Plan d'échantillonnage et calculs de pondérations

ALICE GUÉGUEN, RÉMI SITTA

28 MAI 2012

SOMMAIRE

1 INTRODUCTION

Constances a pour objectif d'être à la fois un outil de recherche et de santé publique. La population qui participe à la cohorte doit donc répondre aux contraintes/critères suivantes :

- *La population échantillonnée doit être « représentative » de la population cible (adultes âgés de 18 à 69 ans affiliés au Régime général de Sécurité sociale des départements inclus dans Constances), afin de pouvoir produire des estimations utiles pour la santé publique et ne pas souffrir de biais de sélection lors d'analyses étiologiques. La notion de représentativité est définie ci-dessous.*
- *La population cible doit pour cela pouvoir être prise en compte exhaustivement, et toute personne de cette population doit être en situation de pouvoir suivre tout le processus impliqué par Constances (visite d'inclusion, suivi actif et passif,..), et donc entre autres être éligible à un examen de santé.*
- *La population cible doit pouvoir permettre l'extrapolation des résultats descriptifs à la France entière (métropole) pour les assurés du Régime Général au sens large.*

De plus, l'échantillon constitué par la cohorte Constances doit avoir une structure proportionnelle à celle de la population cible pour les critères d'âge, de sexe et de catégorie sociale (PCS).

Résumé du circuit : L'inclusion des participants est prévue sur une durée de 5 ans. Des assurés éligibles sont sélectionnés dans une base exhaustive, puis invités par courrier à venir dans un CES donné pour participer à Constances et/ou effectuer un EPS. Les assurés participants et une partie des autres assurés invités seront suivis de manière passive dans les bases de données administratives.

Rappels sur la notion de représentativité et la prise en compte du volontariat

Lorsque l'on construit un échantillon à partir d'une population donnée, et que l'on veut s'assurer de ce qui est convenu d'appeler sa représentativité, la théorie impose : a) l'existence d'un plan de sondage probabiliste ; et b) d'estimateurs associés à ce plan (i.e. qui tiennent compte du plan de sondage). Un plan de sondage probabiliste peut se définir essentiellement par l'existence et la connaissance pour chaque individu de la population cible de sa probabilité d'inclusion dans l'échantillon. Ces probabilités doivent impérativement être strictement positives : dans le cas contraire, il est assez intuitif de voir que les échantillons constitués par de tels plans ne pourront représenter que la partie de la population formée par les individus à probabilités d'inclusion non nulles. La connaissance des probabilités d'inclusion de tous les individus d'un échantillon assure de pouvoir construire des estimations « représentatives » pour la population cible.

Cependant, ce schéma théorique va être perturbé dès lors que l'on se retrouve face à de la non-réponse de certains des individus sélectionnés par le plan de sondage. Cette non-réponse est quasi inéluctable dans n'importe quelle enquête nécessitant d'une part de contacter des personnes échantillonnées et d'autre part d'obtenir de leur part un consentement de participation : la non réponse étant ainsi la résultante d'un échec de contact ou d'un refus de volontariat. On est alors obligé de considérer la non-réponse comme une phase d'échantillonnage supplémentaire, dont les probabilités individuelles sont cette fois-ci inconnues. La seule façon de procéder consiste à estimer au mieux ces probabilités, tout en sachant qu'il persistera une erreur possible. Une hypothèse majeure reste cependant que ces probabilités de volontariat ne sont jamais nulles.

Les probabilités utilisées au final pour donner des estimations pertinentes seront donc le produit des probabilités d'inclusion connues et des probabilités de réponse estimées.

2 POPULATION CIBLE

Constances se greffe sur l'activité des Centres d'Examens de Santé (CES). Ceux-ci proposent des examens périodiques de santé (EPS) aux bénéficiaires affiliés au Régime général de la Sécurité sociale au sens large, c'est-à-dire le Régime Général au sens strict ainsi que les Sections Locales Mutualistes (SLM). Les 17 CES qui participent à Constances dans 16 départements sont des CES volontaires, et non tirés au sort parmi la centaine des CES présents sur le territoire français. La population ciblée par Constances est donc constituée de personnes qui de par leur affiliation ont droit à un EPS dans un de ces CES Constances, i.e. qui sont soit affiliées à l'une des CPAM concernées (Tableau 1), soit affiliées à l'une des SLM participantes (Tableau 2) et résidant dans un des départements Constances. De plus, nous nous sommes restreints aux personnes âgées de 18 à 69 ans.

Cette population est évidemment dynamique (changement de département de résidence, de régime d'affiliation, limites d'âge atteintes et décès). Or, l'inclusion se faisant sur 5 ans avec 5 tirages au sort, il est naturel de définir une population cible qui concerne cette période en entier, tout en choisissant une définition très précise. Nous avons donc décidé de fixer la population cible à l'union des 5 populations cible de chaque vague, et que celles-ci seront définies par les assurés qui respectent les critères d'inclusion à la date du tirage au sort de la vague donnée. Ainsi, pour la vague 1, la population cible est l'ensemble des assurés qui sont d'après leur statut d'affiliation et de résidence éligibles à Constances au jour du tirage (8/11/2010), et dont l'âge calculé au 31/12/2010 est de 18 à 69 ans.

Tableau 1. Centres d'examens de Santé

Département	CES	CPAM concernées
16	Angoulême	161
33	Bordeaux	331
59	Lille	594, 595, 624
69	Lyon	691
13	Marseille	131
54	Nancy	542
30	Nîmes	301
45	Orléans	451
75	Paris (CPAM+IPC)	751
64	Pau	642
86	Poitiers	861
35	Rennes	351
22	Saint-Brieuc	221
44	Saint-Nazaire	441
31	Toulouse	311
37	Tours (La Riche)	371

Tableau 2. Sections Locales Mutualistes participant à Constances

SLM	
MGEN	Mutuelle générale de l'Éducation nationale
MG	Mutuelle générale
MFP	Mutualité fonction publique
MNH	Mutuelle nationale des hospitaliers
LMDE	La Mutuelle des étudiants
MNT	Mutuelle nationale territoriale
CAMIEG	Caisse d'assurance maladie des industries électrique et gazière

3 BASE DE SONDAGE

Le tirage au sort est effectué par la Caisse nationale d'assurance vieillesse (Cnav). Celle-ci construit la base de sondage à partir du RNIAM (Répertoire national inter régimes des bénéficiaires de l'assurance maladie), qui contient de manière exhaustive la liste des assurés sociaux. Cette procédure assure une erreur de couverture quasi nulle entre la population cible et la base de sondage : en effet, aux erreurs de mises à jour de la base près, la base de sondage correspond exactement à la population cible à la date du tirage au sort.

Cependant, les informations contenues dans le RNIAM ne permettent d'identifier le CES d'affectation que pour les assurés du Régime général au sens strict : en effet pour ces derniers le code CPAM indique de quel CES l'assuré dépend, alors que pour les assurés des SLM cela n'est pas le cas, car le code CPAM a des sens variables selon la SLM, voire est unique et donc seulement national[1]. Ainsi, la base de sondage comprend initialement des sujets en surnombre (ceux affiliés à une SLM et ne résidant pas dans un département Constances) et qui sont ôtés dans un deuxième temps par le Tiers de Confiance (TC) quand on connait leur département de résidence.

Il y aura au total 5 tirages au sort correspondant aux 5 vagues d'inclusion. Pour éviter toute interférence d'une vague sur l'autre, les assurés sociaux ont donc été subdivisés à l'avance en 5 groupes, grâce à la clé de leur Numéro d'inscription au répertoire (NIR) qui est unique et stable dans le temps, et qui comprend des valeurs entre 1 et 97. Le tirage au sort de la première vague a ainsi été effectué avec les personnes ayant les clés entre 1 et 19. Par la suite, les bases de sondage concerneront donc les personnes ayant les clés entre 20 et 38 pour l'année 2, puis 39 à 57 pour l'année 3, 58-76 pour l'année 4 et enfin 77-95 pour l'année 5. Le pilote a été effectué avec les clés 96 et 97.

Pour la première vague d'inclusion, la base de sondage contient les assurés du RNIAM, dont la clé NIR est comprise entre 1 et 19, dont l'année de naissance est comprise entre 1941 et 1992 inclus, et qui au 08/11/2010 sont, soit affiliés au régime général au sens strict et appartiennent à une des CPAM du tableau 1, soit affiliés à l'une des 7 SLM du tableau 2.

4 TIRAGE AU SORT

4.1 OBJECTIFS

Le tirage au sort consiste à attribuer à chaque assuré de la base de sondage une *probabilité d'inclusion* (probabilité d'être tiré au sort) de sorte que l'échantillon obtenu contienne le nombre adéquat d'assurés permettant d'atteindre le nombre de participants visé. D'autre part, en plus du NIR et de l'information RG strict ou SLM, le RNIAM contient la date de naissance et le sexe des assurés. De plus, la Cnav dispose de l'accès à l'historique de carrière des bénéficiaires du RG strict, donc à leur PCS. Le tirage au sort utilise de manière optimale ces informations disponibles dans la base de sondage de manière à ce que l'échantillon obtenu contienne le nombre adéquat d'assurés permettant d'atteindre le nombre de participants visé dans chaque strate d'âge, de sexe et de PCS.

4.2 PROCÉDURE

4.2.1 Nombre de participants potentiels à atteindre

Le nombre de participants à atteindre a été défini pour chaque CES en prenant le cinquième de leur activité (en volume d'examens périodiques de santé (EPS) réalisés). Pour la vague 1 (démarrage de Constances) le tirage au sort a bien respecté ces nombres ; cependant, en accord avec la CNAMTS, le nombre de personnes à invitation a été diminué de moitié dans un premier temps afin d'assurer une montée en charge progressive des inclusions (tableau 3).

[1] Ce problème a également une incidence sur la manière dont le tirage au sort est effectué.

Le tirage au sort s'effectue différemment pour le RG strict et pour les SLM. Pour les assurés du RG strict, le tirage au sort est fait avec une stratification sur les 16 départements, i.e. de manière indépendante département par département, avec des taux de sondage différents (en fonction de l'activité des CES). Pour les assurés des SLM, le tirage au sort est fait au niveau national, sans distinction de département (puisque cette information n'est pas disponible dans la base de sondage) avec une stratification par SLM. Il faut donc auparavant répartir le nombre de participants potentiels à atteindre entre RG strict et SLM. La part nationale globale des effectifs des SLM dans le RG est de 18,9%[2], mais on ne connaît pas leur répartition selon les départements Constances. On diminue donc le nombre de participants prévus de chaque CES, au prorata de ce ratio, de la part « réservée » aux SLM. La somme de ces parts donne le nombre de participants potentiels à utiliser lors du tirage au sort pour les SLM, et chaque CES conserve une part de 81.1% pour les assurés du RG strict.

Cette procédure implique que le nombre de tirés au sort des SLM n'est pas fixe, mais aléatoire. Elle implique de plus que la répartition géographique entre les assurés des SLM devrait être bien respectée. Or, si elle n'est pas respectée pour le RG strict, c'est justement parce que les CES ont des volumes d'activité (rapportés à leur population éligible) variables. On s'attend toujours à ce que chaque CES atteigne 81.1% de son nombre de participants avec des assurés du RG strict. Mais le nombre d'invités issus des SLM peut éventuellement être variable d'un CES à l'autre. Si le ratio SLM/RG strict est réellement de 18.9% pour chaque CES, il n'y aura que des faibles fluctuations dues à l'aléa du tirage au sort. Si au contraire ce ratio devait s'avérer relativement variable, alors il est possible que certains CES aient plus de demandes de participation qu'ils ne peuvent accepter, alors que d'autres CES en aient moins qu'attendues. Pour que les CES puissent à la fois atteindre le nombre de participants définis et répondre à toutes les demandes, il faudra que lors des tirages au sort ultérieurs, l'on s'efforce de respecter les ratios SLM/RG strict CES par CES, qui seront estimés grâce aux données de la vague 1. De même, il faudra ultérieurement estimer la répartition entre les différentes SLM pour les seuls départements Constances, et non plus au niveau national. En appliquant alors cette répartition lors de la stratification, on assurera au mieux CES par CES le respect de cette répartition entre SLM.

4.2.2 Utilisation des informations disponibles au moment du tirage au sort

Par ailleurs, les informations disponibles dans la base de sondage serviront à stratifier davantage le tirage au sort : plutôt que de sélectionner un nombre donné d'assurés dans toute la base de sondage, celle-ci sera découpée en sous-populations dans lesquelles on sélectionnera un nombre prédéfini d'assurés par sondage aléatoire simple. Ceci réduit la variabilité de l'échantillonnage et permet d'améliorer les estimations produites. D'autre part, un des objectifs visés est d'obtenir une allocation proportionnelle du nombre de participants entre ces strates. Comme on s'attend à des taux de participation relativement variables entre ces strates, les probabilités d'inclusion seront ajustées en fonction des taux attendus de participation. La probabilité de sélection d'un assuré sera ainsi modulée proportionnellement à sa probabilité (estimée *a priori*) de participer[3]. Ceci permettra d'espérer obtenir un échantillon de participants dont les effectifs par strates seront proportionnels à ceux de la base de sondage.

La stratification sera obtenue en croisant les informations disponibles sur le sexe, la classe d'âge (18-24, 25-29, 30-39, 40-49, 50-59, 60-65) et la typologie d'activité professionnelle (TAP). Celle-ci est issue du croisement entre le statut activité et la PCS enregistrés dans les bases de la Cnav (SNGI et SNGC) pour le compte du Régime général (RG). Dans ces bases, le dernier enregistrement en cours permet de connaître le statut d'activité ainsi que la PCS déclarée par l'employeur. Celle-ci n'est pas forcément renseignée, auquel cas on utilisera la dernière PCS connue (si elle existe), de la même manière que pour les inactifs. Il se peut aussi pour certains assurés qu'il n'y ait pas d'enregistrement

[2] Source : RNIAM, juin 2009

[3] Ces probabilités *a priori* ont été d'estimées à partir de l'Enquête Décennale santé 2002-2003 et de l'enquête Prévalence des hépatites B et C en France en 2004.

du tout, comme par exemple pour des assurés ayant-droits qui n'ont jamais cotisé au Régime Général. La TAP comporte donc 10 classes issues du croisement entre le statut actif/inactif et la PCS qui vaut 3, 4, 5 ou 6, en plus des classes « Actif - PCS inconnue » et « Aucun enregistrement disponible ». Elle n'est de plus disponible que pour les assurés du RG strict, et les assurés des SLM ne seront stratifiés que sur les critères d'âge et de sexe.

Il y aura donc au total 2 (sexe) x 6 (âge) x 10 (TAP) x 16 (départements) = 1 920 strates pour le RG strict et 2 (sexe) x 6 (âge) x 7 (SLM) =84 strates pour les SLM.

Tableau 3. Nombre de participants à atteindre par Centres d'examens de Santé pour la vague 1

Département	CES	Nombre de volontaires attendus
16	Angoulême	900
33	Bordeaux	1100
59	Lille	1400
69	Lyon	700
13	Marseille	1000
54	Nancy	1500
30	Nîmes	600
45	Orléans	1000
75	Paris (CPAM+IPC)	6000
64	Pau	1100
86	Poitiers	800
35	Rennes	1200
22	Saint-Brieuc	1200
44	Saint-Nazaire	850
31	Toulouse	1100
37	Tours (La Riche)	1200
	Total	**21650**

4.3 NOTATIONS NÉCESSAIRES AU CALCUL DES PROBABILITÉS D'INCLUSION

On définit ici les notations pour les quantités qui interviendront dans le calcul des probabilités d'inclusion : nombres d'assurés de la base de sondage, nombres de participants à atteindre, probabilités *a priori* de participation et probabilité *a priori* d'être dans un département Constances pour un assuré SLM. Ce calcul est fait lors du tirage au sort, car il utilise aussi bien des quantités prédéfinies que des quantités calculées directement sur la base de sondage.

4.3.1 Nombre d'assurés dans la base de sondage

a est le nombre d'affiliés au RG strict.

a_d est le nombre d'affiliés au RG strict du département d .

b est le nombre d'affiliés SLM.

b_s est le nombre d'affiliés dans la SLM s.

4.3.2 Nombre d'assurés par strate dans la base de sondage

Pour le RG strict, on a 120 strates par département, résultant du croisement des variables sexe, âge (6 classes) et TAP (10 codes) ; pour les SLM, les strates sont au nombre de 10 par SLM et résultent du croisement des variables sexe et âge seulement.

a_{dk} est le nombre d'affiliés au RG strict de la strate k du département. $\sum_k a_{dk} = a_d$

b_{sk} est le nombre d'affiliés de la strate k dans la SLM s. $\sum_k b_{sk} = b_s$

4.3.3 *Nombre de participants potentiels à atteindre*

v est le nombre total de participants à atteindre (RG strict + SLM) pour la vague 1 de Constances. La valeur de v est le total apparaissant dans le tableau 3.

v_d est le nombre de participants à atteindre (RG strict + SLM) pour le département d. Les valeurs de v_d sont donnés le tableau 3. $\sum_d v_d = v$.

w_d est le nombre de participants du RG strict à atteindre dans le département d. On l'a estimé par la part réservée au RG strict pour le département d :

$$w_d = 0.811\, v_d.$$

w_{dk} est le nombre de participants du RG strict de la strate k à atteindre dans le département d. Pour que l'objectif de proportionnalité des strates soit respecté, il est égal à :

$$w_{dk} = w_d \frac{a_{dk}}{a_d}.$$

w_s est le nombre de participants à atteindre pour la SLM s. On l'a estimé par le produit de la part réservée aux SLM ($0.189\ v$) et de la proportion d'assurés à la SLM s parmi le total des assurés SLM ($\frac{b_s}{b}$), à défaut de pouvoir appliquer la proportion réellement désirée (mais inconnue) d'assurés SLM éligibles à Constances.:

$$w_s = 0.189\ v\ \frac{b_s}{b}.$$

w_{sk} est le nombre de participants de la strate k à atteindre pour la SLM s. Pour que l'objectif de proportionnalité des strates soit respecté (au mieux), il devra être égal à :

$$w_{sk} = w_s \frac{b_{sk}}{b_s}.$$

4.3.4 *Probabilités a priori de participation*

Pour la vague 1, ces probabilités de participation inconnues ont été estimées à partir de l'Enquête Décennale santé 2002-2003 et de l'enquête Prévalence des hépatites B et C en France en 2004. On a fait l'hypothèse que les probabilités de participation par âge, sexe et TAP étaient indépendantes du CES ou de la SLM d'affiliation. Elles sont données dans le tableau 4. Pour les vagues ultérieures, ces probabilités seront réestimées avec les données des vagues précédentes.

p_{dk} est la probabilité de participer à Constances pour les assurés au RG strict de la strate k du département d .

p_{sk} est la probabilité de participer à Constances pour les assurés de la strate k de la SLM s.

4.3.5 *Probabilité a priori d'être dans un département Constances pour un assuré SLM*

Pour les assurés des SLM, il faut anticiper la restriction aux seuls départements Constances. Cette restriction a été estimée pour la vague 1 à un taux de 27 %, grâce au ratio observé pour le RG dans le RNIAM pour les tranches d'âges éligibles[4] . Pour les vagues ultérieures, ce ratio sera lui aussi évalué et éventuellement rectifié pour chaque strate grâce aux données des vagues précédentes.

r_{sk} est la probabilité pour les assurés de la strate k de la SLM s d'être dans un département Constances.

[4] Données de 2006.

4.4 CALCUL DES PROBABILITÉS D'INCLUSION

4.4.1 Pour un assuré du RG strict

t_{dk} est la probabilité pour un assuré de de la strate k du département d d'etre inclus dans l'échantillon tiré au sort. Il s'agit ici de déterminer t_{dk} de manière à bien atteindre w_{dk}.

La relation $w_{dk} = a_{dk} t_{dk} p_{dk}$ exprime le fait que pour un département et une strate donnés, le nombre de participants à atteindre (w_{dk}) sera égal au nombre de participants attendus qui est donné par le produit du nombre d'assurés (a_{dk}) par la probabilité d'être tiré au sort(t_{dk}) et la probabilité de participer (p_{dk}). En remplaçant w_{dk} par $w_d \frac{a_{dk}}{a_d}$, puis w_d par $0.811\ v_d$ on obtient :

$$t_{dk} = 0.811 \frac{v_d}{a_d} \frac{1}{p_{dk}}.$$

Tous ces paramètres sont connus lors du tirage au sort, et on peut donc en déduire exactement t_{dk}.

Tableau 4 Probabilités *a priori* de participation (en %)

			Régime strict		SLM	
Age	**Code TAP**	**Activité et dernière CS**	**Hommes**	**Femmes**	**Hommes**	**Femmes**
18-24 ans	0	Actif CS NR	8	10,5		
	1	Actif CS3	8,5	8,7		
	2	Actif CS4	8,8	10		
	3	Actif CS5	8,9	11,2		
	4	Actif CS6	7,4	9,1	7,1	9,1
	5	Inactif CS3	9	9		
	6	Inactif CS4	9,3	9,9		
	7	Inactif CS5	9,5	11,6		
	8	Inactif CS6	7,9	8,8		
	9	Aucune info	6,8	7		
25-29 ans	0	Actif CS NR	8,2	8,6		
	1	Actif CS3	8,4	8,9		
	2	Actif CS4	9	10		
	3	Actif CS5	6,9	7		
	4	Actif CS6	8,2	9,3	7,5	8,2
	5	Inactif CS3	9,1	9,3		
	6	Inactif CS4	9,5	10		
	7	Inactif CS5	8,7	8,1		
	8	Inactif CS6	8,5	8,8		
	9	Aucune info	8	7,6		
30-39 ans	0	Actif CS NR	9,3	9,3		
	1	Actif CS3	10,2	9,5		
	2	Actif CS4	10,2	9,9		
	3	Actif CS5	8,7	8,9		
	4	Actif CS6	8,3	8,8	9,1	8,6
	5	Inactif CS3	9,3	9,1		
	6	Inactif CS4	9,4	9,4		
	7	Inactif CS5	8,9	8,1		
	8	Inactif CS6	7,4	7,4		
	9	Aucune info	8,3	6,7		
40-49 ans	0	Actif CS NR	9,5	9,4		
	1	Actif CS3	12,7	12,2		
	2	Actif CS4	10,7	11,4		
	3	Actif CS5	7,4	7,5		
	4	Actif CS6	7,3	8,6	9,3	9,2
	5	Inactif CS3	10	10,1		
	6	Inactif CS4	9,6	10,5		
	7	Inactif CS5	8,6	8		
	8	Inactif CS6	7,4	7,3		
	9	Aucune info	8,2	8,8		
50-59 ans	0	Actif CS NR	10,5	9,3		
	1	Actif CS3	12	9,6		
	2	Actif CS4	11,8	10		
	3	Actif CS5	9,5	8,9		
	4	Actif CS6	8,4	8,3	11,1	9,7
	5	Inactif CS3	12	10		
	6	Inactif CS4	12,2	11,4		
	7	Inactif CS5	9,9	10,1		
	8	Inactif CS6	7,7	7,1		
	9	Aucune info	9,2	9,4		
60-69 ans	0	Actif CS NR	9,7	9,3		
	1	Actif CS3	10,1	9,3		
	2	Actif CS4	9,4	9,2		
	3	Actif CS5	9,2	9,3		
	4	Actif CS6	9,4	9,1	11,6	8,5
	5	Inactif CS3	12	9,1		
	6	Inactif CS4	12,1	9,3		
	7	Inactif CS5	9,3	8,1		
	8	Inactif CS6	8,4	7,6		
	9	Aucune info	9,2	8,1		

4.4.2 *Pour un assuré SLM*

t_{sk} est la probabilité pour un assuré de de la strate k de la SLM s d'etre inclus dans l'échantillon tiré au sort. Il s'agit de même ici de déterminer t_{sk} de manière à bien atteindre w_{sk}.

La relation $w_{sk} = b_{sk} t_{sk} r_{sk} p_{sk}$ exprime le fait que pour une SLM et une strate données, le nombre de participants à atteindre (w_{sk}) doit être égal au produit du nombre d'assurés (b_{sk}) par la probabilité d'être tiré au sort (t_{sk}), la probabilité d'appartenir à un département Constances (r_{sk}), et la probabilité de participer (p_{sk}). En remplaçant w_{sk} par $w_s \frac{b_{sk}}{b_s}$, puis w_s par $0.189\ v\ \frac{b_s}{b}$ on obtient :

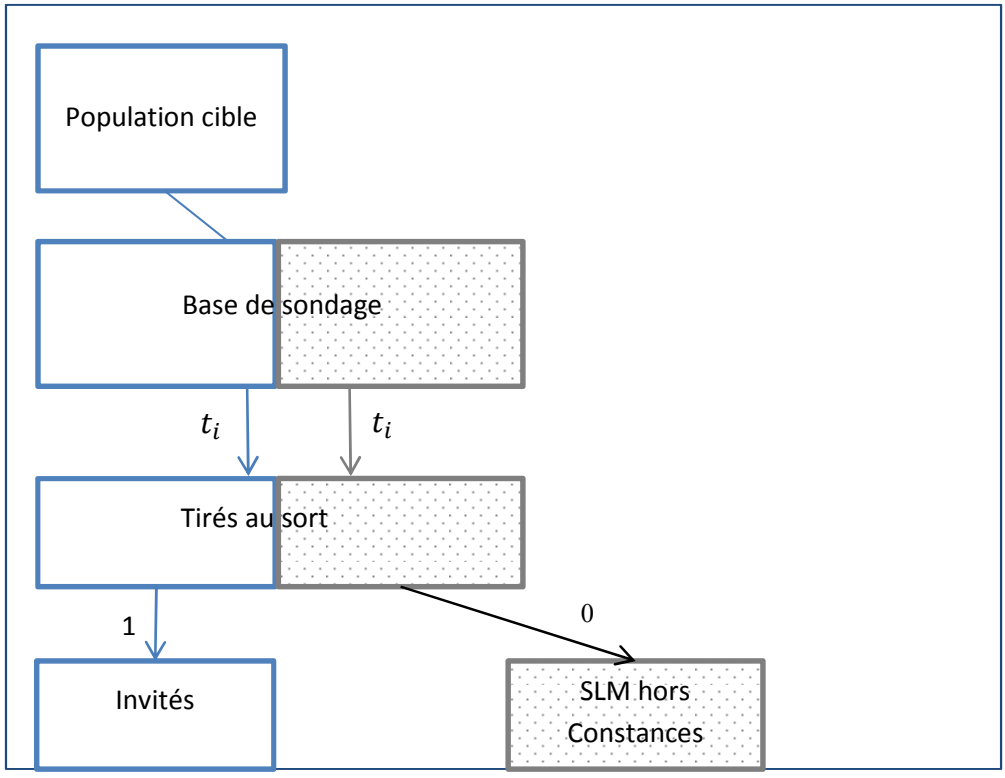
$$t_{sk} = 0.189\ \frac{v}{b} \frac{1}{r_{sk}} \frac{1}{p_{sk}}.$$

Ceci est le taux de sondage qu'il faudrait pouvoir appliquer afin d'atteindre w_{sk}. Or la valeur r_{sk} est inconnue et on impose donc à toutes les strates une valeur unique de 27 % (la valeur observée au niveau national).

5 DU TIRAGE AU SORT À L'INVITATION

La figure 1 résume les paragraphes précédents. La base de sondage contient des sujets en surnombre par rapport à la population cible : les personnes affiliées à une SLM Constances, mais ne résidant pas dans un département Constances. Le problème engendré par cette surcourverture de la base de sondage sera résolu après le tirage au sort : les sujets affiliés à une SLM et ne résidant pas dans un département Constances seront ôtés de la liste des tirés au sort par le Tiers de confiance et ne seront pas invités. Les pondérations des sujets invités seront égales à l'inverse des probabilités d'inclusion calculées au paragraphe précédent : $d_i = 1/t_i$.

Figure 1. De la population cible à l'invitation



6 DE L'INVITATION À LA PARTICIPATION

6.1 DESCRIPTION DES FLUX

La figure 2 représente le diagramme de flux de l'envoi de la lettre d'invitation à la participation effective à Constances. Les personnes tirées au sort (restreintes pour les SLM aux personnes effectivement éligibles au vu de leur adresse postale), sont invitées à participer à Constances (rectangle E). Ces invitations se font par courrier, par 10 vagues successives réparties sur une année. Dans la lettre d'invitation, les personnes sont informées de la mise en place de Constances, et on leur propose d'y participer. Si elles désirent participer à Constances, elles renvoient leur coupon-réponse au CES vers lequel elles sont orientées. Elles sont informées que même si elles ne participent pas, elles pourront être suivies de façon passive dans les bases administratives, à moins qu'elles ne signifient leur refus *via* le coupon-réponse.

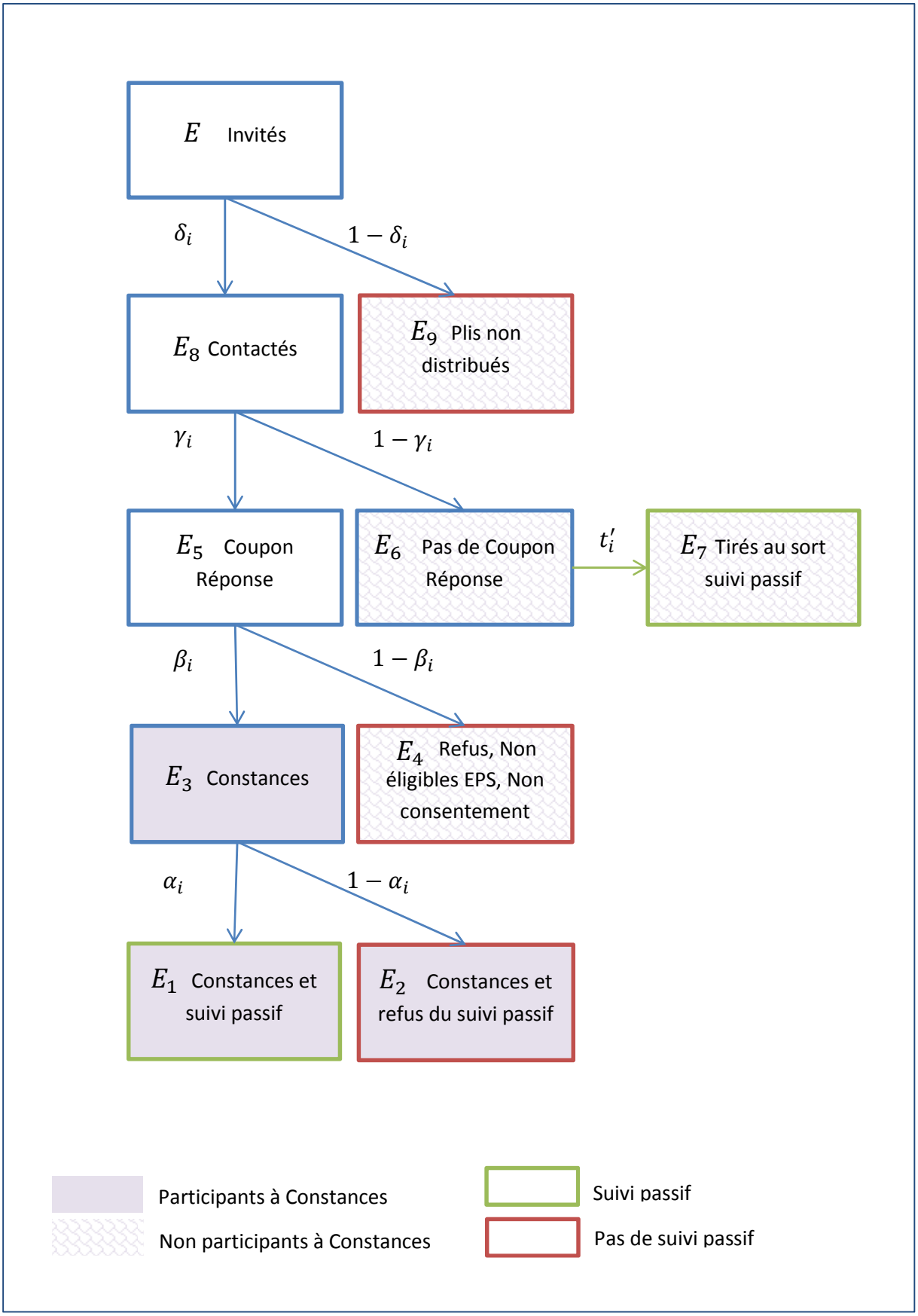
Parmi les personnes à qui une lettre d'invitation a été envoyée (rectangle E), il y aura des personnes à qui celle-ci n'aura pas été distribuée. Ces « plis non distribués » (rectangle E_9) seront renvoyés par la poste à l'Equipe Constances. Ces personnes sont non participantes. Du fait qu'elles n'ont pu être informées de Constances, il ne sera pas possible pour des raisons éthiques de les suivre dans les bases administratives.

Si les personnes contactées (rectangle E_8) ne renvoient pas de coupon-réponse (rectangle E_6), ce qui représente la plupart des invités, elles seront non participantes ; il sera possible de les suivre à travers les bases administratives. Mais pour des raisons de lourdeur informatique, il a été décidé de ne suivre qu'un échantillon de ces non participants, tiré au sort aléatoirement (rectangle E_7).

Parmi les personnes ayant renvoyé leur coupon réponse au CES (rectangle E_5), celles qui ne souhaitent pas participer à Constances et demandent à ne pas être suivies dans les bases administratives (une partie du rectangle E_4) ne seront pas participantes, et pour ces personnes il n'y aura pas de suivi à travers les bases administratives. De plus, parmi les personnes volontaires pour participer à Constances, certaines ne le pourront pas car elles ne seront plus éligibles à un EPS (changement d'affiliation ou de département de résidence), du fait du délai entre le tirage au sort et la réception de l'invitation (une autre partie du rectangle E_4). Enfin, certaines personnes ayant passé un EPS Constances ne donneront finalement pas leur consentement concernant la transmission des données de leur EPS à l'équipe Constances (une dernière partie du rectangle E_4). Toutes ces personnes (rectangle E_4) ne seront donc pas participantes. Pour des soucis de simplicité dans l'exposé des calculs de pondération pour non-participation, on considère ici qu'il n'y aura pas de suivi passif pour l'ensemble des personnes, même si certaines d'entre elles pourront l'avoir accepté.

Les autres personnes volontaires (rectangle E_3) seront les participants à Constances : elles ont la possibilité d'accepter le suivi passif (rectangle E_1) ou de le refuser (rectangle E_2).

Figure 2 : De l'envoi de la lettre d'invitation à la participation à Constances



6.2 CALCUL DES PONDÉRATIONS POUR NON-PARTICIPATION

6.2.1 Principe du calcul

A l'issue du tirage au sort, on obtient un échantillon E de taille n. Chaque sujet de E est affecté d'une pondération initiale connue d_i $i = 1, \dots, n$. L'échantillon E des invités est représentatif de la population cible si les sujets de E sont affectés de leur pondération initiale.

Les symboles accolés aux flèches de la figure 2 représentent des probabilités : α_i représente la probabilité que le sujet i de E_3 (participant à Constances) accepte le suivi passif et $1 - \alpha_i$ la probabilité qu'il le refuse ; β_i représente la probabilité qu'un sujet i de E_5 (ayant renvoyé son coupon-réponse) participe effectivement à Constances ; γ_i représente la probabilité qu'un sujet i de E_8 (contacté) renvoie son coupon-réponse ; δ_i représente la probabilité qu'un sujet i de E (invité) soit contacté.

La probabilité pour qu'un sujet i, invité de E, ait accepté Constances et le suivi passif est égale au produit de la probabilité qu'il ait été contacté (δ_i), de la probabilité qu'il ait renvoyé son coupon réponse (γ_i), de la probabilité qu'il ait accepté de participer à Constances (β_i) et enfin de la probabilité qu'il ait accepté le suivi passif (α_i). La pondération qui doit lui être attribuée pour que les sujets de E_1 soient représentatifs des invités Eest donc :

$$w_i = \frac{1}{\alpha_i}\frac{1}{\beta_i}\frac{1}{\gamma_i}\frac{1}{\delta_i},$$

pondération que l'on multiplie ensuite par d_i pour que les sujets de E_1 soient représentatifs de la population cible. Les probabilités α_i, β_i, γ_i et δ_i sont inconnues et doivent être estimées.

6.2.2 Principe de l'estimation

Soit une population A dont les sujets sont répartis par un mécanisme inconnu entre deux sous-populations A1 et A2, et π_i la probabilité que le sujet i soit dans A1. On cherche à estimer ce mécanisme inconnu par ces probabilités π_i. Pour cela, il est nécessaire de faire l'hypothèse qu'il existe des informations auxiliaires notées x_i, prédictrices de ce mécanisme, telles que celui-ci soit aléatoire conditionnellement à ces informations.

Lorsque ces informations sont disponibles sur toute la population A, on pourra directement estimer ces probabilités π_i, par exemple en posant un modèle f tel que $\pi_i = f(x_i, \theta)$, et en estimant θ par maximum de vraisemblance. On en déduira ainsi $\hat{\pi}_i = f(x_i, \hat{\theta})$.

Lorsque ces informations ne sont pas disponibles que sur toute la population A, mais qu'il existe un sous-échantillon B1 qui soit représentatif de A1 grâce à des pondérations *pond1*, et de même un sous-échantillon B2 qui soit représentatif de A2 grâce à des pondérations *pond2*, alors on pourra estimer de même $\hat{\pi}_i$ par $(x_i, \hat{\theta})$, où $\hat{\theta}$ sera estimé en utilisant cette fois-ci les pondérations *pond1* et *pond2* (par exemple via des équations estimantes).

6.2.3 Estimations des pondérations pour non-participation

Les informations auxiliaires disponibles sont issues de trois sources : la base de sondage qui permet de connaître x_{1i} , les bases administratives x_{2i} et les données enregistrées par Constances x_{3i}.

Les probabilités α_i seront estimées à partir des sujets de E_3, en comparant E_1 et E_2. On ne dispose des informations x_{2i} pour aucun des sujets de E_2, mais toutes les autres informations x_{1i} et x_{3i} sont disponibles pour la totalité de E_1 et E_2. On estimera donc α_i par $\hat{\alpha}_i = \hat{f}(x_{1i}, x_{3i})$ avec tous les sujets de E_1et E_2.

Les probabilités β_i seront estimées à partir des sujets de E_5 en comparant E_3 et E_4. Pour tous les sujets de E_4, on ne dispose que des informations de la base de sondage x_{1i}, qui sont toujours connues pour tous les sujets. On estimera donc β_i par $\hat{\beta}_i = \hat{f}(x_{1i})$ avec tous les sujets de E_3 et E_4.

Les probabilités γ_i seront estimées à partir des sujets de E_8 en comparant E_5 et E_6. Pour les sujets de E_6, on ne dispose pour tous les sujets que des informations de la base de sondage x_{1i}, mais on dispose pour les sujets du sous-échantillon E_7 des informations des bases administratives x_{2i}. De plus, on dispose d'une pondération $1/t_i'$ qui rend E_7 représentatif de E_6. Parallèlement, pour tous les sujets de E_5, on ne dispose que des informations de la base de sondage x_{1i}, mais on dispose pour les sujets du sous-échantillon E_1 de la totalité des informations x_{1i}, x_{2i}, x_{3i}. De plus, on utilisera la pondération $1/(\hat{\alpha}_i\hat{\beta}_i)$ pour rendre les sujets de E_1 représentatif de E_5. On estimera donc γ_i par $\hat{\gamma}_i = \hat{f}(x_{1i}, x_{2i})$ avec les sujets de E_1 et E_7 pondérés.

Les probabilités δ_i seront estimées à partir des sujets de E en comparant E_8. et E_9. Pour tous les sujets de E_9 on ne dispose que des informations de la base de sondage x_{1i}, qui sont aussi connues pour E_8. On estimera donc β_i par $\hat{\beta}_i = \hat{f}(x_{1i})$ avec tous les sujets de E_8 et E_9.

7 CALCUL DES PONDÉRATIONS POUR ATTRITION

Ce calcul sera similaire à celui pour non-participation. Il faudra cependant distinguer plusieurs situations d'attrition, en particulier l'attrition totale (perdus de vue), où les sujets demandent à ce qu'aucune information les concernant ne soit plus collectée, et l'attrition partielle, où les sujets arrêtent totalement de répondre au suivi actif, mais continuent à être suivi passivement dans les bases de données administratives. L'expérience de la cohorte Gazel monte que l'on peut s'attendre à ce que l'attrition totale soit relativement rare, et qu'on rencontre des situations où l'information pour un sujet (ou une partie de celle-ci) manque par intermittence : il est attendu que les questionnaires annuels puissent être manquants certaines années, puis retournés à nouveau les années suivantes[5].

On estimera donc, comme précédemment, pour les participants les probabilités de transition vers chacun de ces différents cas de figure. La quantité d'information disponible sera ainsi relativement variable selon les cas envisagés. Mais en tout état de cause, en plus des informations utilisées lors des calculs pour la participation, on disposera toujours des nombreuses données issues de l'inclusion.

8 CALAGE

Lorsque l'on connaît des paramètres pour la population cible (moyennes, totaux, proportions, etc..), le calage sur marges permet d'utiliser cette information pour modifier les pondérations de l'échantillon, de manière à ce que les estimations fournies avec les nouvelles pondérations donnent exactement les valeurs connues.

On pourra donc utiliser cette méthode de deux manières. D'une part, si l'on connaît des valeurs pour la population cible, on pourra modifier les pondérations pour améliorer encore les estimations. Ceci est particulièrement intéressant pour corriger d'éventuelles erreurs dues aux cas pour lesquels nous n'avions que peu d'informations (plis non distribués, refus explicite de suivi passif, invités devenus

[5] Goldberg et al. Health problems were the strongest predictors of attrition during follow up of the GAZEL cohort. J Clin Epid. 2006;59;1213-1221.

non éligibles ultérieurement). D'autre part, il serait possible de vouloir extrapoler des analyses sur une population plus large que la population cible, comme par exemple passer des assurés des départements Constances à ceux de toute la France métropolitaine.

Plusieurs sources d'informations peuvent ici être envisagées. La plus évidente est l'EGB (Echantillon Généraliste de Bénéficiaires) qui collecte prospectivement, pour un large échantillon aléatoire de tous les assurés du Régime Général, quasiment la même information que le suivi passif de Constances. On peut donc l'envisager comme source à la fois pour améliorer les pondérations pour non-participation (en se restreignant aux assurés des départements Constances aux dates d'inclusion) et pour donner des calculs extrapolés à la France métropolitaine.

9 ÉCHANTILLONNAGE ENVISAGÉ POUR LES VAGUES ULTÉRIEURES

La procédure décrite ci-dessus est celle qui sera appliquée pour la vague 1 de Constances. Les coûts entraînés par cette procédure sont importants, car il est nécessaire d'inviter environ deux millions de personnes pour arriver au total à 200 000 participants dans Constances si on attend une participation de l'ordre de 10 %. On peut donc envisager de moduler cette procédure, en recrutant moins de volontaires par tirage au sort, et d'ajouter en parallèle une stratégie de recrutement plus ciblée : en l'occurrence, il s'agit de proposer de participer à Constances aux personnes déjà convoquées dans un CES. En effet, on sait que le taux d'acceptation de Constances est très bon chez ces personnes, comme cela a été testé au cours de la phase pilote. Il ne serait pas valide de ne proposer Constances qu'à des personnes convoquées, pour de multiples raisons, essentiellement parce qu'en termes de représentativité, cela aurait exclu les personnes ayant une probabilité nulle d'être convoquées dans un CES (comme les bénéficiaires qui ignorent leur droit à passer un EPS, voire l'existence même des CES). De plus, pour constituer une cohorte de non participants, il est préférable de sélectionner ceux-ci parmi des personnes invitées, et donc informées d'un éventuel recueil passif d'informations les concernant.

Cependant, couplée avec des invitations sur tirage au sort - ce qui assure une couverture totale de la population cible - cette stratégie devient valide dès lors que l'on peut raisonnablement estimer : i) les probabilités d'être contacté selon chaque mode de recrutement : tirage au sort, convocation ou les deux ; ainsi que ii) les probabilités de participation à Constances *via* chacun des modes. Ces estimations sont plus ou moins complexes à réaliser selon la manière dont sont coordonnés les deux modes de recrutement. Une option envisageable serait de séparer la population cible aléatoirement (*via* la clé NIR par exemple) en deux sous-populations, qui seraient dédiées chacune à un des deux modes de recrutement. Il n'y aurait alors plus d'interférence entre les deux, et les calculs à effectuer seront proches de ceux qui sont décrits pour la vague 1 :

- pour la population échantillonnée par tirage au sort (à partir donc d'une base de sondage échantillonnée avec un taux r), on pourra procéder comme précédemment en multipliant les poids obtenus par 1/r ;
- pour la population échantillonnée par convocation, on pourra par exemple effectuer un calage des pondérations initiales (1/(1-r)) grâce aux deux types d'informations connues sur les invités par tirage au sort : suivi passif seulement pour non-participants, toutes les données Constances pour les participants.

Si cette option n'était pas possible, par exemple à cause de contraintes pratiques pour les CES, alors les calculs deviendrait plus complexes, mais toujours possibles. Il faudrait en particulier calculer

explicitement la probabilité qu'un assuré soit convoqué pour EPS, puis la probabilité qu'il accepte ou non Constances. Quoiqu'il en soit, même si plusieurs pistes sont envisageables, il sera toujours d'effectuer un calage sur marges pour les tous les participants directement.

En dehors du problème de représentativité, le problème de la précision des indicateurs obtenus à partir de Constances reste posé. Comme la proportionnalité de l'échantillon ne serait certainement plus respectée, certains domaines de la population seraient en faible effectif, et donc mal investigués. Une solution, dont la faisabilité opérationnelle doit être vérifiée, serait de contrôler les quotas de personnes à inclure parmi celles déjà convoquées selon le sexe, l'âge et la PCS. En tout état de cause, ce point est important à prendre en compte lors du choix du prorata à appliquer entre sujets inclus par tirage au sort et par convocation.