithme utilisé en routine à Public Health England. L'algorithme est lancé quotidiennement pour évoluer 12 000 signaux (à l'échelle nationale, à l'échelle régionale qui représente 15 sites et à l'échelle locale qui réprésente 152 sites). A l'échelle nationale est utilisé l'algorithme Farrington-flexible mais à l'échelle locale, des régressions de Poisson à effet mixtes sont utilisées. Notre algorithme permet d'introduire les avantages de l'algorithme de Farrington (sous-pondération des alertes passées et introduction de la saisonnalité) et offre donc une meilleure sensibilité. Notre modèle est de plus validée par simulations.

Pour pouvoir être utilisé quotidiennement, des travaux pour améliorer les temps de calcul du modèle devront sûrement être menés puisque les temps de calcul peuvent être assez prohibitifs. Ces travaux pourraient consister en une parallélisation des calculs (en utilisant le logiciel déjà développé), en une stratification des analyses (les calculs sont assez longs dans notre cas puisque l'ensemble des 1400 entreprises sont inclues dans le même modèle) ou en l'utilisation d'autres méthodes. Récemment, des développements ont été menés pour évaluer l'intérêt des réseaux neuronaux pour estimer les modèles linéaires mixtes généralisés : ces résultats montrent une meilleure performance pour l'évaluation des paramètres et, surtout, un véritable gain en temps de calcul [103].

Le modèle n'ayant été validé que par simulation, il serait intéressant d'évaluer en pratique si les pics identifiés sont véritablement des pics et si nous arrivons à les expliquer. Des travaux pourraient aussi être menés pour évaluer si une stratification pourrait être avantageuse pour les résultats de notre modèle : au lieu de faire un modèle pour l'ensemble de nos 1600 entreprises, serait-il plus profitable d'avoir un modèle par secteur d'activité ? Cela pourrait permettre d'ajuster plus précisément sur les caractéristiques socio-démographiques des entreprises.

Enfin, notre modèle a été développé pour pouvoir analyser des entreprises de plus de 50 salariés. Pour des entreprises de plus petite taille, il faudrait réfléchir à des méthodes différentes puisque l'arrêt maladie est, dans ce cas, un événement relativement rare. Les méthodes inspirées des processus de contrôle statistique pourraient être plus appropriées pour ces analyses.

Système de monitoring des arrêts maladie La plupart des outils présentés dans ce mémoire sont implémentés ou en cours d'implémentation dans le système de monitoring des arrêts maladie proposé par Malakoff Humanis.

A court terme, le modèle de surveillance doit être implémenté et il reste à déterminer comment seront vraiment utilisés les alertes du modèle : est-ce que les alertes seront communiquées à un consultant en prévention qui devra contacter les entreprises et leur présenter des plans d'action ? est-ce que les entreprises seront simplement contactés par un mail qui les invitera à creuser plus en profondeur leur situation d'absentéisme ? Ces questions restent ouvertes et seront répondues dans les prochains mois.

A moyen terme, des enquêtes seront lancées dans les entreprises pour pouvoir déterminer les déterminants d'absence spécifiques à chacun. Il serait intéressant d'analyser ces données afin d'évaluer l'efficacité et la stabilité en pratique de notre méthode basée sur les forêts aléatoires. Une application souhaitée de ces méthodes serait de pouvoir créer un modèle qui proposerait aux entreprises les services les plus adaptées à leurs problématiques. Aujourd'hui, des travaux sont effectués pour pouvoir associer chacun des items du questionnaire à des services. Les services mis en avant pourront être ceux qui préviennent des problématiques génératrices d'arrêts maladie (mais pas uniquement).

Enfin, nous l'avons déjà souligné, il serait très intéressant d'évaluer l'impact de ces services. Nous espérons que la mise à disposition d'un tableau de bord des arrêts maladie aux entreprises peut permettre aux entreprises d'identifier des problèmes et d'agir en circonstances, ce qui pourrait réduire leur absentéisme. De même, nous espérons que les services qui auront été identifiés à partir de notre hiérarchie des facteurs d'arrêt maladie auront un impact positif.

Conclusion

La numérisation et le partage systématique des données d'arrêt maladie offrent de belles opportunités pour la prévention en santé au travail et nous avons tenté, dans ce mémoire, de profiter de ce qui était immédiatement à notre disposition pour développer un éventail d'outils à l'usage des entreprises.

Nous avons tout d'abord réfléchi aux mécanismes qui pouvaient expliquer les arrêts maladie : quels sont leurs déterminants et peut-on détecter des signaux précurseurs d'arrêt chez le salarié ? Les premiers travaux ont permis d'identifier des facteurs clé pour la mise en place de plans d'action efficaces. Les méthodes utilisées permettent d'esquiver certains problèmes inhérents aux données d'arrêts maladie comme leur grand nombre de prédicteurs et pourraient aussi être utiles pour identifier des déterminants dans des contextes différents. Pour préciser encore plus ces plans d'action, nous nous sommes ensuite penchés sur les données administratives d'arrêts maladie afin d'évaluer si des signaux précurseurs d'arrêts maladie pouvaient être identifiés dans des données plus pauvres en variables. Nous avons tout d'abord décrit les trajectoires d'absence des salariés avec des absences de longue durée et identifier certains types de trajectoires qui peuvent mener avec un risque accru vers des arrêts graves. Nous avons finalement affiné cette analyse par une modélisation multi-état qui nous a permis d'évaluer quantitativement ces risques accrus d'absence lorsque ces événements se répètent. Cette dernière analyse pourrait aussi nourrir des outils de projection de l'absence des entreprises et leur permettre ainsi d'anticiper plus efficacement l'absence de leurs salariés.

Nous avons ensuite réfléchi à l'échelle de l'entreprise : chaque salarié risque un arrêt maladie, mais ce risque peut être parfois accru dans certaines entreprises où les salariés sont surexposés à des facteurs de risque. Pour répondre à cette problématique, nous avons procédé en deux étapes. Premièrement, les entreprises ont pour habitude d'évaluer leur absentéisme en se comparant à des valeurs repère. Ces repères sont souvent imprécis et peuvent mener à des erreurs d'interprétations. Nous avons donc proposé une typologie d'entreprise afin de prendre en compte les spécificités d'absence

liées à leur structure sociodémographique et à leur activité. Deuxièmement, pour pouvoir identifier automatiquement les entreprises potentiellement en excès d'arrêt maladie et leur proposer des services, nous avons développé un modèle de surveillance pour identifier des pics d'absence inexpliqués. Ce modèle est adapté de modèles de surveillance épidémiologique principalement utilisés pour surveiller des données syndromiques en y intégrant des spécificitées liées à nos données. Ces travaux nous ont aussi mené à une réflexion plus globale sur la surveillance des arrêts maladie puisque nous avons montré que ces données pouvaient être très efficaces pour détecter les épidémies de grippe saisonnières.

Les méthodologies développées dans ce mémoire pourraient être utilisées dans d'autres domaines. Notamment, le modèle de surveillance des arrêts maladie semble être très adapté au contexte de surveillance syndromique. Enfin, nous espérons que ces travaux ont pu fournir des outils utiles et pratiques pour la prévention des arrêts maladie et pourront servir aux entreprises visées.

Bibliographie

[1] E. Vingård, K. Alexanderson et A. Norlund, "Chapter 9. Consequences of being on sick leave :," *Scandinavian Journal of Public Health*, nov. 2016, publisher : SAGE PublicationsSage UK : London, England. [En ligne]. Disponible : https://journals.sagepub.com/doi/10.1080/14034950410021899

[2] M. Ockander et T. Timpka, "A female lay perspective on the establishment of long-term sickness absence," *International Journal of Social Welfare*, vol. 10, n°. 1, p. 74–79, 2001, _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1111/1468-2397.00154. [En ligne]. Disponible : https://onlinelibrary.wiley.com/doi/abs/10.1111/1468-2397.00154

[3] P. Qin, E. Agerbo, N. Westergård-Nielsen, T. Eriksson et P. B. Mortensen, "Gender differences in risk factors for suicide in Denmark," *The British Journal of Psychiatry : The Journal of Mental Science*, vol. 177, p. 546–550, déc. 2000.

[4] A. Maladie, "Améliorer la qualité du système de santé et maîtriser les dépenses. Propositions de l'Assurance Maladie pour 2020." juill. 2019. [En ligne]. Disponible : https://assurance-maladie.ameli.fr/sites/default/files/2019-07_rapport-propositions-pour-2020_assurance-maladie.pdf

[5] A. C. Group, "7ème baromètre de l'Absentéisme," 2015. [En ligne]. Disponible : http://presse.ayming.com/mobilerelease.aspx?ID=34695

[6] E. Demou, S. Smith, A. Bhaskar, D. F. Mackay, J. Brown, K. Hunt, S. Vargas-Prada et E. B. Macdonald, "Evaluating sickness absence duration by musculoskeletal and mental health issues : a retrospective cohort study of Scottish healthcare workers," *BMJ open*, vol. 8, n°. 1, p. e018085, 2018.

[7] M. Virtanen, J. Ervasti, J. Head, T. Oksanen, P. Salo, J. Pentti, A. Kouvonen, A. Väänänen, S. Suominen, M. Koskenvuo, J. Vahtera, M. Elovainio, M. Zins, M. Goldberg et M. Kivimäki,

"Lifestyle factors and risk of sickness absence from work : a multicohort study," *The Lancet. Public Health*, vol. 3, n°. 11, p. e545–e554, 2018.

[8] W. Beemsterboer, R. Stewart, J. Groothoff et F. Nijhuis, "A literature review on sick leave determinants (1984-2004)," *International Journal of Occupational Medicine and Environmental Health*, vol. 22, n°. 2, p. 169–179, 2009.

[9] H. de Vries, A. Fishta, B. Weikert, A. Rodriguez Sanchez et U. Wegewitz, "Determinants of Sickness Absence and Return to Work Among Employees with Common Mental Disorders : A Scoping Review," *Journal of Occupational Rehabilitation*, vol. 28, n°. 3, p. 393–417, 2018.

[10] J. Petersen, L. Kirkeskov, B. B. Hansen, L. M. Begtrup, E. M. Flachs, M. Boesen, P. Hansen, H. Bliddal et A. I. Kryger, "Physical demand at work and sick leave due to low back pain : a cross-sectional study," *BMJ open*, vol. 9, n°. 5, p. e026917, 2019.

[11] A. Alipour, M. Ghaffari, B. Shariati, I. Jensen et E. Vingard, "Four-year incidence of sick leave because of neck and shoulder pain and its association with work and lifestyle," *Spine*, vol. 34, n°. 4, p. 413–418, févr. 2009.

[12] J. Bué, T. Coutrot, N. Guignon et N. Sandret, "Les facteurs de risques psychosociaux au travail," *Revue francaise des affaires sociales*, n°. 2, p. 45–70, 2008, publisher : La Documentation française. [En ligne]. Disponible : https://www.cairn.info/revue-francaise-des-affaires-sociales-2008-2-page-45.htm

[13] F. W. O'Reilly et A. B. Stevens, "Sickness absence due to influenza," *Occupational Medicine (Oxford, England)*, vol. 52, n°. 5, p. 265–269, août 2002.

[14] C. Pollak, "The impact of a sick pay waiting period on sick leave patterns," *The European journal of health economics : HEPAC : health economics in prevention and care*, vol. 18, n°. 1, p. 13–31, janv. 2017.

[15] S. Thewissen, D. MacDonald, C. Prinz et M. Stricot, "The critical role of paid sick leave in the COVID-19 health and labour market crisis," juill. 2020. [En ligne]. Disponible : https://voxeu.org/article/paid-sick-leave-during-covid-19-health-and-labour-market-crisis

[16] OCDE, "L'investissement dans le capital humain," *Du bien-être des nations, le rôle du capital humain et social*, 2001.

[17] A. D. Langmuir, “The surveillance of communicable diseases of national importance,” *The New England Journal of Medicine*, vol. 268, p. 182–192, janv. 1963.

[18] C. P. Farrington, N. J. Andrews, A. D. Beale et M. A. Catchpole, “A Statistical Algorithm for the Early Detection of Outbreaks of Infectious Disease,” *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, vol. 159, n°. 3, p. 547–563, 1996.

[19] T. Duchemin, A. Bar-Hen, R. Lounissi, W. Dab et M. N. Hocine, “Hierarchizing Determinants of Sick Leave : Insights From a Survey on Health and Well-being at the Workplace,” *Journal of Occupational and Environmental Medicine*, vol. 61, n°. 8, p. e340–e347, 2019.

[20] H. Brborović, Q. Daka, K. Dakaj et O. Brborović, “Antecedents and associations of sickness presenteeism and sickness absenteeism in nurses : A systematic review,” *International Journal of Nursing Practice*, vol. 23, n°. 6, p. e12598, 2017, _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1111/ijn.12598. [En ligne]. Disponible : https: //onlinelibrary.wiley.com/doi/abs/10.1111/ijn.12598

[21] M. M. Davey, G. Cummings, C. V. Newburn-Cook et E. A. Lo, “Predictors of nurse absenteeism in hospitals : a systematic review,” *Journal of Nursing Management*, vol. 17, n°. 3, p. 312–330, 2009, _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2834.2008.00958.x. [En ligne]. Disponible : https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2834.2008.00958.x

[22] L. Daouk-Öyry, A.-L. Anouze, F. Otaki, N. Y. Dumit et I. Osman, “The JOINT model of nurse absenteeism and turnover : A systematic review,” *International Journal of Nursing Studies*, vol. 51, n°. 1, p. 93–110, janv. 2014. [En ligne]. Disponible : http://www.sciencedirect.com/science/article/pii/S0020748913002009

[23] I. A. Steenstra, J. H. Verbeek, M. W. Heymans et P. M. Bongers, “Prognostic factors for duration of sick leave in patients sick listed with acute low back pain : a systematic review of the literature,” *Occupational and Environmental Medicine*, vol. 62, n°. 12, p. 851–860, déc. 2005, publisher : BMJ Publishing Group Ltd Section : Original article. [En ligne]. Disponible : https://oem.bmj.com/content/62/12/851

[24] S. Michie et S. Williams, “Reducing work related psychological ill health and sickness absence : a systematic literature review,” *Occupational and Environmental Medicine*, vol. 60, n°. 1, p. 3–9, janv. 2003, publisher : BMJ Publishing Group Ltd Section : Review. [En ligne]. Disponible : https://oem.bmj.com/content/60/1/3

[25] V. Čikeš, H. Maškarin Ribarić et K. Črnjar, "The Determinants and Outcomes of Absence Behavior : A Systematic Literature Review," *Social Sciences*, vol. 7, n°. 8, p. 120, août 2018, number : 8 Publisher : Multidisciplinary Digital Publishing Institute. [En ligne]. Disponible : https://www.mdpi.com/2076-0760/7/8/120

[26] K. R. Parkes, "Relative weight, smoking, and mental health as predictors of sickness and absence from work," *The Journal of Applied Psychology*, vol. 72, n°. 2, p. 275–286, mai 1987.

[27] M. Wooden, M. Bubonya et D. Cobb-Clark, "Sickness absence and mental health : evidence from a nationally representative longitudinal survey," *Scandinavian Journal of Work, Environment & Health*, vol. 42, n°. 3, p. 201–208, 2016.

[28] C. Høgsbro, M. Davidsen et J. Sørensen, "Long-term sickness absence from work due to physical inactivity : A registry-based study," *Scandinavian Journal of Public Health*, vol. 46, n°. 3, p. 306–313, mai 2018.

[29] D. Antai, A. Oke, P. Braithwaite et D. S. Anthony, "A 'Balanced' Life : Work-Life Balance and Sickness Absence in Four Nordic Countries," *The International Journal of Occupational and Environmental Medicine*, vol. 6, n°. 4, p. 205–222, 2015.

[30] C. Aagestad, H. A. Johannessen, T. Tynes, H. M. Gravseth et T. Sterud, "Work-related psychosocial risk factors for long-term sick leave : a prospective study of the general working population in Norway," *Journal of Occupational and Environmental Medicine*, vol. 56, n°. 8, p. 787–793, août 2014.

[31] C. Aagestad, R. Tyssen, H. A. Johannessen, H. M. Gravseth, T. Tynes et T. Sterud, "Psychosocial and organizational risk factors for doctor-certified sick leave : a prospective study of female health and social workers in Norway," *BMC public health*, vol. 14, p. 1016, sept. 2014.

[32] M. Melchior, N. Krieger, I. Kawachi, L. F. Berkman, I. Niedhammer et M. Goldberg, "Work Factors and Occupational Class Disparities in Sickness Absence : Findings From the GAZEL Cohort Study," *American Journal of Public Health*, vol. 95, n°. 7, p. 1206–1212, juill. 2005. [En ligne]. Disponible : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1449341/

[33] J. I. Halonen, M. Kivimäki, T. Oksanen, P. Virtanen, M. J. Virtanen, J. Pentti et J. Vahtera, "Waterborne Outbreak of Gastroenteritis : Effects on Sick Leaves and Cost of Lost Workdays," *PLOS ONE*, vol. 7, n°. 3, p. e33307, mars 2012, publisher : Public Library of Science. [En ligne]. Disponible : https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0033307

[34] S. J. Pocock, “Relationship between sickness absence and meteorological factors.” *British Journal of Preventive & Social Medicine*, vol. 26, n°. 4, p. 238–245, nov. 1972. [En ligne]. Disponible : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC478727/

[35] T. Lesuffleur, J.-F. Chastang, N. Sandret et I. Niedhammer, “Psychosocial factors at work and sickness absence : results from the French national SUMER survey.” *American journal of industrial medicine*, 2014.

[36] T. Vuorio, S. Suominen, H. Kautiainen et P. Korhonen, “Determinants of sickness absence rate among Finnish municipal employees,” *Scandinavian Journal of Primary Health Care*, vol. 37, n°. 1, p. 3–9, janv. 2019. [En ligne]. Disponible : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6452821/

[37] E. Björk Brämberg, K. Holmgren, U. Bültmann, H. Gyllensten, J. Hagberg, L. Sandman et G. Bergström, “Increasing return-to-work among people on sick leave due to common mental disorders : design of a cluster-randomized controlled trial of a problem-solving intervention versus care-as-usual conducted in the Swedish primary health care system (PROSA),” *BMC public health*, vol. 18, n°. 1, p. 889, 2018.

[38] E. P. Brouwers, B. G. Tiemens, B. Terluin et P. F. Verhaak, “Effectiveness of an intervention to reduce sickness absence in patients with emotional distress or minor mental disorders : a randomized controlled effectiveness trial,” *General Hospital Psychiatry*, vol. 28, n°. 3, p. 223–229, mai 2006. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/pii/S0163834306000363

[39] M. J. Suárez et C. Muñiz, “Unobserved heterogeneity in work absence,” *The European journal of health economics : HEPAC : health economics in prevention and care*, vol. 19, n°. 8, p. 1137–1148, nov. 2018.

[40] G. Heinze, C. Wallisch et D. Dunkler, “Variable selection – A review and recommendations for the practicing statistician,” *Biometrical Journal. Biometrische Zeitschrift*, vol. 60, n°. 3, p. 431–449, mai 2018. [En ligne]. Disponible : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5969114/

[41] C. Strobl, A.-L. Boulesteix, A. Zeileis et T. Hothorn, “Bias in random forest variable importance measures : Illustrations, sources and a solution,” *BMC Bioinformatics*, vol. 8, n°. 1, p. 25, janv. 2007. [En ligne]. Disponible : https://doi.org/10.1186/1471-2105-8-25

[42] T. Hothorn, K. Hornik, C. Strobl et A. Zeileis, "party : A Laboratory for Recursive Partytioning," juin 2020. [En ligne]. Disponible : https://CRAN.R-project.org/package=party

[43] C. Molnar, *Interpretable Machine Learning*, 2019. [En ligne]. Disponible : https://christophm.github.io/interpretable-ml-book/

[44] S. M. Lundberg et S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," dans *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan et R. Garnett, édit. Curran Associates, Inc., 2017, p. 4765–4774. [En ligne]. Disponible : http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[45] A. Gabadinho, G. Ritschard, N. S. Müller et M. Studer, "Analyzing and Visualizing State Sequences in R with TraMineR," *Journal of Statistical Software*, vol. 40, n°. 1, p. 1–37, avr. 2011, number : 1. [En ligne]. Disponible : https://www.jstatsoft.org/index.php/jss/article/view/v040i04

[46] C. H. Elzinga, *Sequence Analysis : Metric Representations of Categorical Time Series*, 2007.

[47] G. Saporta, *Probabilités, analyse des données et statistique*. Editions TECHNIP, août 2011, google-Books-ID : VFoGF97GPiwC.

[48] M. Maechler, P. Rousseeuw, A. Struyf, M. Hubert et K. Hornik, "cluster : Cluster Analysis Basics and Extensions," 2019.

[49] T. Duchemin et M. N. Hocine, "Analysing sickness absence data using semi-Markov models," Barcelona, Spain, 2018.

[50] R. J. Reis, M. Utzet, P. F. La Rocca, F. B. Nedel, M. Martín et A. Navarro, "Previous sick leaves as predictor of subsequent ones," *International Archives of Occupational and Environmental Health*, vol. 84, n°. 5, p. 491–499, juin 2011.

[51] J. Pedersen, J. B. Bjorner, H. Burr et K. B. Christensen, "Transitions between sickness absence, work, unemployment, and disability in Denmark 2004-2008," *Scandinavian Journal of Work, Environment & Health*, vol. 38, n°. 6, p. 516–526, nov. 2012.

[52] I. Oyeflaten, S. A. Lie, C. M. Ihlebæk et H. R. Eriksen, "Multiple transitions in sick leave, disability benefits, and return to work. - A 4-year follow-up of patients participating in a work-related rehabilitation program," *BMC public health*, vol. 12, p. 748, sept. 2012.

[53] A. Burdorf et C. T. J. Hulshof, “Modelling the effects of exposure to whole-body vibration on low-back pain and its long-term consequences for sickness absence and associated work disability,” *Journal of Sound and Vibration*, vol. 298, n°. 3, p. 480–491, déc. 2006. [En ligne]. Disponible : http://www.sciencedirect.com/science/article/pii/S0022460X06004755

[54] M. Lindeboom et M. Kerkhofs, “Multistate models for clustered duration data - An application to workplace effects on individual sickness absenteeism,” *Review of Economics and Statistics*, vol. 82, n°. 4, p. 668–684, nov. 2000, publisher : MIT Press Journals. [En ligne]. Disponible : https://research.vu.nl/en/publications/multistate-models-for-clustered-duration-data-an-application-to-w

[55] A. Król et P. Saint-Pierre, “SemiMarkov : An R Package for Parametric Estimation in Multi-State Semi-Markov Models,” *Journal of Statistical Software*, vol. 66, n°. 1, p. 1–16, août 2015, number : 1. [En ligne]. Disponible : https://www.jstatsoft.org/index.php/jss/article/view/v066i06

[56] V. Damuzzo, L. Agnoletto, L. Leonardi, M. Chiumente, D. Mengato et A. Messori, “Analysis of Survival Curves : Statistical Methods Accounting for the Presence of Long-Term Survivors,” *Frontiers in Oncology*, vol. 9, juin 2019. [En ligne]. Disponible : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6558210/

[57] A. nationale pour l’amélioration des conditions de travail (Anact), *10 questions sur l’absentéisme*, 2016. [En ligne]. Disponible : https://www.anact.fr/10-questions-sur-labsenteisme

[58] F. Juglaret, “Indicateurs et tableaux de bord pour la prévention des risques en santé-sécurité au travail,” phdthesis, Ecole Nationale Supérieure des Mines de Paris, déc. 2012. [En ligne]. Disponible : https://pastel.archives-ouvertes.fr/pastel-00819109

[59] P. Mook, C. Joseph, P. Gates et N. Phin, “Pilot scheme for monitoring sickness absence in schools during the 2006/07 winter in England : can these data be used as a proxy for influenza activity ?” *Eurosurveillance*, vol. 12, n°. 12, p. 11–12, déc. 2007, publisher : European Centre for Disease Prevention and Control. [En ligne]. Disponible : https://www.eurosurveillance.org/content/10.2807/esm.12.12.00755-en

[60] C. K. Cheng, B. J. Cowling, E. H. Lau, L. M. Ho, G. M. Leung et D. K. Ip, “Electronic School Absenteeism Monitoring and Influenza Surveillance, Hong Kong,” *Emerging Infectious Diseases*, vol. 18, n°. 5, p. 885–887, mai 2012.

[61] T. Duchemin, J. Bastard, P. A. Ante-Testard, R. Assab, O. S. Daouda, A. Duval, J.-P. Garsi, R. Lounissi, N. Nekkab, H. Neynaud, D. R. M. Smith, W. Dab, K. Jean, L. Temime et M. N. Hocine, "Monitoring sick leave data for early detection of influenza outbreaks," *medRxiv*, p. 2020.05.28.20115782, mai 2020, publisher : Cold Spring Harbor Laboratory Press.

[62] S. Unkel, C. P. Farrington, P. H. Garthwaite, C. Robertson et N. Andrews, "Statistical methods for the prospective detection of infectious disease outbreaks : a review," *Journal of the Royal Statistical Society : Series A (Statistics in Society)*, vol. 175, n°. 1, p. 49–82, 2012.

[63] G. Bédubourg et Y. Le Strat, "Evaluation and comparison of statistical methods for early temporal detection of outbreaks : A simulation-based study," *PloS One*, vol. 12, n°. 7, p. e0181227, 2017.

[64] D. L. Buckeridge, H. Burkom, M. Campbell, W. R. Hogan et A. W. Moore, "Algorithms for rapid outbreak detection : a research synthesis," *Journal of Biomedical Informatics*, vol. 38, n°. 2, p. 99–113, avr. 2005.

[65] D. G. Enki, P. H. Garthwaite, C. P. Farrington, A. Noufaily, N. J. Andrews et A. Charlett, "Comparison of Statistical Algorithms for the Detection of Infectious Disease Outbreaks in Large Multiple Surveillance Systems," *PLOS ONE*, vol. 11, n°. 8, p. e0160759, août 2016.

[66] P. Farrington et N. Andrews, "Outbreak detection : application to infectious disease surveillance," dans *Monitoring the Health of Populations : Statistical Principles and Methods for Public Health Surveillance*, R. Brookmeyer et D. F. Stroup, édit. New York, NY, USA : OUP USA, déc. 2003, p. 203–231. [En ligne]. Disponible : http://www.us.oup.com/us/catalog/general/subject/Medicine/EpidemiologyBiostatistics/?view=usa&ci=9780195146493

[67] M. L. Jackson, A. Baer, I. Painter et J. Duchin, "A simulation study comparing aberration detection algorithms for syndromic surveillance," *BMC Medical Informatics and Decision Making*, vol. 7, n°. 1, p. 6, mars 2007. [En ligne]. Disponible : https://doi.org/10.1186/1472-6947-7-6

[68] A. Noufaily, R. A. Morbey, F. J. Colón-González, A. J. Elliot, G. E. Smith, I. R. Lake et N. McCarthy, "Comparison of statistical algorithms for daily syndromic surveillance aberration detection," *Bioinformatics (Oxford, England)*, vol. 35, n°. 17, p. 3110–3118, 2019.

[69] M. Salmon, D. Schumacher et M. Höhle, "Monitoring Count Time Series in R : Aberration Detection in Public Health Surveillance," *Journal of Statistical Software*, vol. 70, n°. 1, p. 1–35, mai 2016.

[70] G. Shmueli et H. Burkom, "Statistical Challenges Facing Early Outbreak Detection in Biosurveillance," *Technometrics*, vol. 52, n°. 1, p. 39–51, févr. 2010, publisher : Taylor & Francis. [En ligne]. Disponible : https://amstat.tandfonline.com/doi/abs/10.1198/TECH.2010.06134

[71] C. Sonesson et D. Bock, "A review and discussion of prospective statistical surveillance in public health," *Journal of the Royal Statistical Society : Series A (Statistics in Society)*, vol. 166, n°. 1, p. 5–21, 2003, _eprint : https ://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/1467-985X.00256. [En ligne]. Disponible : https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/1467-985X.00256

[72] D. F. Stroup, G. D. Williamson, J. L. Herndon et J. M. Karon, "Detection of aberrations in the occurrence of notifiable diseases surveillance data," *Statistics in Medicine*, vol. 8, n°. 3, p. 323–329 ; discussion 331–332, mars 1989.

[73] R. E. Serfling, "Methods for current statistical analysis of excess pneumonia-influenza deaths," *Public Health Reports*, vol. 78, n°. 6, p. 494–506, juin 1963. [En ligne]. Disponible : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1915276/

[74] R. A. Parker, "Analysis of surveillance data with poisson regression : A case study," *Statistics in Medicine*, vol. 8, n°. 3, p. 285–294, 1989, _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.4780080309. [En ligne]. Disponible : https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4780080309

[75] A. Noufaily, D. G. Enki, P. Farrington, P. Garthwaite, N. Andrews et A. Charlett, "An improved algorithm for outbreak detection in multiple surveillance systems," *Statistics in Medicine*, vol. 32, n°. 7, p. 1206–1222, mars 2013.

[76] J. Manitz et M. Höhle, "Bayesian outbreak detection algorithm for monitoring reported cases of campylobacteriosis in Germany," *Biometrical Journal. Biometrische Zeitschrift*, vol. 55, n°. 4, p. 509–526, juill. 2013.

[77] K. Kleinman, R. Lazarus et R. Platt, "A Generalized Linear Mixed Models Approach for Detecting Incident Clusters of Disease in Small Areas, with an Application to Biological Terrorism," *American Journal of Epidemiology*, vol. 159, n°. 3, p. 217–224, févr. 2004, publisher : Oxford Academic. [En ligne]. Disponible : https: //academic.oup.com/aje/article/159/3/217/79584

[78] R. A. Morbey, A. J. Elliot, A. Charlett, N. Q. Verlander, N. Andrews et G. E. Smith, "The application of a novel 'rising activity, multi-level mixed effects, indicator emphasis' (RAMMIE)

method for syndromic surveillance in England," *Bioinformatics (Oxford, England)*, vol. 31, n°. 22, p. 3660–3665, nov. 2015.

[79] L. Stern et D. Lightfoot, "Automated outbreak detection : a quantitative retrospective analysis." *Epidemiology and Infection*, vol. 122, n°. 1, p. 103–110, févr. 1999. [En ligne]. Disponible : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2809594/

[80] S. C. Wieland, J. S. Brownstein, B. Berger et K. D. Mandl, "Automated real time constant-specificity surveillance for disease outbreaks," *BMC Medical Informatics and Decision Making*, vol. 7, juin 2007. [En ligne]. Disponible : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1919360/

[81] J. Zhang, F.-C. Tsui, M. M. Wagner et W. R. Hogan, "Detection of outbreaks from time series data using wavelet transform," *AMIA ... Annual Symposium proceedings. AMIA Symposium*, p. 748–752, 2003.

[82] B. Y. Reis et K. D. Mandl, "Time series modeling for syndromic surveillance," *BMC medical informatics and decision making*, vol. 3, p. 2, janv. 2003.

[83] T. M. Rath, M. Carreras et P. Sebastiani, "Automated Detection of Influenza Epidemics with Hidden Markov Models," dans *Advances in Intelligent Data Analysis V*, ser. Lecture Notes in Computer Science, M. R. Berthold, H.-J. Lenz, E. Bradley, R. Kruse et C. Borgelt, édit. Berlin, Heidelberg : Springer, 2003, p. 521–532.

[84] F. F. Gan, "Design of Optimal Exponential CUSUM Control Charts," *Journal of Quality Technology*, vol. 26, n°. 2, p. 109–124, avr. 1994, publisher : Taylor & Francis _eprint : https ://doi.org/10.1080/00224065.1994.11979511. [En ligne]. Disponible : https://doi.org/10.1080/00224065.1994.11979511

[85] L. Hutwagner, W. Thompson, G. M. Seeman et T. Treadwell, "The bioterrorism preparedness and response Early Aberration Reporting System (EARS)," *Journal of Urban Health : Bulletin of the New York Academy of Medicine*, vol. 80, n°. Suppl 1, p. i89–i96, mars 2003. [En ligne]. Disponible : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3456557/

[86] F. F. Gan, "Monitoring observations generated from a binomial distribution using modified exponentially weighted moving average control chart," *Journal of Statistical Computation and Simulation*, vol. 37, n°. 1-2, p. 45–60, oct. 1990, publisher : Taylor

& Francis _eprint : https ://doi.org/10.1080/00949659008811293. [En ligne]. Disponible : https://doi.org/10.1080/00949659008811293

[87] A. Schuh, J. A. Camelio et W. H. Woodall, "Control charts for accident frequency : a motivation for real-time occupational safety monitoring," *International Journal of Injury Control and Safety Promotion*, vol. 21, n°. 2, p. 154–162, avr. 2014, publisher : Taylor & Francis _eprint : https ://doi.org/10.1080/17457300.2013.792285.

[88] M. Kulldorff, "Prospective time periodic geographical disease surveillance using a scan statistic," *Journal of the Royal Statistical Society : Series A (Statistics in Society)*, vol. 164, n°. 1, p. 61–72, 2001, _eprint : https ://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/1467-985X.00186. [En ligne]. Disponible : https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/1467-985X.00186

[89] A. Guillou, M. Kratz et Y. L. Strat, "An extreme value theory approach for the early detection of time clusters. A simulation-based assessment and an illustration to the surveillance of Salmonella," *Statistics in Medicine*, vol. 33, n°. 28, p. 5015–5027, 2014, _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.6275. [En ligne]. Disponible : https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.6275

[90] P. A. Rogerson, "Monitoring point patterns for the development of space–time clusters," *Journal of the Royal Statistical Society : Series A (Statistics in Society)*, vol. 164, n°. 1, p. 87–96, 2001, _eprint : https ://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/1467-985X.00188. [En ligne]. Disponible : https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/1467-985X.00188

[91] A. Lawson, A. Clark et C. V. Rodeiro, "Developments in General and Syndromic Surveillance for Small Area Health Data," *Journal of Applied Statistics*, vol. 31, n°. 8, p. 951–966, oct. 2004, publisher : Taylor & Francis _eprint : https ://doi.org/10.1080/0266476042000270568. [En ligne]. Disponible : https://doi.org/10.1080/0266476042000270568

[92] P. Sebastiani, K. D. Mandl, P. Szolovits, I. S. Kohane et M. F. Ramoni, "A Bayesian dynamic model for influenza surveillance," *Statistics in Medicine*, vol. 25, n°. 11, p. 1803–1816 ; discussion 1817–1825, juin 2006.

[93] W. Ku, R. H. Storer et C. Georgakis, "Disturbance detection and isolation by dynamic principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 30, n°. 1, p. 179–196, nov. 1995. [En ligne]. Disponible : http://www.sciencedirect.com/science/article/pii/0169743995000763

[94] M. Kulldorff, F. Mostashari, L. Duczmal, W. Katherine Yih, K. Kleinman et R. Platt, "Multivariate scan statistics for disease surveillance," *Statistics in Medicine*, vol. 26, n°. 8, p. 1824–1833, avr. 2007.

[95] A. Flahault, T. Blanchon, Y. Dorléans, L. Toubiana, J. F. Vibert et A. J. Valleron, "Virtual surveillance of communicable diseases : a 20-year experience in France," *Statistical Methods in Medical Research*, vol. 15, n°. 5, p. 413–421, oct. 2006.

[96] D. D. Lenaway et A. Ambler, "Evaluation of a school-based influenza surveillance system." *Public Health Reports*, vol. 110, n°. 3, p. 333–337, 1995. [En ligne]. Disponible : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1382129/

[97] J. Szecsenyi, H. Uphoff, S. Ley et H. D. Brede, "Influenza surveillance : experiences from establishing a sentinel surveillance system in Germany." *Journal of Epidemiology & Community Health*, vol. 49, n°. Suppl 1, p. 9–13, août 1995, publisher : BMJ Publishing Group Ltd Section : Research Article. [En ligne]. Disponible : https://jech.bmj.com/content/49/Suppl_1/9

[98] T. Duchemin et M. N. Hocine, "Modeling sickness absence data : A scoping review," *PLOS ONE*, vol. 15, n°. 9, p. e0238981, sept. 2020, publisher : Public Library of Science. [En ligne]. Disponible : https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0238981

[99] I. H. Montano, G. Marques, S. G. Alonso, M. López-Coronado et I. de la Torre Díez, "Predicting Absenteeism and Temporary Disability Using Machine Learning : a Systematic Review and Analysis," *Journal of Medical Systems*, vol. 44, n°. 9, p. 162, août 2020. [En ligne]. Disponible : https://doi.org/10.1007/s10916-020-01626-2

[100] A. Leplège et J. Coste, *Mesure de la santé perceptuelle et de la qualité de vie : méthodes et applications*. Paris : ESTEM, 2002, oCLC : 53802063.

[101] J. M. Gran, S. A. Lie, I. Øyeflaten, r. Borgan et O. O. Aalen, "Causal inference in multi-state models–sickness absence and work for 1145 participants after work rehabilitation," *BMC Public Health*, vol. 15, n°. 1, p. 1082, oct. 2015. [En ligne]. Disponible : https://doi.org/10.1186/s12889-015-2408-8

[102] N. Maltzahn, R. Hoff, O. O. Aalen, I. S. Mehlum, H. Putter et J. M. Gran, "A hybrid landmark Aalen-Johansen estimator for transition probabilities in partially non-Markov multi-state models," *arXiv :2007.00974 [stat]*, août 2020, arXiv : 2007.00974. [En ligne]. Disponible : http://arxiv.org/abs/2007.00974

[103] M.-N. Tran, N. Nguyen, D. Nott et R. Kohn, "Bayesian Deep Net GLM and GLMM," *Journal of Computational and Graphical Statistics*, vol. 29, n°. 1, p. 97–113, janv. 2020, publisher : Taylor & Francis. [En ligne]. Disponible : https://amstat.tandfonline.com/doi/abs/10.1080/10618600.2019.1637747

Annexe A

Statistiques descriptives provenant du Baromètre Santé et bien-être au Travail

	General population	No SL	Very short-term SL	Short-term SL	Long-term SL
n	32327	20665	4128	5187	2347
Sex					
Female	46%	44%	50%	50%	51%
Male	54%	56%	50%	50%	49%
Age					
<30y	21%	20%	29%	20%	14%
30-39y	30%	29%	36%	31%	25%
40-44y	15%	15%	14%	15%	16%
45-49y	13%	14%	8%	13%	14%
50-55y	14%	14%	9%	14%	20%
?56y	8%	8%	5%	7%	12%
Living alone with children					
Yes	94%	94%	95%	94%	92%
No	6%	6%	5%	6%	8%
Having a dependent family member					
Yes	15%	15%	13%	14%	18%
No	85%	85%	87%	86%	82%
Full-time or part-time job					
Full-time	88%	87%	90%	89%	84%
Part-time	12%	13%	10%	11%	16%
Shift work					
Yes	22%	21%	18%	26%	32%
No	78%	79%	82%	74%	68%
At work, do you have to stay standing or in a painful posture :					
Most of the time	27%	26%	21%	33%	43%
Occasionally	17%	17%	15%	16%	19%
Rarely	21%	22%	22%	20%	17%
Never	34%	35%	41%	30%	21%

At work, do you have to carry heavy load :					
Most of the time	34%	35%	41%	30%	21%
Occasionally	15%	14%	12%	19%	29%
Rarely	22%	22%	18%	23%	25%
Never	24%	24%	23%	23%	20%
At work, do you have to perform repetitive movement :					
Most of the time	30%	28%	27%	36%	47%
Occasionally	24%	25%	24%	25%	22%
Rarely	21%	22%	22%	19%	16%
Never	24%	26%	27%	19%	15%
At work, do you have to work in front of a screen :					
Most of the time	59%	59%	70%	57%	45%
Occasionally	16%	16%	12%	15%	18%
Rarely	10%	9%	7%	11%	14%
Never	16%	16%	11%	17%	23%
At work, you have to work in the heat/the coldness :					
Most of the time	19%	18%	15%	23%	30%
Occasionally	21%	21%	18%	22%	23%
Rarely	22%	22%	24%	22%	20%
Never	38%	39%	43%	33%	27%
At work, do you have to work in a noisy environment :					
Most of the time	26%	23%	24%	34%	40%
Occasionally	27%	28%	27%	28%	25%
Rarely	26%	27%	27%	22%	20%
Never	21%	23%	21%	16%	14%
At work, do you breathe toxic products and dust :					
No	29%	27%	26%	35%	39%
Yes	71%	73%	74%	65%	61%
At work, do you handle toxic or dangerous products :					

No	20%	19%	17%	23%	28%
Yes	80%	81%	83%	77%	72%
At work, do you risk severe fall :					
No	19%	18%	15%	22%	30%
Yes	81%	82%	85%	78%	70%
At work, do you work on machine that could injure you ?					
No	24%	22%	20%	30%	36%
Yes	76%	78%	80%	70%	64%
I can learn new things with my job.					
Completely agree	21%	22%	23%	19%	17%
Slightly agree	7%	7%	6%	9%	14%
Slightly disagree	49%	50%	51%	46%	41%
Completely disagree	22%	21%	20%	26%	28%
My job asks me to work very quickly or very intensively					
Completely agree	23%	21%	23%	26%	33%
Slightly agree	47%	48%	49%	47%	42%
Slightly disagree	25%	26%	24%	23%	20%
Completely disagree	5%	6%	4%	4%	5%
My job asks me long period of concentration.					
Completely agree	26%	25%	29%	27%	30%
Slightly agree	50%	51%	52%	48%	45%
Slightly disagree	20%	20%	17%	21%	20%
Completely disagree	4%	4%	2%	4%	6%
My job is physically tiring.					
Completely agree	19%	17%	14%	24%	35%
Slightly agree	26%	27%	23%	27%	28%
Slightly disagree	35%	36%	40%	32%	24%
Completely disagree	19%	20%	23%	17%	13%
My job is nervously tiring.					

Completely agree	26%	23%	25%	31%	39%
Slightly agree	43%	43%	46%	43%	39%
Slightly disagree	26%	28%	25%	22%	17%
Completely disagree	5%	6%	4%	4%	4%
I am satisfied of my work.					
Completely agree	21%	22%	19%	19%	16%
Slightly agree	57%	58%	60%	53%	49%
Slightly disagree	17%	15%	17%	22%	23%
Completely disagree	5%	4%	4%	7%	12%
At work, I have opportunities to take decision.					
Completely agree	27%	28%	25%	25%	23%
Slightly agree	47%	48%	49%	45%	41%
Slightly disagree	20%	18%	20%	22%	22%
Completely disagree	7%	6%	6%	8%	14%
At work, I have difficulties to handle priorities.					
Completely agree	7%	7%	8%	8%	9%
Slightly agree	29%	29%	33%	30%	27%
Slightly disagree	45%	46%	44%	45%	43%
Completely disagree	18%	18%	15%	17%	20%
I feel acknowledged by my superiors.					
Completely agree	14%	15%	13%	11%	10%
Slightly agree	44%	47%	47%	40%	32%
Slightly disagree	28%	26%	29%	32%	32%
Completely disagree	14%	12%	12%	18%	27%
At work, I have opportunities to develop my professional skills.					
Completely agree	18%	18%	18%	15%	13%
Slightly agree	47%	49%	48%	42%	37%
Slightly disagree	26%	25%	26%	31%	31%
Completely disagree	9%	8%	7%	11%	19%

At work, there is a good atmosphere					
Completely agree	24%	26%	25%	21%	17%
Slightly agree	55%	56%	57%	54%	48%
Slightly disagree	15%	14%	15%	18%	23%
Completely disagree	5%	4%	4%	6%	12%
At work, I have a well-fitted workstation.					
Completely agree	15%	16%	15%	13%	11%
Slightly agree	58%	59%	62%	54%	47%
Slightly disagree	21%	19%	20%	26%	30%
Completely disagree	5%	5%	4%	6%	12%
I am always trying to improve my way of working.					
Completely agree	25%	25%	24%	25%	25%
Slightly agree	62%	62%	63%	62%	60%
Slightly disagree	2%	2%	1%	2%	3%
Completely disagree	11%	11%	11%	11%	11%
I am confident in the future of my company.					
Completely agree	14%	15%	14%	12%	11%
Slightly agree	49%	51%	51%	46%	37%
Slightly disagree	26%	24%	26%	28%	31%
Completely disagree	11%	9%	9%	14%	21%
My company carry about the well-being of its workers.					
Completely agree	9%	10%	8%	7%	6%
Slightly agree	44%	46%	45%	38%	33%
Slightly disagree	34%	32%	35%	39%	36%
Completely disagree	14%	12%	12%	17%	25%
Security is a priority for my company.					
Completely agree	22%	23%	19%	22%	21%
Slightly agree	48%	49%	49%	46%	44%
Slightly disagree	22%	21%	25%	24%	23%

Completely disagree	7%	7%	7%	8%	12%
At work, do you have lack of attention, lower vigilance					
Never	16%	19%	11%	11%	12%
Sometimes	66%	65%	70%	69%	62%
Often	14%	13%	15%	16%	19%
Very often	4%	3%	4%	4%	7%
At work, you come only to show up :					
Never	52%	55%	48%	47%	47%
Sometimes	35%	33%	39%	39%	34%
Often	9%	9%	9%	10%	12%
Very often	4%	3%	4%	4%	7%
Do you have difficulties to conciliate work and personal life ?					
Completely agree	7%	6%	7%	8%	11%
Slightly agree	26%	24%	27%	28%	28%
Slightly disagree	46%	46%	49%	45%	41%
Completely disagree	22%	24%	17%	20%	19%
How long is your transport time between work and home (round trip) ?					
Less than 1 hour	67%	69%	63%	65%	65%
Between 1 and 2 hours	26%	25%	30%	27%	28%
More than 2 hours	7%	7%	7%	8%	7%
How do you consider your health condition ?					
Very good	11%	14%	11%	6%	4%
Good	59%	63%	61%	52%	34%
Average	26%	21%	25%	36%	43%
Bad	4%	2%	3%	5%	16%
Very bad	0%	0%	0%	0%	3%
Do you have a chronic disease ?					
Yes	20%	15%	19%	27%	48%
No	80%	85%	81%	73%	52%

Do you have a disability ?					
Yes	6%	4%	4%	8%	23%
No	94%	96%	96%	92%	77%
Over the last 12 months, did you have head pain :					
Permanently	2%	1%	2%	3%	4%
Often	21%	18%	25%	27%	27%
Sometimes	59%	60%	59%	57%	54%
Never	18%	21%	14%	13%	16%
Over the last 12 months, did you have neck, shoulder or arm pain :					
Permanently	5%	4%	5%	7%	16%
Often	23%	20%	24%	30%	31%
Sometimes	45%	46%	47%	44%	37%
Never	27%	30%	24%	19%	17%
Over the last 12 months, did you have back pain :					
Permanently	8%	5%	7%	11%	21%
Often	49%	51%	50%	43%	34%
Sometimes	28%	25%	30%	36%	35%
Never	16%	19%	14%	10%	10%
Over the last 12 months, did you have leg, foot or knee pain :					
Permanently	4%	3%	2%	5%	12%
Often	15%	13%	15%	20%	26%
Sometimes	41%	40%	43%	41%	38%
Never	40%	44%	40%	33%	23%
Over the last 12 months, did you have body pain (in parts other than head, arm, shoulder, neck, back, leg, foot and knee) :					
Permanently	2%	1%	2%	2%	6%
Often	54%	58%	53%	46%	37%
Sometimes	8%	6%	7%	10%	16%
Never	37%	35%	39%	42%	40%
Do you take antidepressants, anxiolytics or sleeping pills :					

Never	87%	90%	87%	81%	71%
Less than once a month	5%	4%	6%	7%	7%
Once a month	1%	1%	1%	2%	2%
More than once a month	2%	1%	2%	3%	3%
At least once a week	5%	3%	4%	7%	18%
Over the last 12 months, did you have sleep issues					
Permanently	7%	6%	7%	10%	17%
Often	22%	19%	22%	28%	33%
Sometimes	48%	49%	51%	47%	38%
Never	22%	26%	20%	15%	12%
Over the last 12 months, did you have lack of tonicity :					
Permanently	4%	3%	3%	6%	12%
Often	20%	16%	23%	28%	32%
Sometimes	60%	61%	62%	58%	47%
Never	16%	20%	12%	9%	8%
Business sector					
Consulting and engineering	5%	5%	7%	3%	2%
Construction industry	8%	8%	8%	9%	9%
Commerce	16%	16%	17%	17%	17%
Manufacturing	25%	24%	24%	29%	28%
Health	9%	9%	8%	9%	11%
Services	23%	24%	23%	20%	20%
Transport, telecom and energy	14%	14%	13%	13%	13%
Size of the company					
More than 1000 workers	39%	37%	42%	40%	41%
49 to 999 workers	24%	23%	25%	28%	28%
10 to 49 workers	12%	13%	12%	12%	12%
Less than 10 workers	25%	27%	20%	19%	19%
French administrative regions					

Alsace	7%	7%	5%	6%	7%
Auvergne Rhônes Alpes	10%	10%	11%	11%	12%
Bourgogne Franche Comté	13%	13%	13%	13%	12%
Bretagne	3%	3%	3%	3%	3%
Centre Val de Loire	7%	7%	9%	8%	6%
Corse	1%	1%	1%	1%	1%
Ile de France	13%	12%	16%	15%	11%
Normandie	11%	11%	12%	10%	9%
Nord Pas De Calais Picardie	6%	5%	5%	6%	6%
Nouvelle Aquitaine	7%	7%	8%	7%	9%
Occitanie	11%	12%	9%	10%	11%
Provence Alpes Côte d'Azur	6%	6%	6%	6%	7%
Pays de la Loire	5%	5%	4%	5%	5%
Number of children at home					
0	48%	49%	49%	46%	43%
1	21%	19%	22%	23%	22%
2	21%	22%	21%	21%	23%
3+	10%	10%	8%	10%	12%

TABLE A.1 – Description des données du Baromètre Santé et bien-être au Travail. La table est une annexe de l'article *Hierachizing sick leave determinants* et est donc écrit en anglais.

Annexe B

Publication 4 : surveiller les données d'absence pour détecter les épidémies de grippe

Monitoring sick leave data for early detection of influenza outbreaks

Tom Duchemin[1,2], Jonathan Bastard[1,3,4,5], Pearl Anne Ante-Testard,[1, 4], Rania Assab[1], Oumou Salama Daouda[1], Audrey Duval[1,3,5,7], Jérôme-Philippe Garsi[1], Radowan Lounissi[2], Narimane Nekkab[1, 6], Helene Neynaud[1], David R. M. Smith[1,3,5], William Dab[1], Kevin Jean[1, 4], Laura Temime[1, 4], Mounia N. Hocine[1]

[1] MESuRS laboratory, Conservatoire National des Arts et Métiers, 292 Rue Saint-Martin, 75003 Paris, France
[2] Malakoff Humanis, 21 Rue Laffitte, 75009 Paris, France.
[3] Institut Pasteur, Epidemiology and Modelling of Antibiotic Evasion (EMAE), Paris, France
[4] PACRI unit, Conservatoire National des Arts et Métiers, Institut Pasteur, Paris, France
[5] Université Paris-Saclay, UVSQ, Inserm, CESP, Anti-infective evasion and pharmacoepidemiology team, Montigny-Le-Bretonneux, France
[6]Malaria: Parasites and Hosts, Department of Parasites and Insect Vectors, Institut Pasteur, Paris, France
[7]Biodiversity and Epidemiology of Bacterial Pathogens, Institut Pasteur, Paris, France

Abstract

Background - Workplace absenteeism increases significantly during influenza epidemics. Sick leave records may facilitate more timely detection of influenza outbreaks, as trends in increased sick leave may precede alerts issued by sentinel surveillance systems by days or weeks. Sick leave data have not been comprehensively evaluated in comparison to traditional surveillance methods.

Aim - To study the performance and the feasibility of using a detection system based on sick leave data to detect influenza outbreaks

Methods - Sick leave records were extracted from private French health insurance data, covering on average 209,932 companies per year across a wide range of sizes and sectors. We used linear regression to estimate the weekly number of new sick leave spells from 2010 to 2017 in 12 French regions, adjusting for trend, seasonality and worker leaves. Outbreaks were detected using a 95%-prediction interval. This method was compared to results from the French Sentinelles network, a gold-standard primary care surveillance system currently in place.

Results - Using sick leave data, we detected 92% of reported influenza outbreaks between 2016 and 2017, on average 5.88 weeks prior to outbreak peaks. Compared to the existing Sentinelles model, our method had high sensitivity (89%) and specificity (86%), and detected outbreaks on average 2.5 weeks earlier.

Conclusion - Sick leave surveillance could be a sensitive, specific and timely tool for detection of influenza outbreaks.

1. Introduction

Early outbreak detection is crucial for preparedness and timely public health and medical responses. It provides useful information to physicians, companies, and the public, ensuring proper drug prescription, health service planning, workplace preparedness, and continuity of operations in case of high absenteeism (1), among many other uses.

Most countries face periodic influenza (or “flu”) epidemics that vary in size and severity from year to year (2). Seasonal flu can be highly virulent and, like many respiratory viruses, can spread rapidly through populations highlighting a need for a robust epidemiological surveillance system to detect emerging outbreaks. Surveillance system guidelines developed by the US Centers for Disease Control (CDC) suggest that systems should be simple, reliable, flexible, timely, and readily accepted by diverse individuals and organizations to ensure participation (3).

Flu surveillance systems vary by country and rely on various types of data. National health agencies monitor flu epidemics using healthcare records, medical sentinel systems, pharmaceutical sales and other data sources. Most of these systems rely on data from healthcare settings that rely on patient healthcare seeking behavior or after results of clinical tests, which often reflects those with symptoms or relatively advanced stages of the disease. These systems fail to capture individuals who do not seek medical care, whether due to asymptomatic infection, perceived mildness of infection or a general reluctance to seek care (1,4). Given these gaps, alternative data streams from non-healthcare settings may provide a valuable complement to classical surveillance systems (5).

Human resources data collected at the workplace have received relatively little attention for outbreak detection, but present characteristics that are useful for infectious disease surveillance. In many settings, absenteeism data are routinely collected and centralized either for the use by companies themselves or for health insurance purposes. Due to legal purposes and to the implication of work absence on salaries, these data are comprehensive and reliable. Though a handful of studies have assessed the role of sick leave data for outbreak detection, they did not develop a comprehensive assessment of its performance and robustness. Bollaerts et al., Patterson et al. and Groenewold et al. (1,6,7) assessed the usefulness of work absenteeism surveillance as a tool for early warning systems for influenza. They can also supplement more traditional medical data by providing information about an epidemic’s socioeconomic impact (1,4). As a consequence, the US National Institute for Occupational Safety and Health (NIOSH) has been monitoring health-related workplace absenteeism among full-time workers using data received monthly from the Current Population Survey since 2017 and making this data available online (8).

A great challenge in epidemiological surveillance lies in identifying data streams that allow for sensitive, specific and timely outbreak detection. In the context of outbreak detection, surveillance sensitivity can refer to both (i) the proportion of true cases detected, and (ii) the probability of detecting an outbreak, including the changes in the number of cases over time (3). Surveillance specificity refers to the probability of correctly identifying when an outbreak is not occurring (9). Lastly, surveillance timeliness generally refers to the time difference between an event and its standard reference(5). Some studies have suggested that

absenteeism data along with others such as over-the-counter pharmaceutical sales and emergency visits seem to be more timely than sentinel Influenza-Like Illness (ILI) surveillance (5,6), other traditional flu data sources (7), physician diagnoses (5), and virological data (5).

In this study, we assess the sensitivity, specificity and timeliness of workplace absenteeism data for detection of flu outbreaks in France. Our hypothesis is that monitoring sick-leave data at the workplace might help anticipate outbreaks in a timely manner using routinely collected data. We then compare the performance of a sick-leave based monitoring system to the performance of the national standard surveillance system of influenza in France which is based on ILI data.

2. Material and methods

2.1. Data

2.1.1 Sick-leave data

The study relies on the sick leave record system of the French health insurance company Malakoff Médéric. Malakoff Médéric insures sick leaves for 114,707 to 245,973 French companies in a wide range of sectors, covering between 290,056 and 2,765,400 employees per year. This wide variation is due to the fact that some companies were no longer required to report these data to the insurer after 2015. The companies have in average 18.7 employees per year. Nearly half of companies (46%) were in Services, 36% in Commerce, 12% in Industry and Construction, and 5% in Health.

These data are routinely collected and annually reported in the system DADS (*Déclarations Annuelles de Données Sociales*). For our purposes, we used the weekly incidence rate of sick leave spells (per 100,000 workers) aggregated at the regional level across the 12 administrative regions of metropolitan France over the period 2010-2017.

2.1.2 Workers leave data

The number of workers on non-sick leave (e.g. paid holiday) is not reported in DADS, so the denominator of the weekly sick leave incidence rate was defined as the number of workers actively employed by their company during the observed week, and not the number of workers actually working during the week. To adjust our data, we used data from statistics department of the French Ministry of labour (DARES) to build an indicator (the worker-leave-peak indicator) describing weeks with a peak in sick leave (10). Peaks were identified during the Christmas school holidays (last week of December and first week of January) and during summer (second week of July to third week of August).

2.1.3 Influenza-like illness data

Weekly sick leave incidence was compared to weekly ILI incidence (per 100,000 inhabitants), derived from the French influenza surveillance database of the GP Sentinelles network,

coordinated by Santé Publique France. In 2018, the Sentinelles network was composed of 1,314 general private practitioners and 116 private pediatricians, all voluntary participants and spread widely across the whole of France's territories. Detailed information on this network can be found elsewhere (11). This ILI incidence data is the main source of data used to declare influenza epidemics in France. ILI are defined by Sentinelles as a fever above 39°C, with sudden onset, accompanied by myalgia and respiratory signs.

In addition to providing weekly ILI incidence data, the French Sentinelles network also proposes an ILI-outbreak detection algorithm. The algorithm is based on a Serfling method (12). It has been adapted for routine surveillance of epidemics of ILI in France (13). The method implemented by Sentinelles is based on a periodic regression model including a biannual seasonal effect of clinical influenza, a linear trend and an intercept adjusting for a baseline diagnostic activity (which corresponds to the number of influenza syndromes that would be diagnosed in the absence of influenza virus during the off-season).

2.2. Methods

2.2.1 Identification of influenza outbreak episodes

Dates of influenza outbreak episodes from Sentinelles are publicly unavailable so we trained the model described above on data from 1984 to 2009 to mimic the sentinel system. Weekly ILI incidence rates per 100,000 residents and per region from 2010 to 2017 were then compared to an outbreak detection threshold. This was defined as the upper bound of the 95% prediction interval from this model, and to increase specificity an alert was only declared when this threshold was crossed twice consecutively.

2.2.2 Determination of sick leave outbreak episodes

To detect sick leave outbreak episodes, an algorithm based on the Serfling method was used. The regression includes an intercept to adjust for the baseline sick-leave activity and the worker-leave-peak indicator to adjust for seasonality.

Similarly to the Sentinelles method, an outbreak was declared if the true sick leave incidence rate crossed the 95% prediction interval twice consecutively. An alert was lifted when the incidence fell below the threshold, again for two consecutive weeks (14).

The model was trained on sick leave data from 2010 to 2015. Years 2016 and 2017 were used to evaluate model performance. As timing of ILI outbreaks may vary geographically, analyses were conducted separately for each of mainland France's 12 administrative regions. For simplicity, some results were plotted in the main text for three regions only, chosen to reflect a North-South gradient (respectively Haut-de-France, Ile-de-France and Provence-Alpes-Côte d'Azur). Full results are included as supplementary results.

2.2.3 Criteria for assessing the proposed surveillance system

Evaluation criteria were selected to answer two questions: (i) Does the sick leave model efficiently detect ILI outbreaks? and (ii) How does it compare to the Sentinelles model? Results

are presented for each French administrative region and are also aggregated at the national level.

Performance of the sick leave model to detect ILI outbreaks

To answer the first question, we calculated sensitivity and specificity per outbreak episode to evaluate whether our model correctly detected all ILI outbreaks. The two criteria are:

$$Sensitivity_{per\ episode} = \frac{Number\ of\ Influenza\ outbreaks\ detected\ by\ the\ Sick\ Leave\ model}{Number\ of\ Influenza\ outbreaks}$$

$$Specificity_{per\ episode} = \frac{Number\ of\ Sick\ Leave\ outbreaks\ crossing\ an\ Influentza\ outbreak}{Number\ of\ Sick\ Leave\ outbreaks}$$

To evaluate our model's timeliness, we calculated the detection time we defined as the delay between the outbreak detection of our algorithm compared to the annual influenza outbreak peak. The influenza outbreak peak was defined as the week with the highest number of reported ILI cases between June 1st and May 31st of the subsequent year.

Performance of the sick leave model compared to the Sentinelles model

To evaluate the performance of the sick leave model, we calculated its sensitivity and specificity with respect to the Sentinelles model. Unlike the previous criteria, we define these indicators at the level of the week rather than the episode. The objective is to assess whether our two models are similar. The two criteria are defined as follows:

$$Sensitiviy_{Sentinelles} = \frac{Number\ of\ weeks\ with\ a\ Sentinelles\ alert\ and\ with\ a\ Sick\ Leave\ alert}{Number\ of\ weeks\ with\ a\ Sentinelles\ alert}$$

$$Specificity_{Sentinelles} = \frac{Number\ of\ weeks\ with\ no\ Sentinelles\ alert\ and\ with\ no\ Sick\ Leave\ alert}{Number\ of\ weeks\ with\ no\ Sentinelles\ alert}$$

The timeliness of the sick-leave model compared to the Sentinelles model was also evaluated: the delay between the outbreak detection of the first model compared to the second one is computed.

3. Results

Incidence curves obtained from the Sentinelles surveillance networks from 2010 to 2017 reveal annual peaks of influenza-like illness (ILI), from 163 to 1,290 per 100,000 per week, occurring approximately between December and February (Figures 1 and S1). During the summers, the incidence approaches zero. By comparison, weekly sick leave incidence varied about an annual average of 1,021 to 1,335 per 100,000 per week, depending on the region (Figures 1 and S1). They exhibit greater variability and more peaks per year than ILI incidence. However, based on visual inspection, the highest seasonal peaks tend to coincide with ILI incidence peaks. Moreover, most of the seasonal sick-leave troughs coincide with Christmas

and summer school holidays periods, which can be explained by a decrease of the at-risk population, i.e. an increase in the number of workers on paid leave. Finally, there is no apparent change between 2014 and 2015 despite the strong variation in the volume of workers in the database.

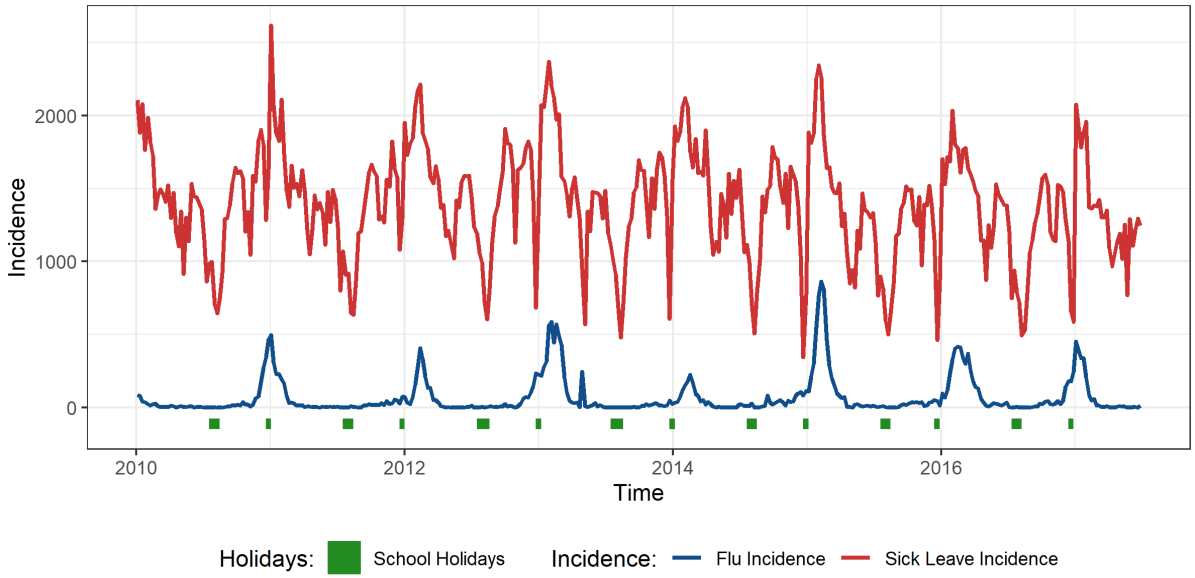

Figure 1: Incidence per 100,000 per week of influenza-like illness and sick leave in Ile-de-France, the most populous region in France, 2010-2017. Christmas and summer school holidays (increased worker leave periods) are shown at the bottom.

For three French regions, Figure 2 presents the incidence of sick leaves and ILI for the 2015-2017 time period. In each region, the Sentinelles surveillance system identified exactly one alert per year, triggered a few weeks before or during the peak of ILI incidence (Figures 2 and S2). The exception was Bretagne, where no alert was identified during winter 2016-2017 (Figure S2). By comparison, the sick leave surveillance system triggered one to three alert episodes per year.

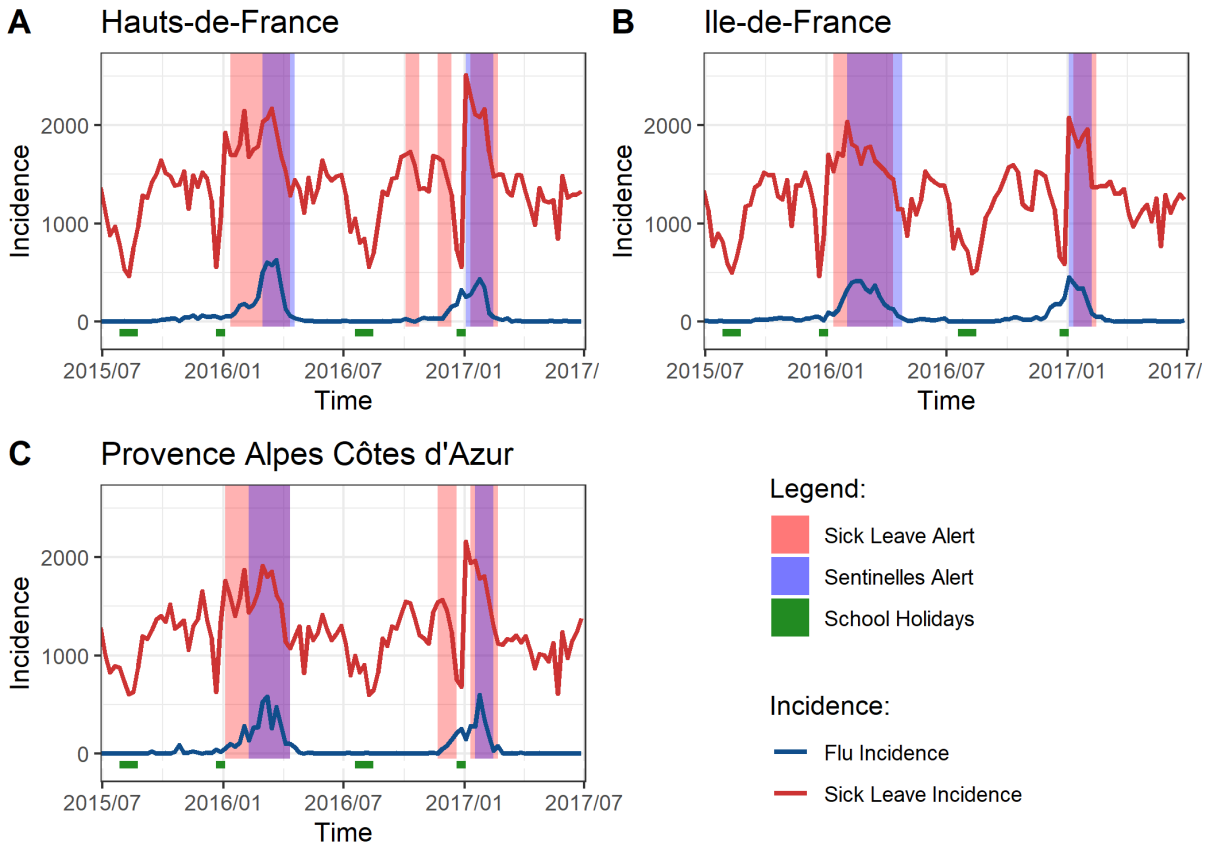

Figure 2: Incidence per 100,000 per week of influenza-like illness and sick leave, 2015-2017, and alerts from the Sentinelles (blue) and the sick-leave (red) models, in three French regions (A: Hauts-de-France, B: Ile-de-France, C: Provence-Alpes-Côte d'Azur). Alerts show as purple when Sentinelles and sick-leave alerts overlap. Christmas and summer school holidays (increased worker leave periods) are shown at the bottom (green).

We assessed the sick leave surveillance system on (i) its ability to detect and anticipate ILI incidence peaks, and (ii) how it compared with the Sentinelles surveillance system. Table 1 summarizes the indicators of its performance regarding ILI peak detection and anticipation in each region, averaged over the two years of the model test. The sensitivity per episode (probability of detection of the ILI outbreak) had a mean of 0.92 (range 0.5-1) across regions, while the specificity per episode had an average of 0.58 (range 0.2-1). The sick leave alert generally occured prior to peak ILI incidence, on average 5.88 weeks (range 2.5-11) before the peak.

Table 1: Performance of the sick-leave model to detect ILI outbreaks: sensitivity per episode (probability of detection), specificity per episode and detection time before ILI peak. For each region, the value of these indicators are averaged over the two years evaluated.

Region	Sensitivity per episode	Specificity per episode	Detection time (weeks) before ILI

			peak
Auvergne-Rhône-Alpes	0.5	0.2	7
Bourgogne-Franche-Comté	0.5	0.2	11
Bretagne	1	1	2.5
Centre-Val de Loire	1	0.5	6.5
Grand Est	1	0.5	5
Hauts-de-France	1	0.5	7
Ile-de-France	1	1	3
Normandie	1	0.33	6
Nouvelle-Aquitaine	1	0.5	6.5
Occitanie	1	1	3
Pays de la Loire	1	0.5	6.5
Provence-Alpes-Côte d'Azur	1	0.66	6.5
Total mean	0.92	0.58	5.88

Table 2 compares the performance of our sick leave model with the Sentinelles surveillance system. We observe a mean sensitivity of 0.89 (range 0.64-1) and a mean specificity of 0.86 (range 0.80-0.95). When both alerts match, the sick leave model alert was always triggered earlier than the Sentinelles model, with an average lead time of 2.5 weeks (range 0.5-4).

Table 2: Performance of the sick-leave model compared to the Sentinelles *model: sensitivity, specificity and detection time before ILI peak. For each region, the value of these indicators are averaged over the two years evaluated.*

Region	Sensitivity	Specificity	Detection time (weeks) before ILI peak
Auvergne-Rhône-Alpes	0.87	0.81	1.5
Bourgogne-Franche-Comté	0.64	0.80	3
Bretagne	1	0.89	3
Centre-Val de Loire	1	0.83	4

Grand Est	0.91	0.84	4
Hauts-de-France	0.83	0.84	3
Ile-de-France	0.82	0.95	1
Normandie	0.93	0.80	1.5
Nouvelle-Aquitaine	1	0.85	4
Occitanie	0.79	0.95	1.5
Pays de la Loire	0.89	0.87	0.5
Provence-Alpes-Côte d'Azur	1	0.87	3
Total mean	0.89	0.86	2.5

4. Discussion

Workplace absenteeism data can be used by public health surveillance systems to detect emerging infectious disease epidemics (15). Despite this, to this date, few health authorities worldwide use absenteeism data to inform outbreak surveillance. Here, we assessed the potential of workplace absenteeism data to monitor influenza and detect epidemics in France, using an adapted statistical method to analyze this data. We applied this method to a comprehensive national database of workplace absenteeism and validated it against the French national surveillance system based on sentinel GPs. Our results suggest that a system based on workplace absenteeism could be highly sensitive and detect influenza epidemics earlier than the current French surveillance system.

We found that the surveillance system we propose would be able to detect outbreaks 5.9 weeks before the peak and about 2.5 weeks before the Sentinelles system. This suggests that sick-leave data could be almost as timely as emergency visits data that has a timeliness of 3 weeks compared to ILI data systems (16). Our findings in this regard are in line with previously published studies in several contexts worldwide. In a French study from 1994, sick-leave data from a large company allowed to detect flu epidemics with up to 2 weeks of advance (9). In a more recent Belgian study, worker absenteeism data from the Belgian Medical Expertise and from the Belgian railway system was shown to start rising 2-3 weeks in advance and to peak 2 weeks in advance, as compared with ILI data from the Belgian sentinel GP surveillance system (6). In the UK, data on workplace absenteeism among employees of Transport for London peaked up to 2 weeks before the NHS ILI surveillance data(7); and monitoring workplace absence due to "cold", "cough" or "influenza" among the staff of a large hospital

organization was shown to allow the detection of flu epidemics with a significant advance of up to 9 weeks (17).

Very few of these previously published studies included an assessment of the sensitivity and specificity of an absenteeism-based surveillance system. However, in the French study from 1994, the sensitivity and specificity of surveillance based on sick-leave data from a large company were estimated at 74% and 67% for the identification of epidemic weeks and 67% and 94% for the detection of epidemics, with an 80% positive predictive value (9). The UK study based on hospital staff absenteeism also noted that the resulting system did not lead to more false positives than the NHS surveillance data in London.

The quality of the developed model depends strongly on the quality of the data collected within companies. Sick-leave data have the advantage of describing quasi-real individual behavior regarding sick-leave and presence at the workplace. These data are in fact used to enter employees' pay and are subsequently fulfilled by obligation in the computer system. Our data may not be representative of French population because it describes data from a health-insurer. For instance, it insures few construction companies because they have their own specific insurer. The data also do not include unemployed people. However, representativeness is not necessarily required to build up an outbreak detection system. In fact, outbreak detection aims to detect any unusual expected number of cases to generate signal alarms. Similarly, the GP Sentinelles network does not include all GPs but a small subset of the same practitioners over time.

Another downside of our data is related to the definition of the sick leave rate. The denominator of this rate is the number of workers and it includes workers that are on holiday. The sick leave rate therefore drops during school holidays and the systems do not detect any alerts during those holidays even if they are included in the statistical model as covariates. This is an issue if the epidemic occurs during the holidays and this is the case during the year 2016-2018 where the peak occurs during the Christmas break for two regions. The model could then be more sensitive if the denominator was the number of employees actually at work.

Furthermore, the model accuracy estimated by the false positive rate is strongly related to the definition of cases to detect. The time series of cases is based only on flu cases. However, other epidemics such as gastroenteritis can influence sick leaves data and could be considered for future works: the sick leaves outbreaks may correspond to other disease and may explain the poor specificity in some cases. The detection algorithm could also be improved if the estimation of expected cases could adjust for any potential past exogenous environmental factors such as a terrorist attack, strikes or unexpected bad weather episodes. The model accuracy is moreover strongly related to the method chosen for the algorithm. The Serfling method may actually not be the more accurate model and was chosen to be consistent with the Sentinelles method. Some other regression-based models are known to be more specific, like the Farrington algorithm (18,19).

Another limitation of our study is linked to the fact that we relied on data that were consolidated on an annual basis, and not in real-time. As such, the system is able to detect outbreak based

on the dates of sick-leaves, but only retrospectively. A proper integration of sick-leave data into a health surveillance system would thus require an effort to ensure sick-leave data are consolidated and made available in real time. Our results suggest that the resulting increased timeliness of a surveillance system including this data stream may justify this effort. Moreover, these data sometimes already exist in near-real time because of local legislation. For instance, French companies must declare sick leave within 5 days.

Many of the previously published studies assessing the potential of workplace absenteeism data for flu surveillance simply provided visual analyses comparing and correlating absenteeism data with ILI surveillance data (6,7). By contrast, in this work, we propose a statistical approach and algorithm to analyze the French workplace absenteeism data, and to raise alarms when outbreaks are detected. This allows us to both propose a complete surveillance system that could be used in practice provided the data is available, and to fully assess the performance of this surveillance system.

Acknowledgement

TD is supported by Association Nationale de la Recherche et de la Technologie and Malakoff Humanis.

JB is supported by the INCEPTION project (PIA/ANR-16-CONV-0005)

PAA is supported by INSERM-ANRS (France Recherche Nord & Sud Sida-HIV Hépatites), grant number ANRS-12377 B104

DS is supported by a Canadian Institutes of Health Research Doctoral Foreign Study Award (Funding Reference Number 164263) as well as the French government through its National Research Agency project SPHINX-17-CE36-0008-01.

Conflict of interest

TD and RL are both employees of Malakoff Humanis.

Authors' contribution

TD, JB, RL: analyzed and extracted the data. TD, JB, PAAT, WD, KJ, LT, MH: Wrote the first draft of the paper. TD, JB, PAAT, RA, OSD, AD, JPG, NN, HN, DRMS, WD, KJ, LT, MNH: designed and discussed statistical analysis and interpreted data. All authors: Read, reviewed and approved the final manuscript.

References

1. Groenewold MR, Konicki DL, Luckhaupt SE, Gomaa A, Koonin LM. Exploring National Surveillance for Health-Related Workplace Absenteeism: Lessons Learned From the 2009 Influenza A Pandemic. Disaster Medicine and Public Health Preparedness. 2013 Apr;7(2):160–6.

2. World Health Organization. WHO Fact sheets, Influenza (Seasonal) [Internet]. 2018 [cited 2020 May 18]. Available from: https://www.who.int/news-room/fact-sheets/detail/influenza-(seasonal)

3. German RR, Lee LM, Horan JM, Milstein RL, Pertowski CA, Waller MN, et al. Updated guidelines for evaluating public health surveillance systems: recommendations from the Guidelines Working Group. MMWR Recomm Rep. 2001 Jul 27;50(RR-13):1–35; quiz CE1-7.

4. Groenewold M, Burrer S, Ahmed F, Uzicanin A. National Surveillance for Health-Related Workplace Absenteeism, United States 2017-18. Online J Public Health Inform [Internet]. 2019 May 30 [cited 2020 May 18];11(1). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6606163/

5. Dailey L, Watkins RE, Plant AJ. Timeliness of Data Sources Used for Influenza Surveillance. J Am Med Inform Assoc. 2007;14(5):626–31.

6. Bollaerts K, Antoine J, Robesyn E, Van Proeyen L, Vomberg J, Feys E, et al. Timeliness of syndromic influenza surveillance through work and school absenteeism. Arch Public Health. 2010 Sep 29;68(3):115–20.

7. Paterson B, Caddis R, Durrheim D. Use of Workplace Absenteeism Surveillance Data for Outbreak Detection. Emerg Infect Dis. 2011 Oct;17(10):1963–4.

8. NIOSH - CDC. Absenteeism in the Workplace [Internet]. 2020 [cited 2020 May 18]. Available from: https://www.cdc.gov/niosh/topics/absences/default.html

9. Quenel P, Dab W, Hannoun C, Cohen JM. Sensitivity, Specificity and predictive Values of Health Service Based Indicators for the Surveillance of Influenza A Epidemics. Int J Epidemiol. 1994 Aug 1;23(4):849–55.

10. DARES. Les congés payés et jours de RTT : quel lien avec l’organisation du travail ? DARES Analyse [Internet]. 2017 Sep [cited 2020 May 18];55. Available from: https://dares.travail-emploi.gouv.fr/IMG/pdf/2017-054.pdf

11. Valleron AJ, Bouvet E, Garnerin P, Ménarès J, Heard I, Letrait S, et al. A computer network for the surveillance of communicable diseases: the French experiment. Am J Public Health. 1986 Nov;76(11):1289–92.

12. Serfling RE. Methods for current statistical analysis of excess pneumonia-influenza deaths. Public Health Rep. 1963 Jun;78(6):494–506.

13. Costagliola D, Flahault A, Galinec D, Garnerin P, Menares J, Valleron AJ. A routine tool for detection and assessment of epidemics of influenza-like syndromes in France. Am J Public Health. 1991 Jan;81(1):97–9.

14. Retel O, Fortin N, Henry V, Hubert B, Faisant M, Casamatta D, et al. Contribution des associations SOS Médecins à une surveillance locale de la grippe saisonnière en France. Bull épidémiol hebd. 2014;(28):466–72.

15. Buehler JW, Hopkins RS, Overhage JM, Sosin DM, Tong V, CDC Working Group. Framework for evaluating public health surveillance systems for early detection of outbreaks: recommendations from the CDC Working Group. MMWR Recomm Rep. 2004 May 7;53(RR-5):1–11.

16. Heffernan R, Mostashari F, Das D, Karpati A, Kulldorff M, Weiss D. Syndromic surveillance in public health practice, New York City. Emerging Infect Dis. 2004 May;10(5):858–64.

17. Drumright LN, Frost SDW, Elliot AJ, Catchpole M, Pebody RG, Atkins M, et al. Assessing the use of hospital staff influenza-like absence (ILA) for enhancing hospital preparedness and national surveillance. BMC Infect Dis. 2015 Mar 1;15:110.

18. Noufaily A, Enki DG, Farrington P, Garthwaite P, Andrews N, Charlett A. An improved algorithm for outbreak detection in multiple surveillance systems. Stat Med. 2013 Mar 30;32(7):1206–22.

19. Farrington CP, Andrews NJ, Beale AD, Catchpole MA. A Statistical Algorithm for the Early Detection of Outbreaks of Infectious Disease. Journal of the Royal Statistical Society Series A (Statistics in Society). 1996;159(3):547–63.

Supplementary material

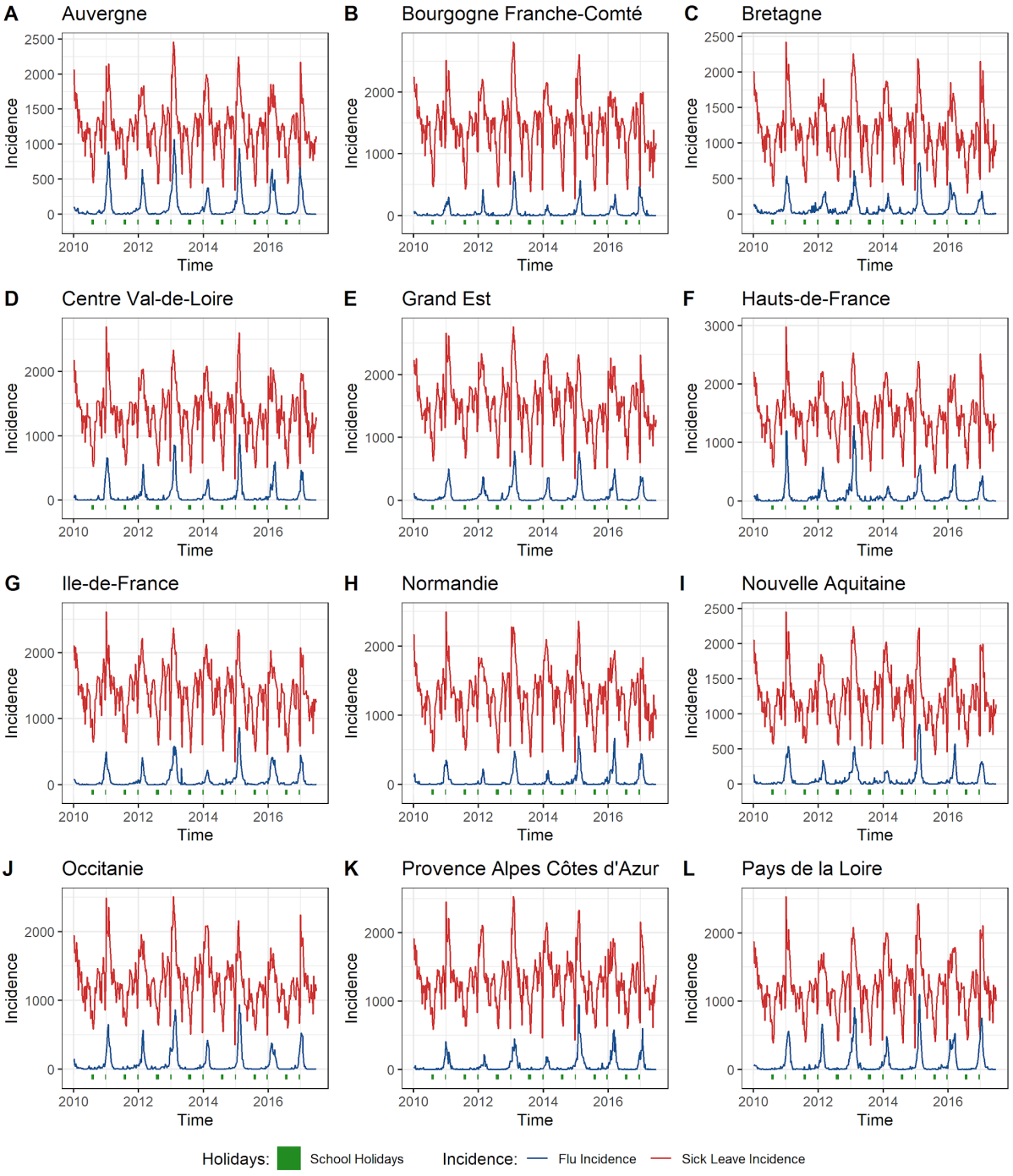

Figure S1: Incidence per 100,000 per week of influenza-like illness and sick leave in twelve French regions, 2010-2017. The Christmas and summer school holidays (increased worker leave periods) are shown at the bottom.

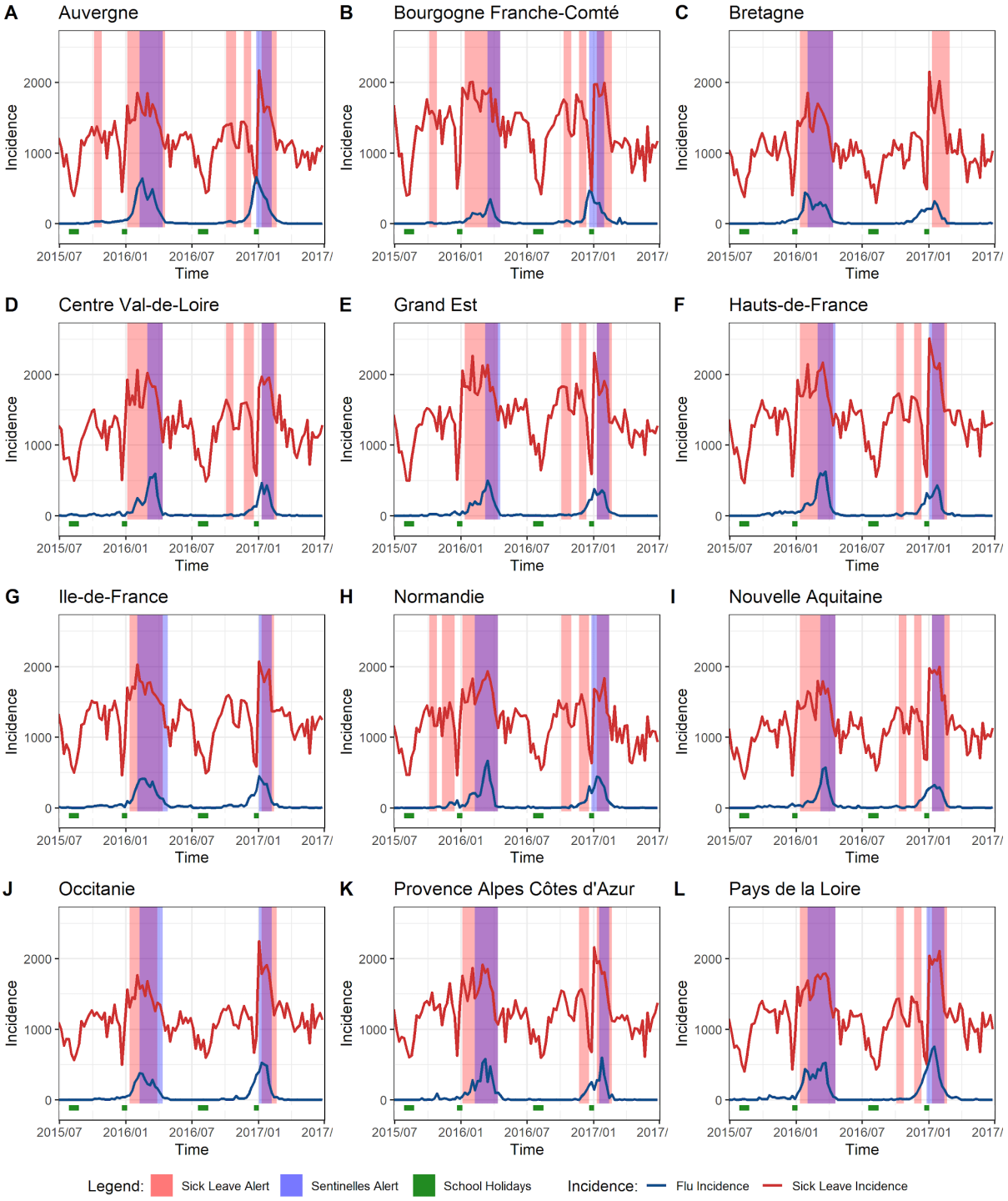

Figure S2: Incidence of influenza-like illness and sick leave, 2015-2017, and alerts from the Sentinelles *and the sick-leave models, in twelve French regions. The Christmas and summer school holidays (increased worker leave periods) are shown at the bottom.*

le cnam

Tom Duchemin

Modélisation et surveillance pour la prévention des arrêts maladie

HESAM UNIVERSITÉ

Résumé : Alors que les arrêts maladie sont le signe d'un mal-être croissant chez les salariés et qu'ils pèsent un coût certain pour la collectivité, la numérisation et le partage systématique des données offrent de belles opportunités pour leur prévention. Nous avons ainsi profité de cette opportunité pour développer un éventail d'outils de prévention basés sur des méthodes d'analyse statistique. Dans un premier temps, ces travaux de thèse proposent une analyse des mécanismes expliquant les arrêts maladie chez le salarié. L'analyse d'une enquête nationale a premièrement permis d'identifier et de hiérarchiser leurs principaux facteurs déterminants grâce à l'algorithme des forêts aléatoires. Ensuite, une analyse de données administratives a identifié des trajectoires d'absence pouvant mener à des arrêts graves grâce à des analyses séquentielles et à de la modélisation multi-état. Dans un second temps, des outils ont été développés afin d'identifier des situations anormales d'arrêt maladie à l'échelle de l'entreprise. Une typologie d'entreprise a premièrement été construite afin de produire des valeurs étalon pour que les entreprises évaluent précisément leur situation. Un algorithme de détection des pics d'absence, adapté de modèles de surveillance épidémiologique, a enfin été développé pour pouvoir identifier automatiquement les entreprises en dérive.

Mots clés : santé au travail, prévention, arrêt maladie, surveillance, forêt aléatoire, statistiques

Abstract : At a time when sick leave is a sign of growing ill-being for workers and a cost burden for the society, the systematic digitalization and distribution of data offers great opportunities for its prevention. We have therefore taken advantage of this opportunity to develop a range of prevention tools based on statistical analysis methods. In a first part, this work proposes an analysis of the mechanisms explaining sick leave among workers. The analysis of a national survey has first identified and prioritised their main determinants using random forest. Then, an analysis of administrative data had helped to identify absence trajectories that could lead to serious sick leaves thanks to sequential analyses and multi-state modelling. In a second step, tools were developed to identify abnormal situations of sick leave at company level. A company typology was first built to produce benchmark values for companies to accurately assess their situation. Finally, an algorithm for identifying absence peaks, adapted from epidemiological surveillance models, was finally developed to automatically identify companies in difficulty.

Keywords : occupational health, prevention, sick leave, surveillance, random forest, statistics