Abandons dans une enquête sur internet : l'exemple de l'inclusion dans la cohorte Coset-MSA

Noémie SOULLIER[1]
Santé publique France, Chargée de projet organisation et suivi des enquêtes

Hugo ROGIE[2]
Santé publique France, Stagiaire

Guilhem DESCHAMPS[3]
Santé publique France, Data scientist

Jean-Luc MARCHAND[4]
Santé publique France, Chef de projet COSET-Indépendants

Béatrice GEOFFROY-PEREZ[5]
Santé publique France, Cheffe de projet COSET-MSA

TITLE

Dropouts of Web surveys: lessons from the recruitment in the French Coset-MSA cohort study

RÉSUMÉ

Les enquêtes par internet sont de plus en plus utilisées y compris dans le domaine de la santé. Elles présentent l'avantage de pouvoir enquêter de nombreuses personnes à faible coût et de disposer de données enregistrées automatiquement qui peuvent éclairer sur les comportements de réponse. Nous nous proposons d'explorer les abandons en cours de questionnaire lors du recrutement de la cohorte française Coset-MSA, à la fois en fonction des profils des répondants mais également de leurs données de connexion. Le programme Coset a pour objectif de décrire et de surveiller l'état de santé de la population selon l'activité professionnelle en France. Entre fin 2017 et début 2018, 270 000 actifs âgés de 18 à 65 ans affiliés à la Mutualité Sociale Agricole (MSA) en 2016 ont été invités à remplir un questionnaire en ligne. À la fin de la collecte, 28 054 personnes avaient répondu au moins aux questions d'identité (sexe et âge) et parmi celles-ci, 2 004 personnes avaient abandonné le remplissage de leur questionnaire avant la fin. Le taux d'abandon était associé à l'âge, à l'état de santé perçu et au niveau d'études. L'abandon était également associé aux données de connexion : plus fréquent parmi les personnes qui s'étaient

1. noemie.soullier@santepubliquefrance.fr
2. hugo.rogie@hotmail.fr
3. guilhem.deschamps@santepubliquefrance.fr
4. Jean-Luc.MARCHAND@santepubliquefrance.fr
5. Beatrice.GEOFFROY-PEREZ@santepubliquefrance.fr

connectées une seule fois, celles qui ne s'étaient jamais connectées le week-end, celles qui s'étaient connectées pour la première fois la nuit et celles qui s'étaient connectées pour la première fois après la seconde relance. Ces paradonnées, disponibles à faible coût pour tous les individus, se révèlent pertinentes pour documenter les abandons lorsqu'on a peu de données auxiliaires. Ce travail offre un nouvel angle de vue sur l'analyse des paradonnées au regard des abandons, forme particulière de non-réponse partielle. Ces résultats pourront servir à orienter la correction de la non-réponse partielle, mais également à améliorer le questionnaire et le protocole de contact.

Mots-clés *: internet, enquête, santé, abandon, paradonnées.*

ABSTRACT

Internet surveys are developing, including in the health field. Internet makes it easy to survey many people at low cost and provides automatically recorded data, which can shed light on response behaviors. We intend to explore the questionnaire drop-outs during the recruitment of the French Coset-MSA cohort, both according to the respondents' characteristics but also according to their connection data. The Coset program aims to describe and follow the health status of the French population according to professional characteristics. Between late 2017 and early 2018, 270,000 working people aged 18 to 65 years and affiliated to the Mutualité Sociale Agricole (MSA) in 2016 were invited to complete an online questionnaire. At the end of the collection, 28,054 people had answered at least the identity questions (sex and age) and among these, 2,004 people had stopped their questionnaire before the end. The drop-out rate was associated with age, perceived health status and education. Drop-out was also associated with connection data: it was more frequent among those who had logged in only once, those who had never logged in on weekends, those who had logged in for the first time at night and those who had logged in for the first time after the second reminder. These paradata, available at low cost for all individuals, prove to be relevant for documenting questionnaire drop-outs when there is few auxiliary data. This work offers a new perspective on the analysis of paradata with regard to questionnaire drop-outs, a particular form of item non-response. These results could be used to guide the correction of item non-response, but also to improve the questionnaire and the contact protocol.

Keywords: *Web, survey, health, drop-out, paradata.*

1. Introduction

En lien avec l'augmentation de l'équipement et de l'utilisation d'internet ces dernières décennies (Legleye *et al.*, 2022), les enquêtes par internet sont de plus en plus utilisées y compris dans le domaine de la santé. Avec le développement de l'accès à internet, la couverture de ces enquêtes est désormais importante (Croutte & Muller, 2021), même si une part de la population reste éloignée d'internet en raison d'un accès lent ou par choix, soit en raison d'une crainte pour la protection de ses données, soit parce qu'elle tire peu de bénéfice d'internet (Felderer & Herzing, 2022). Les enquêtes par internet présentent par ailleurs l'avantage de pouvoir contacter de nombreuses personnes à faible coût (Couper, 2000).

La non-réponse aux enquêtes est classiquement divisée en non-réponse totale (la personne ne répond à aucune question) et non-réponse partielle (la personne ne répond pas à certaines questions), toutes deux étant des indicatrices de qualité d'une enquête (Groves, 1989). La collecte par internet souffre d'un taux de non-réponse totale plus important que pour les autres modes de collecte (Daikeler *et al.*, 2020), quand le taux de non-réponse partielle est quant à lui plutôt similaire entre les modes (Čehovin *et al.*, 2022). Ces deux types de non-réponses sont le résultat de deux mécanismes différents, puisque dans le premier cas la décision est prise avant de commencer le questionnaire, alors que dans le second cas elle est prise au cours du questionnaire. Cependant, il semble exister une relation positive entre les deux au niveau individuel : par exemple, les personnes les plus réticentes à participer à l'enquête sont également plus susceptibles de ne pas répondre à certaines questions (Couper, 1997). Ainsi, les personnes se placeraient sur un *continuum*, alliant leur propension à répondre à l'enquête et à chaque question de celle-ci (Yan & Curtin, 2010). Dans le cadre des enquêtes longitudinales, les individus qui ont émis une non-réponse partielle plus importante à une vague d'enquête sont moins susceptibles de répondre à la vague suivante (Loosveldt et al., 2002).

Tout comme la non-réponse totale, la non-réponse partielle peut être source de biais si les non-répondants diffèrent des répondants et que seuls les répondants sont pris en compte dans l'estimation. Des méthodes d'imputation prenant en compte les caractéristiques des individus peuvent permettre de résoudre ce problème, sous l'hypothèse que les données sont manquantes au hasard conditionnellement aux caractéristiques prises en compte (de Leeuw *et al.*, 2003). Les enquêtes par internet permettent, grâce à des données supplémentaires enregistrées automatiquement telles que les données de connexion (paradonnées), de définir plus finement les comportements de non-réponse (Bosnjak & Tuten, 2001), mais également d'en éclairer la compréhension (Kreuter, 2013).

Nous nous proposons d'explorer les abandons en cours de questionnaire lors du recrutement de la cohorte française Coset-MSA. Ces abandons seront étudiés à la fois en fonction des profils des répondants, mais également de leurs données de connexion. Ces résultats permettront d'identifier les facteurs associés à l'abandon en cours de questionnaire, afin d'orienter la correction de la non-réponse partielle mais également de fournir des éléments pouvant servir à la compréhension de la non-réponse totale aux vagues suivantes, voire à la mise en place de protocoles ciblés.

2. Matériel et méthodes

Le programme Coset (Cohortes pour la Surveillance Epidémiologique en lien avec le Travail) est un dispositif longitudinal pour la surveillance épidémiologique en lien avec le travail (Geoffroy-Perez *et al.*, 2012), s'appuyant sur des données de cohortes concernant les actifs français. Il a pour objectif de surveiller l'état de santé de la population selon l'activité professionnelle en France. Dans ce cadre, deux cohortes ont été constituées par Santé publique France : Coset-Indépendants et Coset-MSA, dont les populations cibles sont les actifs âgés de 18 à 65 ans qui étaient affiliés respectivement au Régime Social des Indépendants et à la Mutualité Sociale Agricole (MSA) en 2016.

Ce travail porte sur le recrutement de la cohorte Coset-MSA par un questionnaire en ligne. Le recrutement de cette cohorte visait à inclure 30 000 cohortistes. Pour cela, 270 000 personnes ont été tirées au sort dans les bases de la MSA et invitées à participer. Les invitations, envoyées en plusieurs vagues, ont commencé en novembre 2017, sous la forme d'un courrier postal envoyé au domicile des personnes. Le courrier comprenait un identifiant et un mot de passe uniques, qui permettaient aux personnes de s'authentifier afin de répondre au questionnaire en ligne. Les personnes n'ayant pas répondu recevaient une première relance postale, puis une deuxième si elles n'avaient toujours pas répondu. Les dernières relances postales ont été envoyées fin mai 2018. L'accès au questionnaire en ligne a été ouvert le 20 novembre 2017 et fermé le 11 juillet 2018. À cette date, 28 054 personnes avaient rempli les questions obligatoires situées sur la première page du questionnaire (sexe et âge) et leurs réponses étaient cohérentes avec l'identité de la personne tirée au sort et invitée.

Le questionnaire se composait des grandes sections suivantes :

- Santé,
- Habitudes et cadre de vie,
- Activité professionnelle actuelle,
- Historique professionnel (autres activités et arrêts de travail).

Pour la partie concernant l'historique professionnel, il était demandé à la personne de renseigner et décrire chaque épisode professionnel (type d'activité, profession, période d'exercice, conditions de travail) et chaque arrêt de travail qu'elle a eu au cours de sa carrière professionnelle. Pour les personnes n'ayant pas eu d'autre emploi que leur emploi actuel ou

n'ayant jamais eu d'arrêt de travail au cours de leur carrière, ces parties n'étaient pas à remplir. L'étude de cette section montre qu'elle a été vraisemblablement mal remplie de manière générale, avec une sous-déclaration du nombre d'épisodes professionnels (1,5 en moyenne contre 2,4 déclarés par questionnaire papier lors de l'étude pilote menée sur un échantillon de 2 363 répondants).

Ainsi, afin de pouvoir bien distinguer les abandons en cours de questionnaire, on s'est intéressé dans cette analyse uniquement à la partie linéaire du questionnaire, c'est-à-dire sans prendre en compte la partie sur l'historique professionnel (voir tableau 1). Sur cette partie contenant 22 pages, le nombre minimal de questions posées à une personne était de 85 et le nombre maximal 267. Il est à noter que les pages 13 à 22 concernaient uniquement les personnes ayant déclaré être en activité au moment du remplissage du questionnaire (n = 26 847, 96%) et les pages 20 à 22 uniquement celles ayant déclaré une activité professionnelle actuelle qui n'était pas une activité de bureau (n = 20 285, 72%).

Dans ce travail, on considère qu'une personne a abandonné définitivement le remplissage de son questionnaire à la page p si elle n'a répondu à aucune des questions qui lui étaient posées à partir de la page p et jusqu'à la page 22.

Les variables étudiées pour l'association avec l'abandon étaient :

- des variables de l'auto-questionnaire rempli par la personne : sexe, âge, état de santé perçu, taille et poids (utilisés pour le calcul de l'indice de masse corporelle), les antécédents de cancer et le niveau d'études le plus élevé atteint par la personne ;
- des variables de la base de sondage : statut professionnel (salarié / non-salarié) et zone géographique (métropole / départements et régions d'outre-mer) au moment du tirage au sort (au dernier trimestre 2017) ;
- des variables de connexion au questionnaire (paradonnées) : une variable couplant le nombre de connexions au délai entre les connexions (une seule connexion / plusieurs connexions le même jour / plusieurs connexions en plusieurs jours), une variable indiquant s'il y avait eu au moins une connexion le week-end, une variable indiquant le créneau horaire de la première connexion (matin 6h-12h / midi 12h-14h / après-midi 14h-18h / soirée 18h-22h / nuit 22h-6h), et une variable indiquant après quel courrier avait eu lieu la première connexion (courrier d'invitation / première relance / seconde relance).

Les différences de taux d'abandon pour une caractéristique donnée ont été testées au moyen du test du khi-deux de Pearson. Des régressions logistiques multivariées ont été réalisées pour analyser l'abandon en fonction de plusieurs variables explicatives. Les analyses ont été effectuées avec le logiciel SAS© version 9.4.

Tableau 1 – *Présentation des différentes pages de la partie linéaire du questionnaire*

Page	Description de la page	Nombre de questions sur la page (min-max)
Page 1 - **Santé générale 1/3**	Sexe, âge, biométrie, état de santé générale perçu	7
Page 2 - **Santé générale 2/3**	Antécédents cardiovasculaires, métaboliques, psychiques	7-16
Page 3 - **Santé générale 3/3**	Antécédents de cancer	1-7
Page 4 - **Santé respiratoire**	Volet santé respiratoire de l'enquête *European Community Respiratory Health Survey* (ECRHS), autres antécédents de maladie respiratoire, allergie nasale	15-32

Page 5 - **Santé musculo articulaire**	Symptômes musculo-articulaires au cours des douze derniers mois	11-32
Page 6 - **Santé - Moral**	Symptomatologie dépressive (échelle *Center for Epidemiologic Studies- Depression* (CES-D))	20
Page 7 - **Santé - Autres**	Maladies infectieuses au cours des 12 derniers mois, troubles de l'audition, eczéma de contact, autres problèmes de santé, recours aux soins, observance	13-15
Page 8 - **Cadre de vie**	Situation familiale et composition du foyer, situation professionnelle du conjoint	4-9
Page 9 - **Rythme de vie**	Chronobiologie, sommeil, rythme alimentaire	3
Page 10 - **Habitudes vie - Tabac**	Consommation de tabac actuelle et passée, utilisation de la cigarette électronique	1-11
Page 11 - **Habitudes vie - Alcool**	Consommation d'alcool : Audit C abrégé	1-4
Page 12 - **Activité professionnelle - Situation actuelle**	Niveau d'études, situation professionnelle actuelle	2-7
Page 13 - **Activité professionnelle - Description 1/2**	Volume horaire, polyactivité, statut et description de l'activité professionnelle actuelle principale (ou dernière activité professionnelle pour les inactifs)	0-18
Page 14 - **Activité professionnelle - Description 2/2**	Description des cultures produites et des animaux d'élevage concernés par cette activité	0-29
Page 15 - **Activité professionnelle - Déplacements**	Déplacements et horaires de travail (travail de nuit, rythme de travail, organisation du travail)	0-17
Page 16 - **Activité professionnelle - Bien-être**	Contact avec le public, déséquilibre efforts/ récompenses (*Effort-Reward Imbalance* (ERI))	0-18
Page 17 - **Activité professionnelle - Efforts 1**	Efforts et contraintes physiques au travail : pénibilité physique (échelle de Borg), contraintes musculosquelettiques	0-16
Page 18 - **Activité professionnelle - Efforts 2**	Contraintes musculosquelettiques : postures	0-8
Page 19 - **Activité professionnelle - Bruits**	Exposition aux bruits	0-3
Page 20 - **Activité professionnelle - Autres expositions 1**	Entretien de machines, de bâtiments, utilisation de solvants, matériaux d'isolation, de construction, de peinture, de soudage en rapport avec cet entretien	0-20
Page 21 - **Activité professionnelle - Autres expositions 2**	Désinfection de bâtiments ou de matériel, activités de brûlage, exposition aux poussières	0-10
Page 22 - **Activité professionnelle - Autres expositions 3**	Utilisation et/ou application de produits phytopharmaceutiques sur les cultures ou les animaux	0-20

3. Résultats

Au total, 2 004 personnes (7%) ont abandonné le remplissage avant la fin du questionnaire. Parmi ces abandons, 43% ont eu lieu dans la partie santé, 9% dans la partie habitudes et cadre de vie, et 48% dans la partie activité professionnelle actuelle.

Les abandons les plus fréquents ont été observés aux pages 5 (Santé musculaire et articulaire)

(*n* = 163, 8% des abandons), 6 (Santé-Moral) (*n* = 349, 17% des abandons), 13 (Description de l'activité professionnelle 1/2) (*n* = 380, 19% des abandons) et 14 (Description de l'activité professionnelle 2/2) (*n* = 199, 10% des abandons). Ainsi, les pages 5-6 de la partie santé totalisaient 25% des abandons et les pages 13-14 de la partie professionnelle totalisaient 29% des abandons. Les pages 5 et 6 comprenaient en moyenne une vingtaine de questions (respectivement 22 et 20) ; les pages 13 et 14 étaient les premières pages s'intéressant en détail à l'activité professionnelle actuelle et comprenaient en moyenne respectivement 14 et 8 questions.

L'âge et le sexe, renseignés sur la première page du questionnaire, sont des données nécessaires à la définition de la population d'étude (vérification de l'identité du participant). Par conséquent, il n'y a pas de données manquantes pour ces variables, ni d'abandon à la page 1. En revanche, pour les autres caractéristiques étudiées, des données manquantes existent : l'analyse se fait parmi les valeurs renseignées et exclut donc une partie des abandons (ceux qui ont eu lieu avant la page sur laquelle la caractéristique est renseignée et ceux qui ont eu lieu après mais pour lesquelles la réponse à la question est manquante).

Le taux d'abandon était identique selon le sexe (Tableau 2). Cependant, les hommes abandonnaient significativement plus tôt dans le questionnaire que les femmes : parmi les hommes qui ont abandonné, 47% se sont arrêtés à la partie « santé », 10% à la partie « habitudes » et 43% à la partie « activité », contre respectivement 37%, 8% et 55% des femmes qui abandonnent.

Le taux d'abandon diminuait avec l'âge, les 18-34 ans abandonnant 1,5 fois plus que les 50-65 ans.

Le taux d'abandon diminuait lorsque l'état de santé perçu était meilleur : ce taux est 1,5 fois moins important chez les personnes déclarant un bon état de santé perçu par rapport à celles se déclarant en mauvais état de santé.

On n'observe pas d'association significative avec l'indice de masse corporelle, ni avec les antécédents de cancer.

Tableau 2 – *Taux d'abandon selon les variables de l'auto-questionnaire*

	Effectif	% d'abandon[1]	Test de khi-deux (*p-value*[2])
Total	**28 054**	**7,1%**	
Sexe [page 1]		**7,1%**	***NS***
Homme	15 990	7,1%	
Femme	12 064	7,2%	
Age [page 1]		**7,1%**	**< 0,001**
18-34 ans	4 695	8,9%	
35-49 ans	10 197	7,0%	
50-65 ans	13 162	6,1%	
Etat de santé perçu (manquant=109[3]) [page 1]		**7,0%**	**< 0,001**
Bon (A-B)	13 756	6,5%	
Moyen (C-D-E)	12 753	7,3%	
Mauvais (F-G-H)	1 436	9,9%	

Indice de masse corporel (manquant=531[4]) [page 1]		**6,8%**	***NS***
Maigre (<18,5) 500	500	8,0%	
Normal (18,5-<25)	13 952	7,0%	
Surpoids (25-<30)	9 359	6,5%	
Obèse (>=30)	3 712	6,7%	
Antécédents de cancer (manquant=160[5]) [page 3]		**6,8%**	***NS***
Non	26 776	6,8%	
Oui	1 118	5,6%	
Niveau d'études (manquant=1227[6]) [page 12]		**3,3%**	**< 0,001**
Jamais scolarisé, Ecole primaire ou Collège	987	7,1%	
Enseignement technique court (CAP, BEP ou équivalent)	5 866	4,3%	
Lycée	4 042	3,3%	
1er cycle de l'enseignement supérieur (Bac + 1 à Bac + 3)	11 563	2,9%	
2e et 3e cycle de l'enseignement supérieur (Bac + 4 et plus)	4 369	2,4%	

1 *calculé parmi les valeurs renseignées*
2 *niveau de significativité du test statistique ; NS = non-significatif (p-value > 0,05)*
3 *dont 43 abandons*
4 *dont 135 abandons*
5 *dont 117 abandons (75 avant la page 3, 24 à la page 3 et 18 après la page 3)*
6 *dont 1 110 abandons (1 033 avant la page 12, 50 à la page 12 et 27 après la page 12)*

Enfin, concernant le niveau d'études, cette information était collectée tardivement dans le questionnaire (page 12) ; l'analyse porte donc uniquement sur les abandons tardifs, c'est-à-dire ceux qui ont eu lieu dans la section « activité professionnelle » du questionnaire (soit 48% des abandons). Parmi ces abandons tardifs, le taux d'abandon était fortement lié au niveau d'études, avec un gradient net : plus le niveau d'études déclaré était faible et plus le taux d'abandon était élevé.

Aussi, afin d'étudier tous les abandons dans un modèle multivarié, il est préférable d'utiliser des informations renseignées pour l'ensemble des répondants. Ces informations étaient de deux ordres : les informations de la base de sondage et les données de connexion.

Les données de connexion sont des marqueurs des conditions de remplissage du questionnaire ; ces conditions pouvant jouer sur la propension à abandonner, elles étaient donc intéressantes à étudier. Par ailleurs, les données de connexion étaient intéressantes car elles étaient associées aux caractéristiques des personnes. Par exemple, les personnes qui déclaraient un mauvais état de santé perçu effectuaient plus souvent plusieurs connexions en plusieurs jours, moins souvent leur première connexion le soir ou la nuit et plus souvent leur première connexion à la suite du premier courrier (invitation). Le niveau d'études était également associé aux données de connexion : les personnes déclarant un niveau d'études supérieur ou égal à Bac + 4 effectuaient plus souvent une seule connexion, plus souvent leur première connexion le matin et plus souvent cette connexion suite au premier courrier (invitation). Ainsi, les données de connexion représentent une alternative pour tenir compte des caractéristiques disponibles uniquement via l'auto-questionnaire mais qui sont entachées de données manquantes.

Les variables de connexion étaient toutes associées à l'abandon en univarié (Tableau 3). Les abandons étaient plus fréquents parmi les personnes qui se connectaient une seule fois, celles

qui ne se connectaient jamais le week-end, celles qui se connectaient pour la première fois la nuit et celles qui se connectaient pour la première fois après la seconde relance.

Tableau 3 – *Taux d'abandon selon les variables de connexions*

	Effectif	% d'abandon	Test de khi-deux (*p-value*)
Nombre de connexions			< 0,001
1 seule connexion	22 471	7,8%	
Plusieurs connexions le même jour	2 655	3,9%	
Plusieurs connexions en plusieurs jours	2 928	5,1%	
Au moins une connexion le week-end			< 0,05
Non	20 441	7,3%	
Oui	7 613	6,7%	
Créneau horaire de la première connexion			< 0,05
Matin (6h-12h)	6 972	6,7%	
Midi (12h-14h)	2 838	7,8%	
Après-midi (14h-18h)	8 754	7,2%	
Soirée (18h-22h)	8 076	6,9%	
Nuit (22h-6h)	1 414	9,1%	
Première connexion après le courrier			< 0,001
1 (invitation)	10 750	6,6%	
2 (relance 1)	12 080	7,2%	
3 (relance 2)	5 224	8,2%	

Concernant les données de la base de sondage, les abandons étaient plus fréquents parmi les non-salariés (7,6% *vs* 6,9% parmi les salariés, $p < 0,05$) et dans les départements et régions d'outre-mer (18,7% *vs* 7,0% en métropole, $p < 0,0001$).

Le modèle logistique multivarié incluait le sexe et l'âge issus de l'auto-questionnaire, le statut professionnel et la zone géographique issus de la base de sondage, et les données de connexion. Une sélection pas-à-pas (*stepwise*) a été appliquée. Toutes choses égales par ailleurs, l'abandon était associé à l'âge, à la zone géographique, au nombre de connexions et au courrier après lequel avait lieu la première connexion (Tableau 4).

Tableau 4 – *Résultats de la régression logistique multivariée expliquant l'abandon*

	Effectif	OR [IC 95%]	*p-value*
Sexe (auto-questionnaire)			NS
Âge (auto-questionnaire)			< 0,0001
18-34	4 695	1,37 [1,22 - 1,55]	
35-49	10 197	1,06 [0,96 - 1,18]	
50-65	13 162	1 Ref	
Statut professionnel (base de sondage)			NS

Zone géographique (base de sondage)			< 0,0001
Métropole	27 610	1 Ref	
Outre-mer	444	3,15 [2,46 - 4,02]	
Nombre de connexions			< 0,0001
1 seule connexion	22 471	1,55 [1,30 - 1,84]	
Plusieurs connexions le même jour	2 655	0,75 [0,58 - 0,97]	
Plusieurs connexions en plusieurs jours	2 928	1 Ref	
Au moins une connexion le week-end			NS
Créneau horaire de la première connexion			NS
Première connexion après le courrier			0,02
1 (invitation)	10 750	1 Ref	
2 (relance 1)	12 080	1,06 [0,96 - 1,18]	
3 (relance 2)	5 224	1,20 [1,06 - 1,36]	

Abréviations : OR = Odds-ratio, IC = Intervalle de confiance, Ref = Référence, NS = non-significatif (p-value > 0,05)

Lorsque l'on ajoutait la variable état de santé perçu au modèle (suppression de 109 individus dont 43 abandons), les mêmes variables restaient associées à l'abandon et la variable état de santé était également associée significativement ($p < 0,0001$) avec un risque d'abandon plus élevé pour les personnes déclarant un mauvais état de santé perçu (OR = 1,71 [1,42 - 2,07] pour un mauvais état de santé *vs* un bon état, et OR = 1,20 [1,09 - 1,32] pour un état de santé moyen *vs* un bon état).

4. Discussion

Ce travail propose une analyse des abandons en cours de questionnaire, dans le cadre d'un questionnaire en ligne destiné à l'inclusion dans une cohorte épidémiologique ciblant les risques professionnels. Cette analyse des abandons a pour intérêt d'éclairer la non-réponse partielle au questionnaire via une seule variable d'intérêt (l'abandon), permettant de tirer des conclusions globales pour l'ensemble du questionnaire. Ce travail offre un nouvel angle de vue en analysant également les paradonnées (ici les données de connexion) au regard des abandons en cours de questionnaire.

L'objectif du questionnaire Coset est de décrire l'état de santé en relation avec la situation professionnelle : ce sont là les deux grandes dimensions d'intérêt. Ainsi, les associations entre l'abandon et l'état de santé ou le niveau d'études interpellent sur le traitement des questions qui sont positionnées en fin de questionnaire : pour ces questions pour lesquelles il y aura eu de nombreux abandons antérieurs, il sera important de traiter rigoureusement la non-réponse partielle au risque que les résultats soient biaisés, par exemple si les personnes en meilleur état de santé remplissent plus souvent la partie professionnelle. Ces résultats sont également à prendre en compte pour la correction de la non-réponse totale à l'enquête. Les données de la cohorte Coset ont été appariées aux données de l'assurance maladie (Système National des Données de Santé, de la Caisse Nationale d'Assurance Maladie), pour les répondants et un échantillon aléatoire des non-répondants, lorsque ces personnes avaient été informées (courrier distribué) et qu'elles ne s'étaient pas opposées à l'appariement (Geoffroy-Perez *et al.*, 2019). Ces données appariées seront utilisées pour étudier les évènements de santé rencontrés par les cohortistes, l'appariement permettant d'éviter un recueil par questionnaire qui aurait été difficile et souffrant généralement de biais de mémoire. Cet appariement a également été mis en place afin de disposer de données de remboursements de soins, utilisées comme proxys

de l'état de santé de la personne, pour la correction de la non-réponse totale. Lors du pilote de l'enquête Coset-MSA, il avait en effet pu être observé que la consultation d'un praticien de ville (médecin généraliste, dentiste ou spécialiste) dans l'année précédant l'enquête était positivement associée à la probabilité de répondre à l'enquête, quand l'hospitalisation de la personne l'année précédente était associée négativement à la probabilité de répondre (Soullier *et al.*, 2018). Les résultats observés ici sur les abandons en cours de questionnaire confortent ce choix d'utiliser cet appariement dans la correction de la non-réponse totale, en particulier compte tenu du faible taux de participation à l'enquête.

L'utilisation de paradonnées telles que les données de connexion d'une enquête par internet est intéressante car ces données sont peu coûteuses et ne souffrent pas de non-réponse partielle (Lynn & Nicolaas, 2010). En effet, ces données sont collectées de manière indirecte lors d'une enquête et ne pèsent donc pas sur le fardeau de réponse du répondant. De plus, elles sont renseignées pour tous les répondants, y compris ceux qui abandonnent le questionnaire en cours de remplissage. Lors du recrutement en ligne de la cohorte Coset-MSA, le nombre de connexions et le courrier après lequel avait lieu la première connexion étaient significativement associés à l'abandon, toutes choses égales par ailleurs. Les paradonnées peuvent donc jouer un rôle dans la correction de la non-réponse partielle, au même titre que les données de la base de sondage et que certaines données déclarées dans l'auto-questionnaire qui ne souffriraient pas de non-réponse partielle. En effet, il semble tout de même important d'inclure des données déclarées essentielles qui ne sauraient être totalement remplacées par les paradonnées, par exemple ici l'état de santé perçu qui contient peu de données manquantes.

L'intérêt d'étudier les abandons est aussi qu'on peut le faire en fonction de variables dont on ne dispose pas forcément pour les non-répondants, que ce soit les variables déclarées dans l'auto-questionnaire ou les paradonnées. Or, on pourrait voir la non-réponse totale comme un abandon qui a lieu avant le début du questionnaire. Ainsi, les facteurs associés à l'abandon pourraient aussi éclairer sur la non-réponse totale, observée à cette vague ou attendue aux vagues suivantes. Par exemple, l'association de l'abandon avec l'âge est identique aux associations connues avec la non-réponse totale : les plus jeunes abandonnent plus souvent, de même qu'ils répondent moins à l'enquête. Il pourrait être intéressant d'orienter des stratégies de contact ou de connexion afin de limiter les abandons et de regarder si ces stratégies permettent également de réduire la non-réponse totale. Coset étant une cohorte, ces adaptations de protocole pourraient servir à limiter l'attrition.

La non-réponse partielle, et donc l'abandon, est un indicateur de la qualité du questionnaire. Aussi le fait que les abandons augmentent avec les relances donne à réfléchir sur le compromis à faire entre non-réponse totale et non-réponse partielle : les relances permettent de diminuer la non-réponse totale mais on peut obtenir des questionnaires moins bien complétés. Aussi, il convient de relancer avec parcimonie le cas échéant.

Il convient également de réfléchir à des adaptations spécifiques des questionnaires internet pour certains territoires, où la fracture numérique se fait sentir. Par exemple, le taux d'abandon est bien plus élevé dans les territoires ultra-marins, alors même que le taux de réponse y est également bien plus faible. Une analyse s'appuyant sur un zonage permettant d'appréhender les territoires urbains et ruraux pourrait apporter un éclairage important sur ce sujet. Cette catégorisation, non disponible au moment de cette analyse, pourra être construite à partir des adresses postales de contact des cohortistes. Pour ces territoires ou pour les groupes de population moins enclins à répondre sur internet, un protocole multimode pourrait également permettre de réduire la non-réponse totale et partielle (de Leeuw, 2005). Compte tenu du nombre important de personnes invitées et du budget contraint de l'enquête, cette possibilité n'a pas été envisagée pour le recrutement de la cohorte, mais pourrait l'être pour les vagues suivantes.

Le remplissage d'un questionnaire résulte d'un compromis entre l'intérêt porté par le répondant et le fardeau engendré par le questionnaire (Galesic, 2006). Un abandon peut donc être interprété comme un fardeau ressenti plus important que l'intérêt porté. Ainsi, même si on peut penser que répondre à un questionnaire en ligne est plus aisé pour les plus jeunes, ils abandonnent plus fréquemment en cours de questionnaire, soit parce que leur vie professionnelle est débutante et que le sujet les intéresse moins, soit parce que le questionnaire leur apparaît trop long et que subjectivement le fardeau de réponse leur paraît plus important. Il en est de même pour l'état de santé : l'état de santé perçu au moment du remplissage est très lié à l'abandon ce qui peut témoigner d'un fardeau de réponse ressenti plus lourd pour les personnes qui se déclarent en moins bon état de santé, même si elles sont intéressées pour évoquer les sujets de santé. On retrouve aussi cela dans les paradonnées : les personnes qui se connectent plusieurs fois abandonnent moins, ce qui peut être interprété comme un intérêt à répondre qui prévaut sur la difficulté à répondre. Une manière d'approcher le fardeau rencontré par le répondant est la durée (Yan & Tourangeau, 2008 ; Zhang & Conrad, 2014 ; Six *et al.*, 2016) ; cette donnée n'est malheureusement pas disponible de manière fiable et pour tous les répondants dans les paradonnées de Coset. Les déconnexions n'étaient en effet pas enregistrées avec certitude, et le temps de remplissage calculé en cas de connexions multiples est donc peu fiable. À titre d'illustration, la durée totale médiane de remplissage du questionnaire calculée parmi les personnes ayant rempli le questionnaire en une seule connexion était de 37 minutes. Le questionnaire était donc d'une durée relativement longue pour une enquête par internet (Revilla & Ochoa, 2017). On remarque par ailleurs que quatre pages du questionnaire concentrent plus de la moitié des abandons. Une analyse plus approfondie pourrait étudier le moment de l'abandon (page) en fonction du nombre de questions (variable dépendant de la page) et de la difficulté des questions (de binaire oui/non la plus simple à question ouverte la plus difficile) sur cette page mais également sur les précédentes. Cette analyse pourra permettre d'envisager des améliorations tenant à l'ergonomie du questionnaire, ce afin de réduire les abandons dans les enquêtes futures.

Dans la continuité de ce travail, on pourra également étudier l'impact de la prise en compte des paradonnées sur la correction de la non-réponse partielle. Cette analyse pourra être complétée en incluant d'autres données auxiliaires (non disponibles au moment de ce travail) appariées aux données d'enquête, telles que les données de l'assurance maladie ou les données professionnelles issues des bases de données de la MSA. L'analyse pourra également être pondérée par les poids de sondage corrigés pour la non-réponse totale (non disponibles au moment de ce travail) (Santin *et al.*, 2014).

En conclusion, l'étude des abandons permet de détecter des difficultés à répondre à l'enquête, liées aux caractéristiques du répondant (âge, état de santé, niveau d'études) mais aussi aux données de connexion (nombre de connexions, créneau horaire). Cette étude permet d'orienter le traitement de la non-réponse partielle et également de la non-réponse totale, mais aussi de proposer des pistes d'amélioration du questionnaire et du protocole de contact pour les vagues futures de la cohorte.

Références

Bosnjak M. and T. Tuten (2001), « Classifying Response Behaviors in Web-Based Surveys », *Journal of Computer-Mediated Communication*, 6(3), https://doi.org/10.1111/j.1083-6101.2001.tb00124.x

Čehovin G., M. Bosnjak, and K. Lozar Manfreda (2022), « Item Nonresponse in Web Versus Other Survey Modes: A Systematic Review and Meta-Analysis », *Social Science Computer Review*, 0(0), https://doi.org/10.1177/08944393211056229

Couper M. P. (1997), « Survey Introductions and Data Quality », *Public opinion quarterly*, 61(2), pp. 317-338.

Couper M. P. (2000), « Web Surveys: A Review of Issues and Approaches », *Public opinion quarterly*, 64, pp. 464-494.

Croutte P. and J. Muller (2021), « Baromètre du numérique », Crédoc.

Daikeler J., M. Bošnjak, and K. Lozar Manfreda (2020), « Web Versus Other Survey Modes: An Updated and Extended Meta-Analysis Comparing Response Rates », *Journal of Survey Statistics and Methodology*, 8(3), pp. 513-539.

de Leeuw E., J. Hox, and M. Huisman (2003), « Prevention and treatment of item nonresponse », *Journal of Official Statistics*, 19, pp. 153-176.

de Leeuw E. D. (2005), « To mix or not to mix data collection modes in surveys », *Journal of Official Statistics*, 21, pp. 233-255.

Felderer B. and J. M. E. Herzing (2022), « What about the Less IT Literate? A Comparison of Different Postal Recruitment Strategies to an Online Panel of the General Population », *Field Methods*, 0(0), https://doi.org/10.1177/1525822X221132940

Galesic M. (2006), « Dropouts on the web: Effects of interest and burden experienced during an online survey », *Journal of Official Statistics*, 22(2), pp. 313-328.

Geoffroy-Perez B., J. Chatelot, G. Santin, L. Benezet, P. Delezire, and E. Imbernon (2012), « Coset : un nouvel outil généraliste pour la surveillance épidémiologique des risques professionnels. Numéro thématique. Surveillance épidémiologique des risques professionnels, quoi de neuf ? », *Bulletin Epidémiologique Hebdomadaire*, 22-23, pp. 276-277.

Geoffroy-Perez B., N. Soullier, P. Delézire, L. Bénézet, G. Deschamps, E. Breuillard, J. Chesneau, and J.-L. Marchand (2019), « Cohortes pour la surveillance épidémiologique en lien avec le travail (Coset). Bilan de la phase d'inclusion de la cohorte Coset-MSA », Santé publique France, Saint-Maurice, 71 p.

Groves R. M. (1989), *Survey Errors and Survey Costs*, New York, Wiley.

Kreuter F. (2013), *Improving Surveys with Paradata: Analytic Uses of Process Information*, Hoboken, New Jersey, John Wiley & Sons, Inc.

Legleye S., A. Nougaret, and L. Viard-Guillot (2022), « L'usage des technologies de l'information et de la communication par les ménages entre 2009 et 2021 », *Insee Focus*, 259, https://www.insee.fr/fr/statistiques/6049348

Loosveldt G., J. Pickery, and J. Billiet (2002), « Item Nonresponse as a Predictor of Unit Nonresponse in a Panel Survey », *Journal of Official Statistics*, 18(4), pp. 545-557.

Lynn P. and G. Nicolaas (2010), « Making Good Use of Survey Paradata », *Survey Practice*, 3, pp. 1-5.

Revilla M. and C. Ochoa (2017), « Ideal and Maximum Length for a Web Survey », *International Journal of Market Research*, 59(5), pp. 557-565.

Santin G., B. Geoffroy, L. Benezet, P. Delezire, J. Chatelot, R. Sitta, J. Bouyer, and A. Gueguen (2014), « In an occupational health surveillance study, auxiliary data from administrative health and occupational databases effectively corrected for nonresponse », *J. Clin. Epidemiol.*, 67(6), pp. 722-730.

Six M., A. Kowarik, and M. Plate (2016), « Using Paradata to Assess the Quality of an Online Questionnaire », Statistics Austria.

Soullier N., B. Geoffroy-Perez, A. Gueguen, L. Bénézet, J. Chatelot, M. Zins, and G. Santin (2018), « Correction de la non-réponse et estimation de prévalences : résultats issus de trois cohortes épidémiologiques ciblant les risques professionnels », « 10ème Colloque francophone sur les sondages » (Lyon, France).

Yan T. and R. Curtin (2010), « The Relation Between Unit Nonresponse and Item Nonresponse: A Response Continuum Perspective », *International Journal of Public Opinion Research*, 22(4), pp. 535-551.

Yan T. and R. Tourangeau (2008), « Fast Times and Easy Questions: The Effects of Age, Experience and Question Complexity on Web Survey Response Times », *Applied Cognitive Psychology*, 22(1), pp. 51-68.

Zhang C. and F. Conrad (2014), « Speeding in Web Surveys: The tendency to answer very fast and its association with straightlining », *Survey Research Methods*, 8(2), pp. 127–135, https://doi.org/10.18148/srm/2014.v8i2.