



INSPQ

INSTITUT NATIONAL
DE SANTÉ PUBLIQUE
DU QUÉBEC

Centre d'expertise
et de référence

Formation statistique sur la surveillance en santé publique

École de santé publique de l'Université de Montréal

Jérémie Sylvain-Morneau
Louis Rochette

Mai 2023

BiESP

BUREAU D'INFORMATION
ET D'ÉTUDES EN SANTÉ
DES POPULATIONS

*Institut national
de santé publique*

Québec 

Plan et horaire de la formation

Première demi-journée: 25 mai de 9h à 12h

- 9:00 Présentation et plan de la formation (Louis)
- 9:10 Contexte entourant la surveillance et domaines de surveillance (Louis)
- 9:20 Introduction aux données administratives (Louis)
- 9:30 Indicateurs associés aux bases de données administratives (Louis)
→ Période de questions
- 10:10 Un cas est-il un cas ? (Louis)
- 10:20 Pause de 15 minutes
- 10:35 Données d'enquêtes (Jérémie) → Période de questions
- 11:40 Survol méthodologique: standardisation (Louis)
- 12:00 Fin de la première partie

Plan et horaire de la formation

Deuxième demi-journée: 26 mai de 9h à 12h

- 9:00 Survol méthodologique: Tests statistiques et intervalles de confiance (J)
→ Période de questions
- 10:00 Survol méthodologique: Association et modélisation statistique (Louis)
→ Période de questions
- 10:45 Pause de 15 minutes
- 11:00 Sujets spéciaux
- Indice de défavorisation (Jérémy)
 - Indice de multimorbidité (Louis)
 - Exemple d'étude de cohorte (Louis)
 - Dénombrement des personnes en situation d'itinérance (Jérémy)
- 12:00 Fin et évaluation

La surveillance en santé publique : de quoi s'agit-il?

www.inspq.qc.ca

Définition de la surveillance

La définition de la surveillance est la suivante :

- Un processus continu de collecte, d'analyse et d'interprétation de données sur des événements de santé spécifiques importants pour la planification, la mise en œuvre, et l'évaluation des pratiques en santé publique, afin de soutenir la décision et pouvoir offrir à la population des services appropriés.
 - Caractère systématique et pérenne
 - Circulation de l'information
 - Servir l'action en santé publique



Objectifs de la surveillance

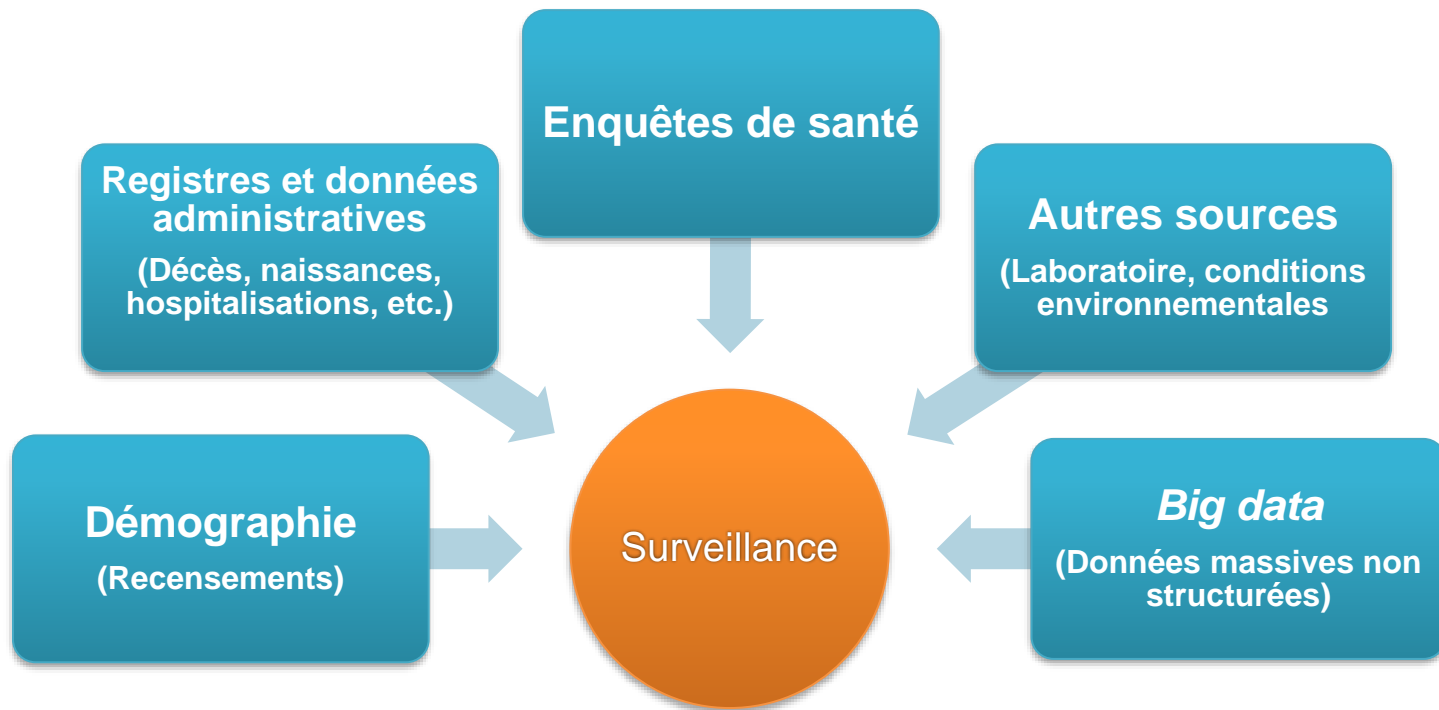
- Dresser un portrait global de l'état de santé de la population
- Observer les tendances et les variations temporelles et spatiales
- Détecter les problèmes en émergence
- Identifier les problèmes prioritaires
- Élaborer des scénarios prospectifs de l'état de santé de la population
- Suivre l'évolution, au sein de la population, de certains problèmes spécifiques de santé et de leurs déterminants.

Rappel des finalités analytiques de la surveillance en santé publique

Question sur le problème	Finalité analytique
Combien de personnes affecte-t-il? Combien compte-t-on de nouveaux cas? Combien en décède-t-il?	Ampleur du phénomène
Le phénomène augmente-t-il, stagne-t-il ou diminue-t-il? Connaît-il des cycles saisonniers?	Histoire naturelle du phénomène et détection de problèmes en émergence
Est-il réparti de façon homogène sur le territoire ? Se concentre-t-il dans l'espace ou se propage-t-il?	Distribution géographique du phénomène
Affecte-t-il certains groupes particuliers de la population ?	Groupes plus à risque, tests d'hypothèses

(Lee et al., 2010)

Les sources de données en surveillance

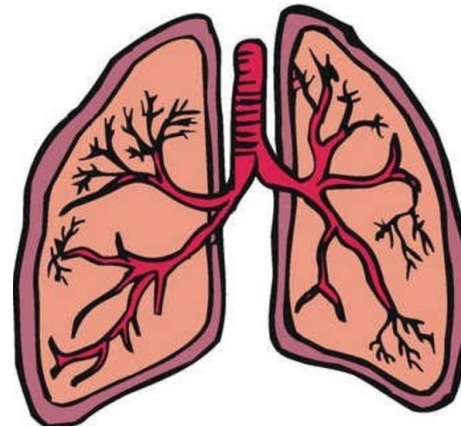


Domaines de surveillance à l'INSPQ

www.inspq.qc.ca

Domaines de surveillance

- La surveillance en santé publique s'applique à une panoplie de domaines. Ceux couverts par le Bureau d'information et d'études en santé des populations (BIESP) de l'INSPQ incluent:
 - Maladies chroniques et troubles mentaux



Domaines de surveillance

- Habitudes de vie



Domaines de surveillance

- Traumatismes



1 866 2773553 1 866 2773553 1 866 2773553
1 866 2773553 1 866 2773553 1 866 2773553
1 866 2773553 **1 866 APPELLE** 1 866 2773553
1 866 2773553 1 866 2773553 1 866 2773553
1 866 2773553 1 866 2773553 1 866 2773553

ça soulage!

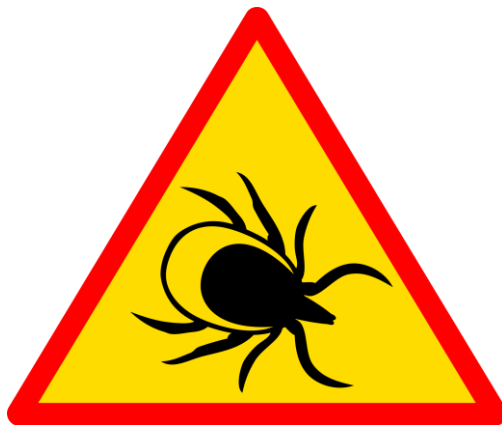
Domaines de surveillance

- Inégalités sociales de santé



Domaines de surveillance

- Outre la surveillance effectuée au BIESP, l'INSPQ s'occupe de la surveillance des:
 - Maladies infectieuses dont bien sûr la COVID-19



Domaines de surveillance

- Outre la surveillance effectuée au BIESP, l'INSPQ s'occupe de la surveillance des:
 - Infections nosocomiales



Campylobacter



Enterococcus faecalis



Clostridium difficile

Domaines de surveillance

- Outre la surveillance effectuée au BIESP, l'INSPQ s'occupe de la surveillance des:
 - Infections transmises sexuellement et par le sang



Domaines de surveillance

- Outre la surveillance effectuée au BIESP, l'INSPQ s'occupe de la surveillance des:
 - Immunisation dont la vaccination contre la COVID-19



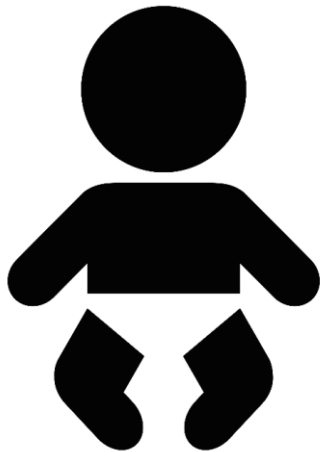
Registres et données administratives

www.inspq.qc.ca

Portée et objet (1)

Constitués à des fins administratives:

- Pour des exigences légales (et le décompte de la population)



Portée et objet (2)

Constitués à des fins administratives:

- Pour le paiement des médecins:

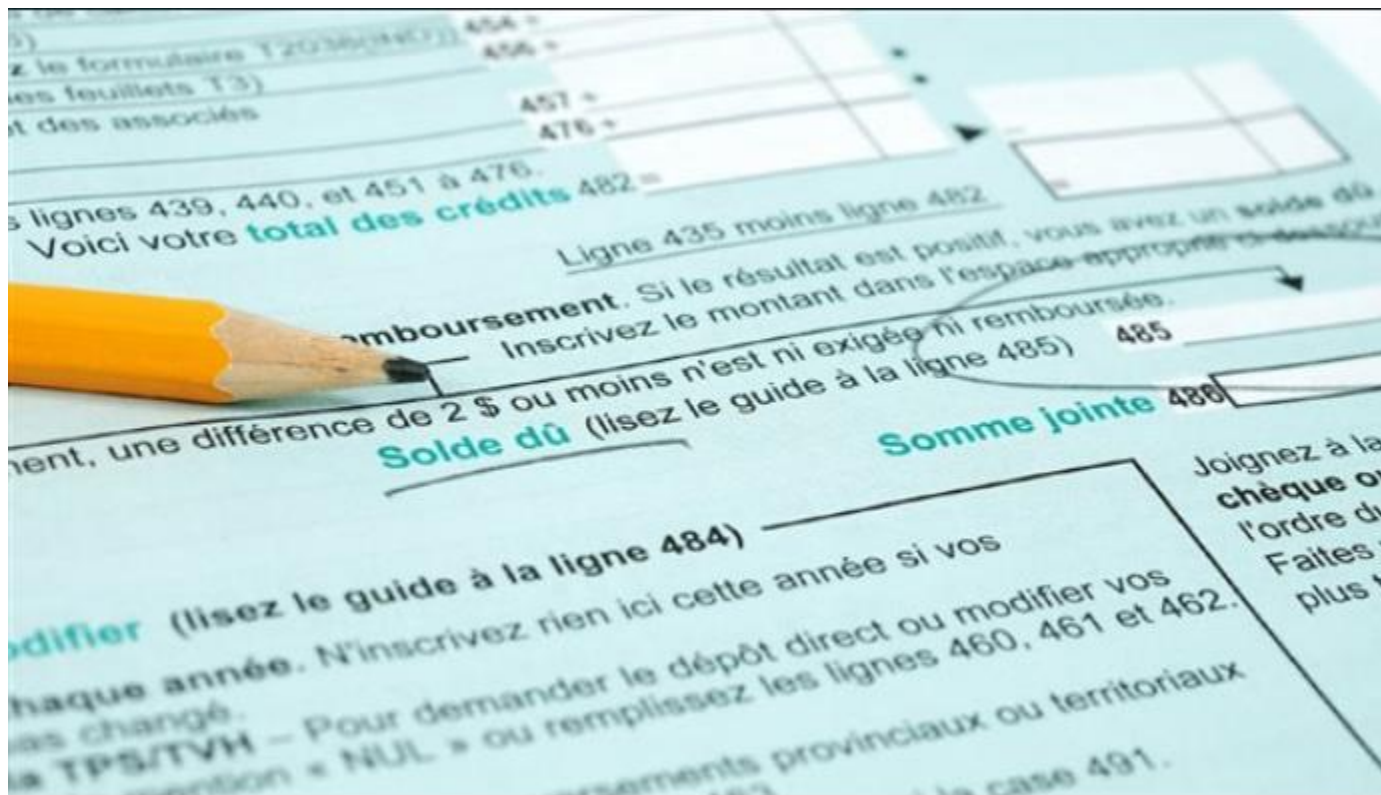


Source de l'image: Québec en tête

Portée et objet (3)

Constitués à des fins administratives:

- Pour des prélèvements obligatoires:



Portée et objet (4)

Constitués à des fins administratives:

- Pour la gestion des avantages sociaux:



Fichiers administratifs utiles en santé

Une liste non-exhaustive:

- Registre des décès, des naissances et mortinaissances
- Fichier des hospitalisations (MED-ÉCHO)
- Fichier des personnes assurées à l'assurance maladie
- Fichier des services médicaux rémunérés à l'acte
- Fichier des services pharmaceutiques
- Registre québécois du cancer
- Registre des traumatismes/suicides
- Registre de vaccination du Québec
- Dossier Santé Québec

Avantages et inconvénients des fichiers ou registres administratifs

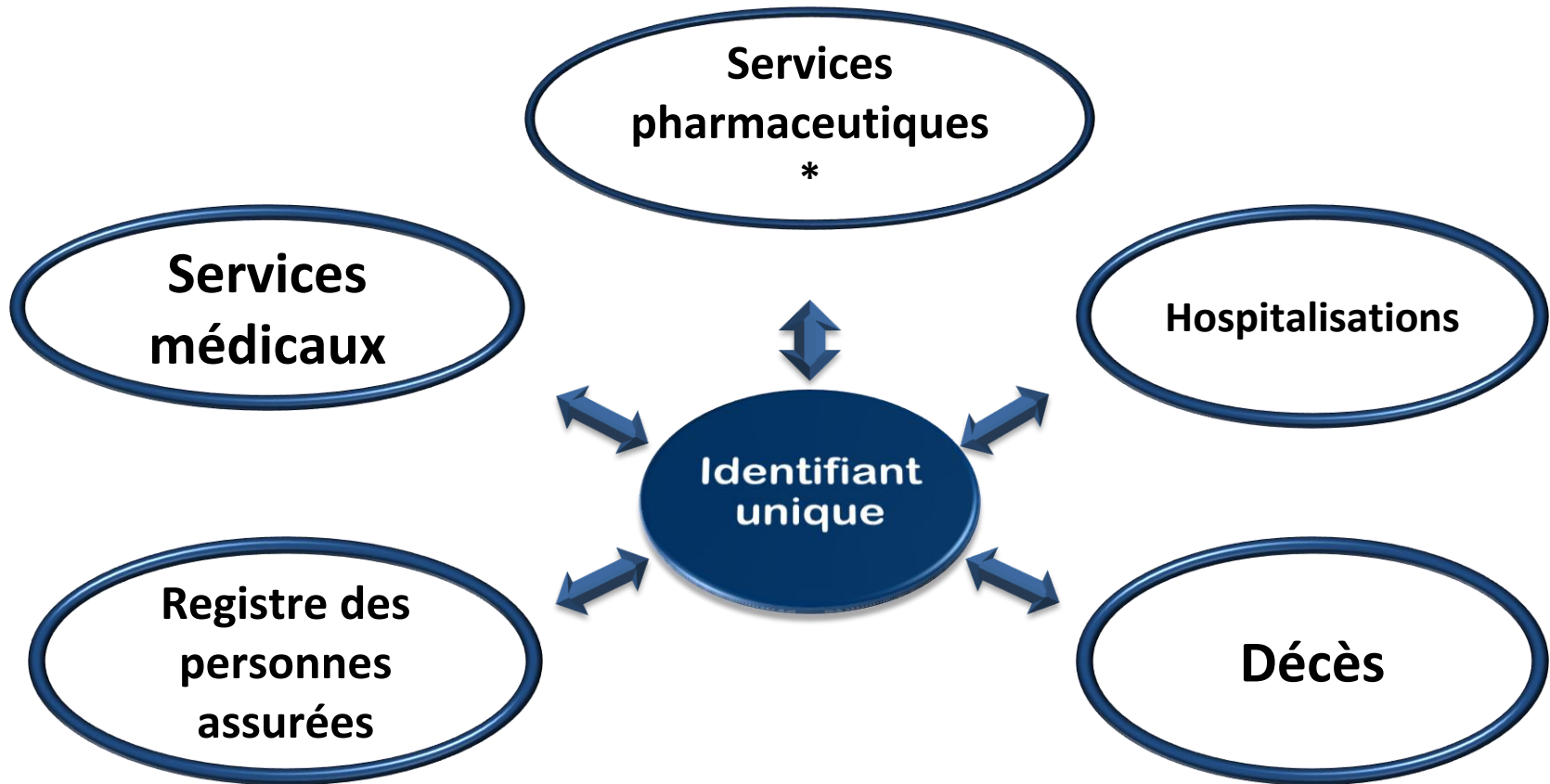
Avantages

- Couvrent généralement bien la population ou la clientèle visée (population totale, étudiants, patients, prestataires, etc.).
- Sont d'une assez bonne qualité.
- Sont mis à jour régulièrement et assez bien documentés.
- Fournissent généralement des estimations précises à différentes échelles territoriales, et avec un bon recul historique.
- Peuvent souvent être jumelés (de manière plus ou moins complète) et permettent donc des analyses complexes et très informatives (parcours de vie, parcours de soins, etc.).
- Pas de coûts supplémentaires pour la cueillette, et aucun fardeau sur les répondants.

Inconvénients

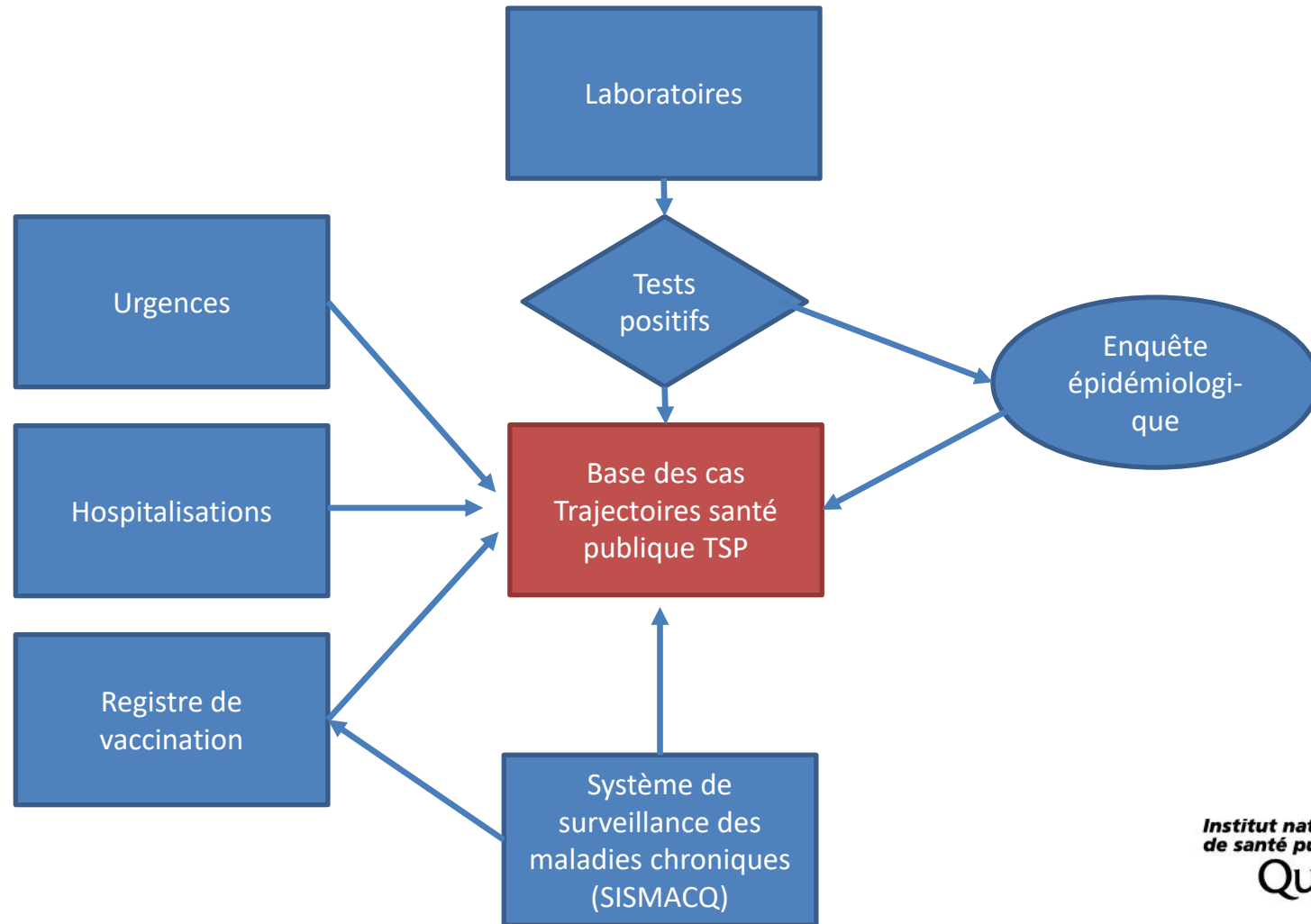
- **N'ont pas été principalement conçus à des fins de surveillance et nous sommes à la merci des propriétaires.**
- Ne contiennent généralement que peu d'information sur les caractéristiques socioéconomiques, les habitudes de vie et les comportements des individus.
- Ne fournissent donc qu'une image partielle d'un phénomène → sous-déclaration les plus importantes pour les problèmes de santé non sévères, au diagnostic plus difficile, ou pour les problématiques socialement sensibles. Au Québec, leur accès et leur utilisation sont souvent régis par des clauses légales très contraignantes.
- La qualité des données peut varier d'un fichier à l'autre et il peut y avoir des incohérences.
- La taille des bases de données peut être massive.

SISMACQ



* Individus sélectionnés seulement

Fichiers jumelés COVID-19



Indicateurs associés aux bases de données administratives

www.inspq.qc.ca

Indicateurs provenant des données administratives

- Les mesures de base simples comme les proportions et les moyennes sont davantage associées aux enquêtes, mais s'appliquent généralement aussi aux données administratives.
- Cependant, le caractère longitudinal et souvent populationnel de ce type de données permet d'aller plus loin.
- Dans un contexte de surveillance de la santé, nous discuterons des mesures de prévalence, d'incidence et de mortalité en introduisant d'abord les concepts de population à risque et de taux.

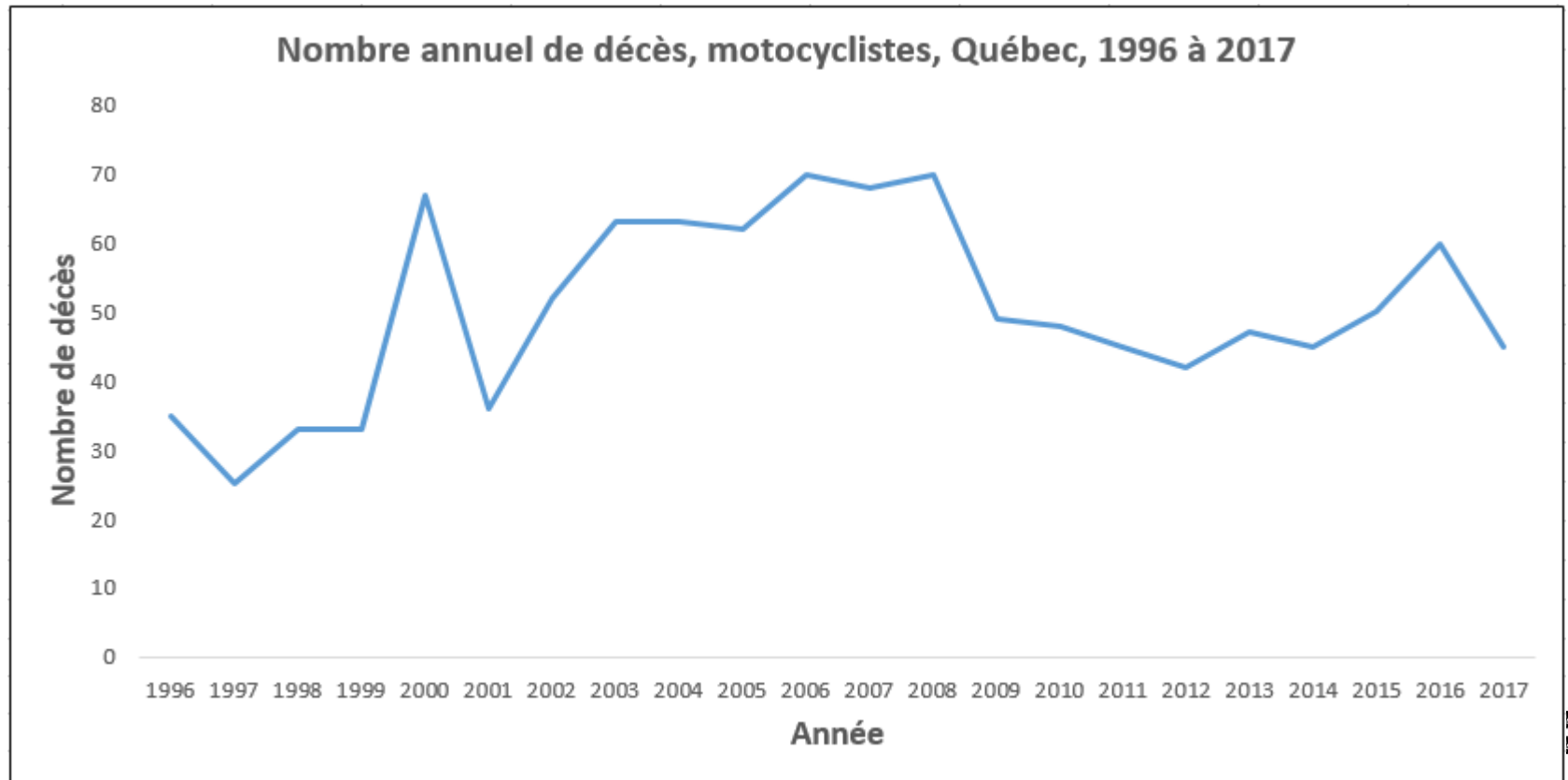
Population à risque (et dénominateur)

Pour calculer les mesures les plus représentatives possibles, le dénominateur doit seulement inclure les individus susceptibles de développer la maladie à l'étude.

- Tous les individus sont à risque de mourir.
- Seuls les individus admissibles à la RAMQ sont à risque de consommer des services rémunérés à l'acte au Québec.
- Seuls les hommes sont à risque de développer un cancer de la prostate.
- Seules les femmes sont à risque de donner naissance.
- Seuls les individus mariés sont à risque de divorcer.

Population à risque (et dénominateur)

Il n'est pas toujours possible de calculer la population réellement à risque:



Population à risque (et dénominateur)

Il n'est pas toujours possible de calculer la population à risque optimale:

- Dans l'exemple précédent, on doit se résoudre à utiliser la population totale en âge de conduire comme dénominateur.
- Une meilleure population à risque serait le nombre d'individus ayant pratiqué la moto au moins une fois dans l'année.
- Dans les faits, l'exposition la plus exacte tiendrait aussi compte du nombre d'heures sur la route au volant d'une moto.

Population à risque (et dénominateur)

Déjà 10 motocyclistes morts sur les routes cette année

f 1,1K

PARTAGEZ SUR FACEBOOK



PARTAGEZ SUR TWITTER



AUTRES

FRÉDÉRIQUE GIGUÈRE

Mardi, 31 mai 2016 06:30
MISE à JOUR Mardi, 31 mai 2016 06:30

L'été ne fait que commencer et déjà un motocycliste a trouvé la mort hier sur les routes du Québec. Une telle nouvelle présage une saison désastreuse pour les amateurs de deux roues.

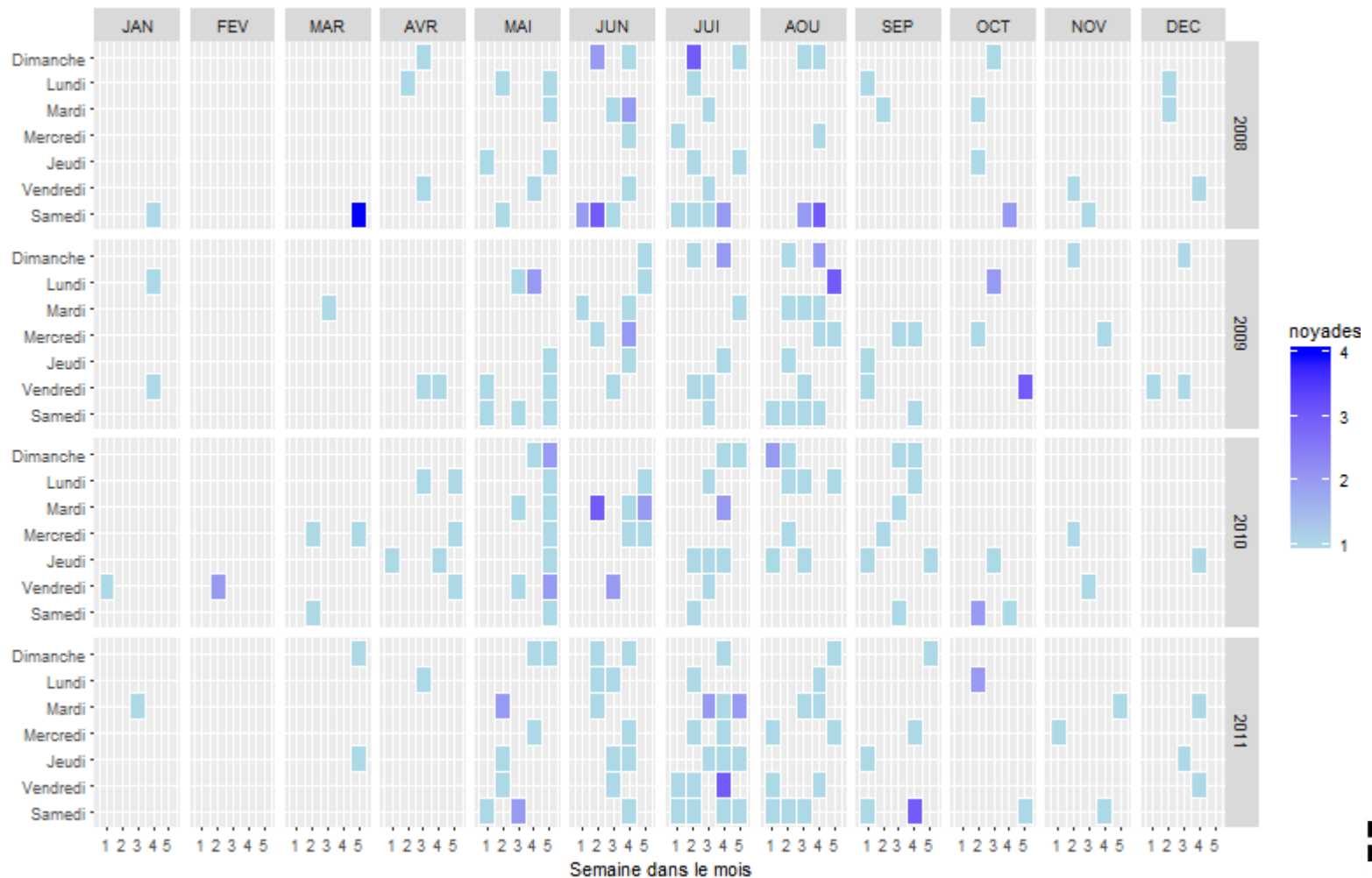
Un homme de 40 ans a été fauché par un poids lourd hier après-midi, dans Lanaudière.

Ces deux décès porte à 10 le nombre de motocyclistes tués au volant de leur engin en avril et mai. On ne dénombrait que sept victimes à pareille date l'an dernier. Pourtant, 2015 affichait le pire bilan depuis le début de la décennie.

● 54 motocyclistes sont décédés, soit 4 décès de plus qu'en 2015

Population à risque (et dénominateur)

Calendrier des noyades au Québec, 2008 à 2011

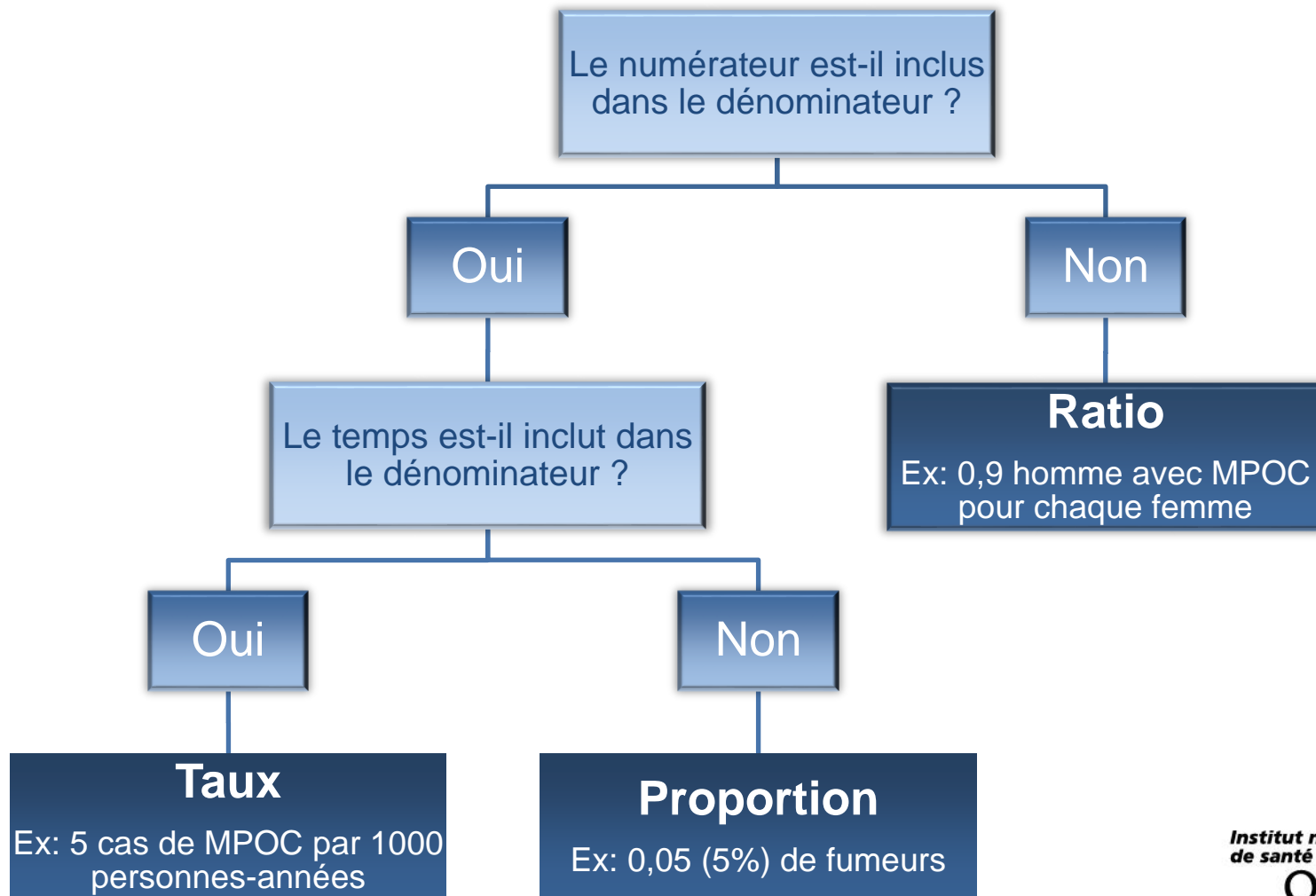


Population à risque (et dénominateur)

Pour la surveillance des maladies chroniques:

- L'ensemble de la population est à risque de développer la plupart des maladies couvertes.
- Sauf les individus ayant déjà la maladie !
- Aussi, les individus non couverts par la RAMQ ne peuvent pas être suivis par les données administratives.
- La surveillance de certaines maladies est restreinte à certains groupes d'âge (ex: peu utile de suivre l'Alzheimer chez les moins de 20 ans).

Sommes-nous en présence d'un taux ?



Mesure de prévalence

La prévalence:

- Répond à la question « Combien d'individus de la population ont la maladie à un moment précis ? »
- En surveillance, à partir de données administratives, la « photo » est prise à la fin de l'année
- Se calcule ainsi (prévalence brute):

$$\frac{\text{Nombre d'individus à risque ayant la maladie}}{\text{Nombre d'individus à risque dans la population à l'étude}}$$

Mesure d'incidence

L'incidence:

- Répond à la question « À quelle vitesse/rythme la maladie se développe-t-elle dans la population ? ».
- Se calcule ainsi (incidence brute):

$$\frac{\text{Nombre de nouveaux cas (malades) sur la période}}{\text{Nombre **d'unités** à risque d'être un cas sur la période}}$$

- C'est donc un taux puisque le temps est une composante majeure. Plus l'intervalle de temps est réduit, plus le taux devient instantané.

Concept de personnes-temps

Le dénominateur d'un taux doit tenir compte du temps:

- Les taux annuels sont majoritairement utilisés dans la littérature scientifique médicale
- « L'unité » au dénominateur devient donc des personnes-années
- Plus rarement, on peut rencontrer des personnes-jours, des personnes-mois ou des personnes-périodes

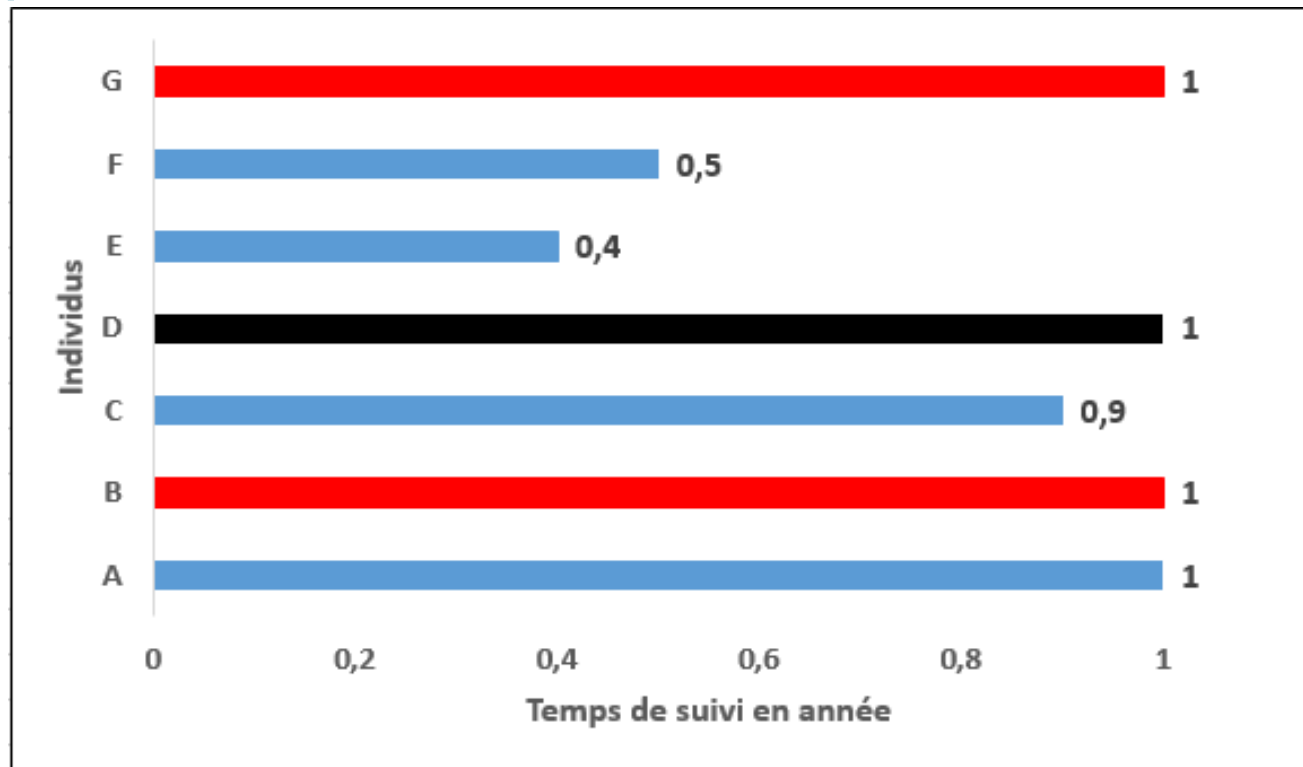
Concept de personnes-temps

Le temps de suivi (à risque) peut varier d'un individu à l'autre pour plusieurs raisons:

- Décès
- Migration
- Devient inadmissible
- Biologiquement non à risque
- Pas exposé
- Immunisé

Concept de personnes-temps

Exemple de calcul du taux d'incidence:

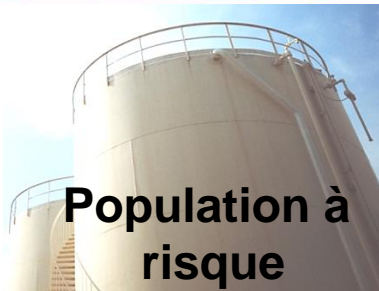


- Cas incident de MPOC
- Non-cas
- Cas ayant déjà une MPOC

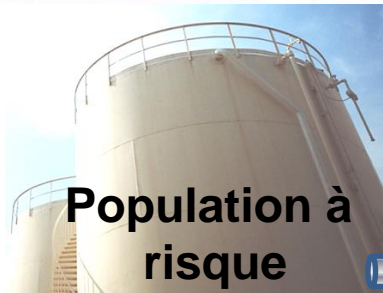
Concept de personnes-temps

- Les taux peuvent être exprimés par personnes-années, par 100 personnes-années, par 1000 personnes-années, etc.
- Ce choix dépend généralement de la rareté de la maladie ou de l'événement à l'étude.
- Peu importe, les mesures sont équivalentes:
1,5 cas par 100 personne – années = 15 cas par 1000 personnes – années

Résumé



Résumé

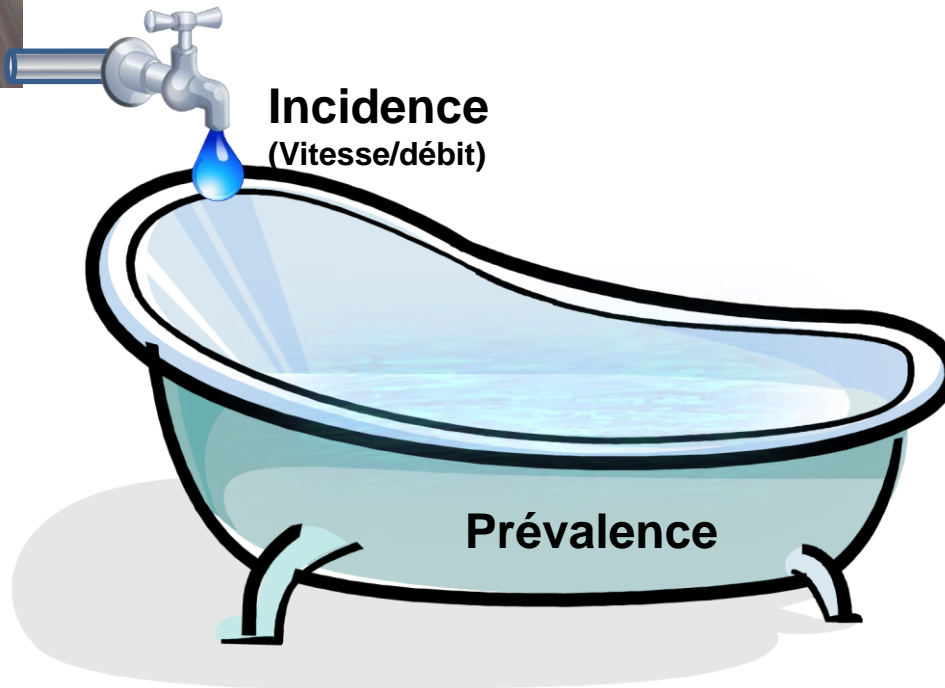
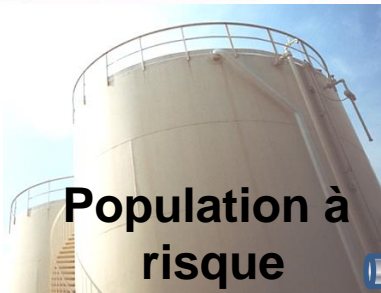


**Population à
risque**

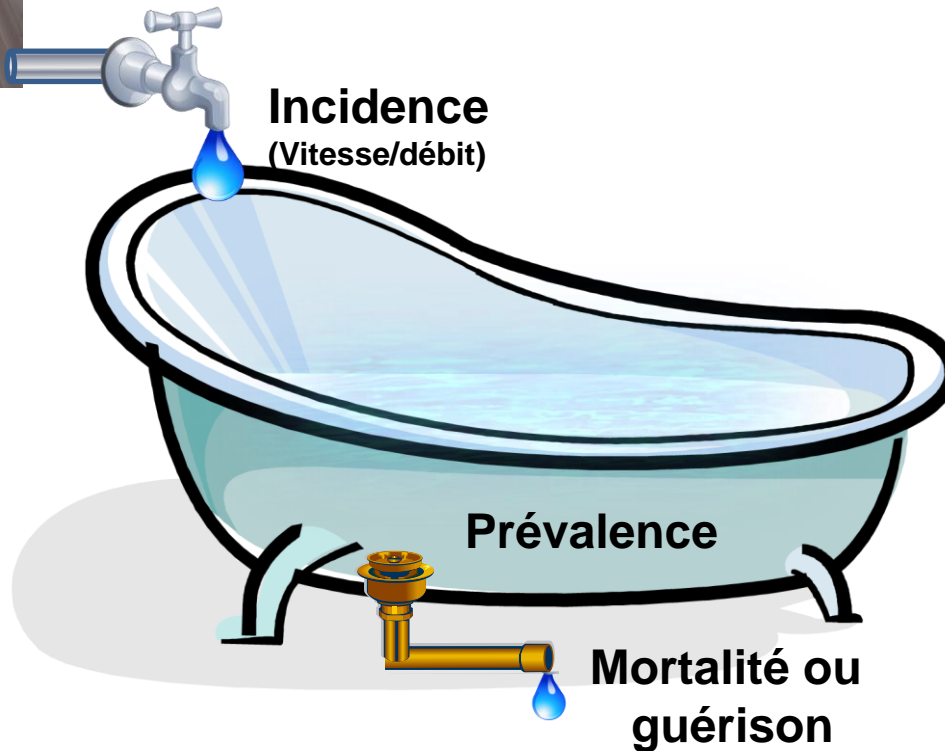
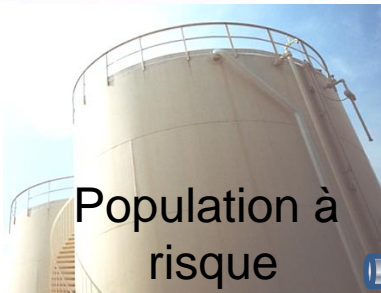


Incidence
(Vitesse/débit)

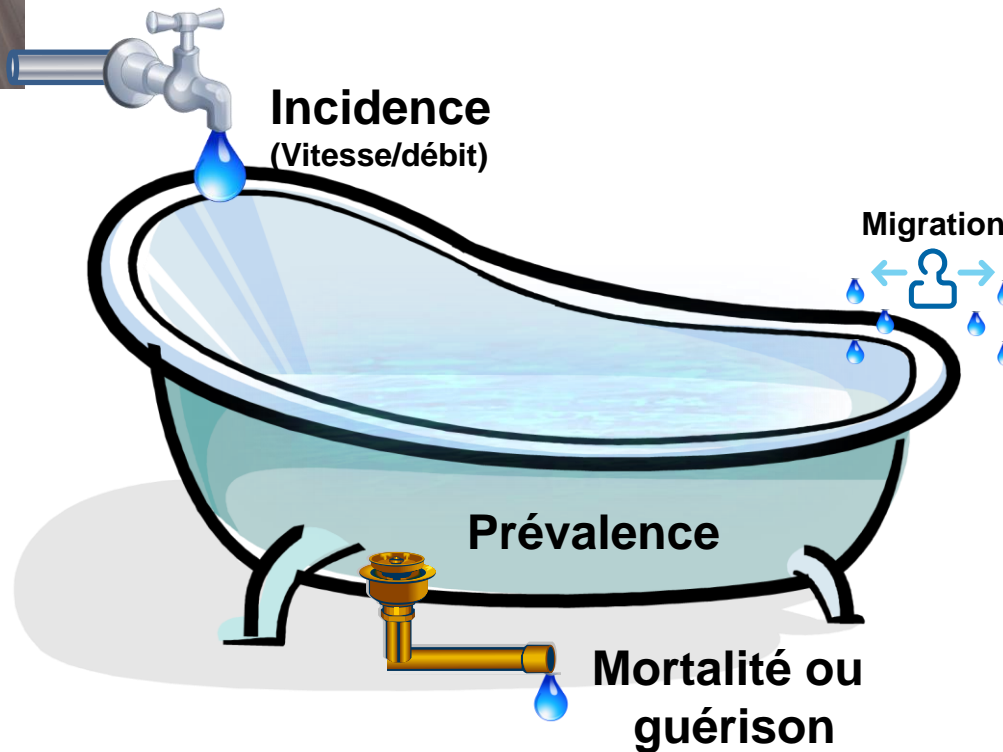
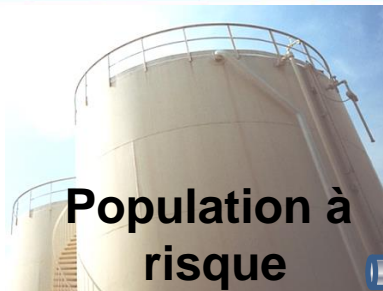
Résumé



Résumé



Résumé



Résumé

Équivalent dans le domaine de la physique:

- Le taux d'incidence est équivalent au concept de vitesse. Même si elle s'exprime souvent en kilomètres à l'heure, la vitesse peut varier à chaque instant. De même, le taux d'incidence s'exprime en personnes-années, mais il peut varier constamment.
- La prévalence est un peu équivalente au concept de distance parcourue. La distance augmente en fonction de la vitesse, alors que la prévalence augmente en fonction de taux d'incidence.
- La mortalité/guérison/migration sont en quelque sorte équivalentes à la friction.

Mise en situation

santé

recherche

Innovation

centre d'expertise et de référence

infectieuses

promotion de sair

santé environnementale

toxicologie

prévention des maladies chroniques

impact des politiques pub

santé au travail

développement des personnes et des communautés

www.inspq.qc.ca

de l'état de santé de la population

microbiologie

sécurité et prévention des traumatismes

recherche

santé au tra

**Institut national
de santé publique**

Québec 

Mise en situation



Le directeur de santé publique de votre région vous demande de documenter la problématique des maladies pulmonaires obstructives chroniques (MPOC) au Québec en lien avec les cas positifs de COVID-19 en utilisant les données de surveillance disponibles. Est-ce que les cas positifs sont plus souvent associés aux gens aux prises avec des maladies chroniques dont les MPOC ?

Il désire également savoir si les gens de sa région suivent les recommandations de la santé publique car il semble que dernièrement, le taux de cas positifs de la COVID-19 est plus important dans sa région comparativement au reste du Québec.

Prévalence: retour sur la mise en situation

Fin 2018-2019

5 028 450 Québécois
de 35 ans et plus
admissibles à la RAMQ

508 270 (10,1%) de
ceux-là ont une MPOC

Prévalence

Fin 2019-2020

5 091 745 Québécois
de 35 ans et plus
admissibles à la RAMQ

518 660 (10,2%) de
ceux-là ont une MPOC

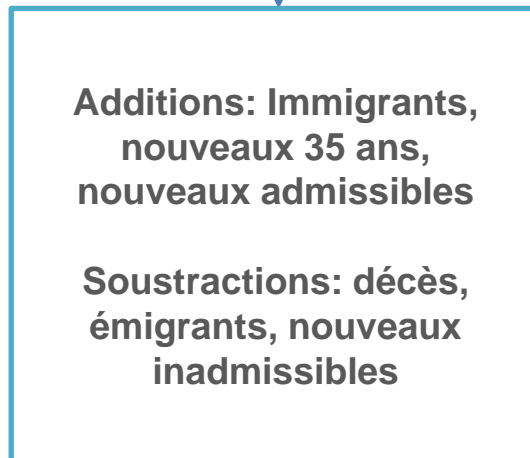
**Hausse du nombre d'individus
dans la population à l'étude
(dénominateur de la prévalence)**

**Hausse du nombre d'individus
avec la maladie
(numérateur de la prévalence)**

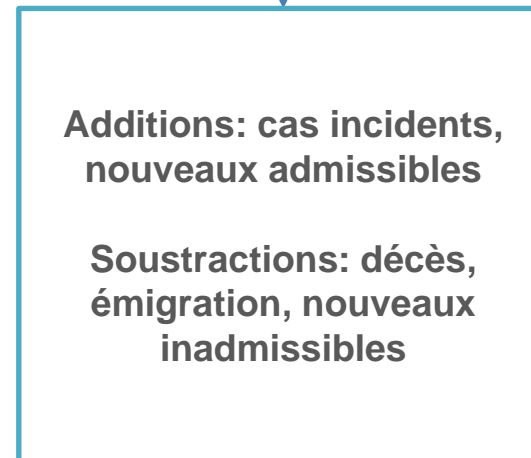


Prévalence: retour sur la mise en situation

**Hausse du nombre d'individus
dans la population à l'étude
(dénominateur de la prévalence)**



**Hausse du nombre d'individus
avec la maladie
(numérateur de la prévalence)**



Incidence: retour sur la mise en situation

Fin 2018-2019

5 028 450 Québécois
de 35 ans et plus
admissibles à la RAMQ

508 270 (10,1%) de
ceux-là ont une MPOC

Fin 2019-2020

5 091 745 Québécois
de 35 ans et plus
admissibles à la RAMQ

518 660 (10,2%) de
ceux-là ont une MPOC

+10 390

Au courant
de 2019-2020

36 955 cas incidents et
26 510 décès pour un
ajout de +10 445

Taux
d'incidence

36 955 cas incidents divisés par
4 610 040 Québécois à risque
(exclut les cas déjà prévalents)

À risque
≈
5 091 745 – 518 660

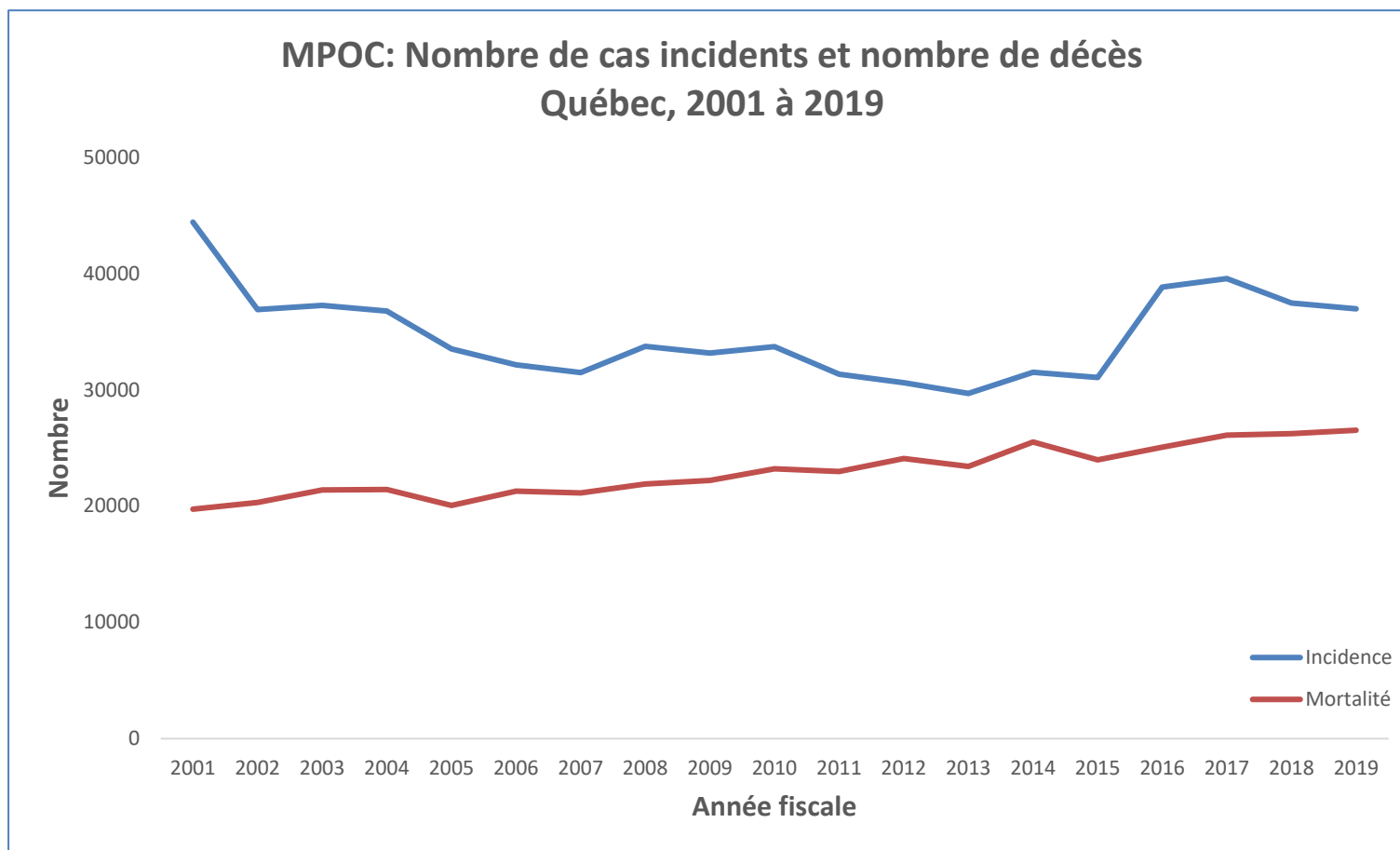
Taux de
mortalité

26 510 décès divisés par
518 660 Québécois à risque
(nombre de cas prévalents)

5,1%
Institut national
de santé publique
Québec



Incidence: retour sur la mise en situation



Données administratives: quand un cas est-il un vrai cas ?

www.inspq.qc.ca

Quand un cas est-il un vrai cas ?

Dans un monde idéal:

- L'accès direct aux dossiers médicaux de tous les individus à risque serait autorisé
- Ces dossiers médicaux seraient soigneusement complétés, à jour, et parfaitement lisibles/compréhensibles
- Les diagnostics auraient été confirmés indépendamment par au moins un autre médecin



Quand un cas est-il un vrai cas ?

Dans la réalité, dans les données administratives:

- Un seul diagnostic de MPOC est-il suffisant pour conclure qu'un individu a développé la maladie pendant l'année ?
- Le diagnostic d'un pneumologue est-il plus valide que celui d'un omnipraticien ?
- Un diagnostic de MPOC associé à une hospitalisation n'a-t-il pas plus de poids qu'un diagnostic lors d'une simple consultation médicale ?
- Les changements apportés à la rémunération des médecins incitent-ils ces derniers à diagnostiquer plus ou moins la maladie ?

Définition de cas et études de validation

Plusieurs définitions de cas sont testés. Des exemples:

Abbréviation et description	Avantages et inconvénients
<u>1P</u> : un seul diagnostic provenant du fichier des services médicaux rémunérés à l'acte	+ Détection de presque tous les cas réels - Détection de plusieurs non-cas
<u>1H</u> : un diagnostic de la maladie lors d'une hospitalisation	+ Peu de non-cas identifiés - Détection de cas graves seulement
<u>1H3P</u> : un diagnostic de la maladie lors d'une hospitalisation OU trois diagnostics provenant du fichier des services médicaux rémunérés à l'acte (trois jours différents)	+ Bon compromis - Est-on trop ou pas assez sévère ?
<u>1H1P1Rx</u> : un diagnostic de la maladie lors d'une hospitalisation OU un diagnostic provenant du fichier des services médicaux rémunérés à l'acte OU une prescription d'un médicament associé à la maladie	+ Combine de l'information de trois sources d'informations - Les services pharmaceutiques ne couvrent pas l'ensemble de la population.

Définition de cas et études de validation

Lors d'une étude de validation:

- Les dossiers médicaux d'un échantillon de la population sont étudiés afin de vérifier la présence de la maladie.
- Ces dossiers médicaux sont considérés comme la référence, le *gold standard* (mesure étalon).
- Chaque définition de cas testée est comparée avec la mesure étalon.
- Différentes mesures de performance sont calculées, permettant de comparer et de choisir la meilleure définition.

Définition de cas et études de validation

Sensibilité, spécificité, valeur prédictive positive (VPP) et valeur prédictive négative (VPN):

		Mesure étalon	
		Malade	Non-malade
Définition de cas	Positif	Vrais positifs (VP)	Faux positifs (FP)
	Négatif	Faux négatifs (FN)	Vrais négatifs (VN)

Définition de cas et études de validation

- La sensibilité est la probabilité que le test (définition de cas) soit positif si la maladie est vraiment présente.

$$Se = \frac{VP}{VP + FN}$$

- La spécificité est plutôt la probabilité d'obtenir un test négatif si la maladie n'est pas présente.

$$Sp = \frac{VN}{VN + FP}$$

Définition de cas et études de validation

- La valeur prédictive positive est la probabilité que la maladie soit présente lorsque le test est positif.

$$VPP = \frac{VP}{VP + FP}$$

- La valeur prédictive négative est la probabilité que la maladie ne soit pas présente lorsque le test est négatif.

$$VPN = \frac{VN}{VN + FN}$$

Définition de cas et études de validation

Aucune définition de cas ne maximise les quatre mesures en même temps, et la meilleure définition varie d'une maladie à l'autre:

Abbréviation et description	Constats généraux
<u>1P</u>	Excellente sensibilité Mauvaise VPP
<u>1H</u>	Excellente VPP Mauvaise sensibilité
<u>1H3P</u>	Sensibilité moyenne VPP moyenne
<u>1H1P1Rx</u>	Sensibilité élevée VPP médiocre Mesures varient selon l'âge

Données provenant d'enquêtes sur la santé

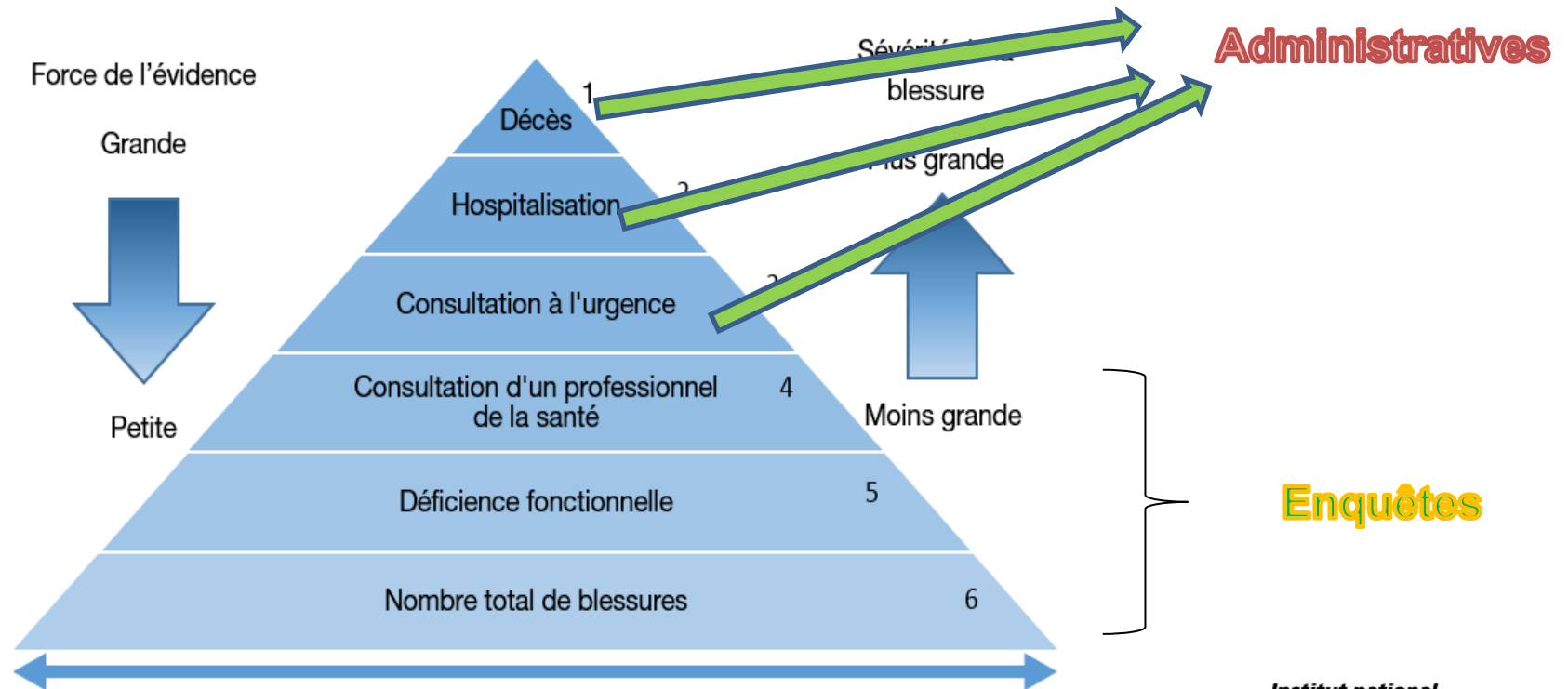
www.inspq.qc.ca

Utilité des enquêtes sur la santé

- Pour recueillir des informations non contenues dans les autres sources de données disponibles :
 - Caractéristiques socio-démographiques
 - Habitudes de vie : alimentation, statut tabagique, activité physique
 - Comportements, attitudes, perceptions
- Pour documenter des phénomènes ou des problèmes de santé émergents (ex. COVID-19)
- Pour rejoindre une population peu présente dans les données administratives
 - Population en situation d'itinérance

Complément aux données administratives

Pyramide des effets des traumatismes sur le système de santé



Utilité des enquêtes sur la santé

- Pour documenter une problématique de santé:
 - Son évolution
 - Les personnes les plus à risque
 - Ses déterminants, sa perception dans la population
 - Son fardeau sanitaire, social et économique



Type d'enquêtes

- Par recensement (Ex. Recensement du Canada de 2021)

- Par échantillonnage

- **Transversale**

VS → Portrait d'une population en un temps précis

- **Longitudinale**

→ Suivi d'un échantillon de répondants à différents moments dans le temps (vagues).

- **Probabiliste**

VS → Échantillons tirés aléatoirement

- **Non probabiliste (non scientifique)**

Ex.: sondage d'opinion, panel Web formé de volontaires

Principales étapes d'une enquête

1. Définir la population-cible

À qui s'adresse-t-elle?

- Enquêtes générales de santé
 - Personnes âgées de 12 (ou de 15) ans et plus vivant en ménages privés au Québec
 - Exclusions : membres des Forces armées, personnes vivant sur les réserves indiennes, dans les institutions et dans certaines régions nordiques
- Populations particulières
 - Jeunes du secondaire, communautés culturelles, Cris, Inuits

Principales étapes d'une enquête

2. Bases de sondage

Comment rejoindre les personnes de la population-cible?

- Base téléphonique. Génération aléatoire de numéros de téléphone (GANT) pour couvrir les numéros confidentiels et les cellulaires.
- Base aréolaire (géographique)
- Liste de noms, d'entreprises. Ex. registre de la RAMQ
- Liste d'adresses
- Liste de courriels – panel Web.

- Mixte : plusieurs bases à la fois.

Principales étapes d'une enquête

3. Plan d'échantillonnage

Comment seront sélectionnés les répondants de l'échantillon?

- Échantillonnage non probabiliste
 - Échantillon à portée purement volontaire. Exemples: Panel Web, sondage d'opinion (télévision, centre d'achats, ...)
 - Quotas
 - Pas possible de calculer des marges d'erreur (mesures de précision)
 - Certains individus de la population n'ont aucune chance d'être sélectionnés

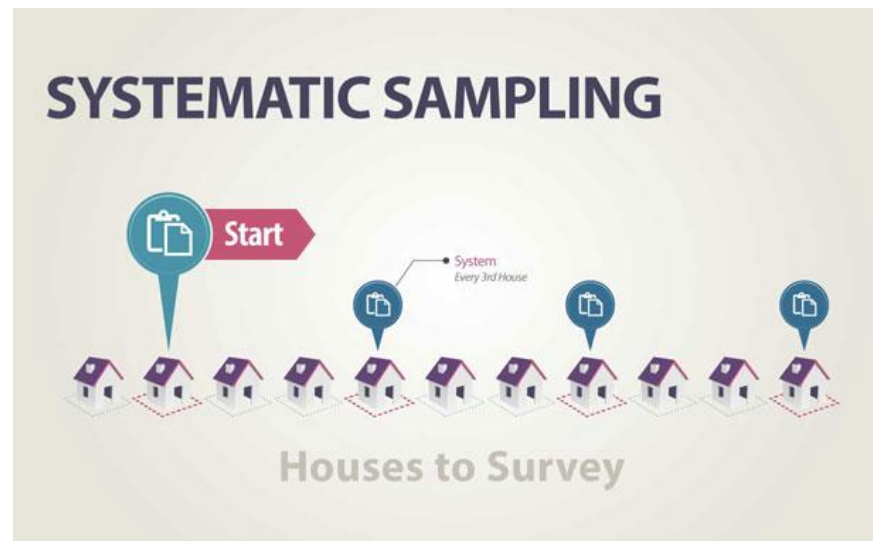
Principales étapes d'une enquête

3. Plan d'échantillonnage

- Échantillonnage probabiliste
 - Aléatoire simple avec ou sans remise



- Systématique

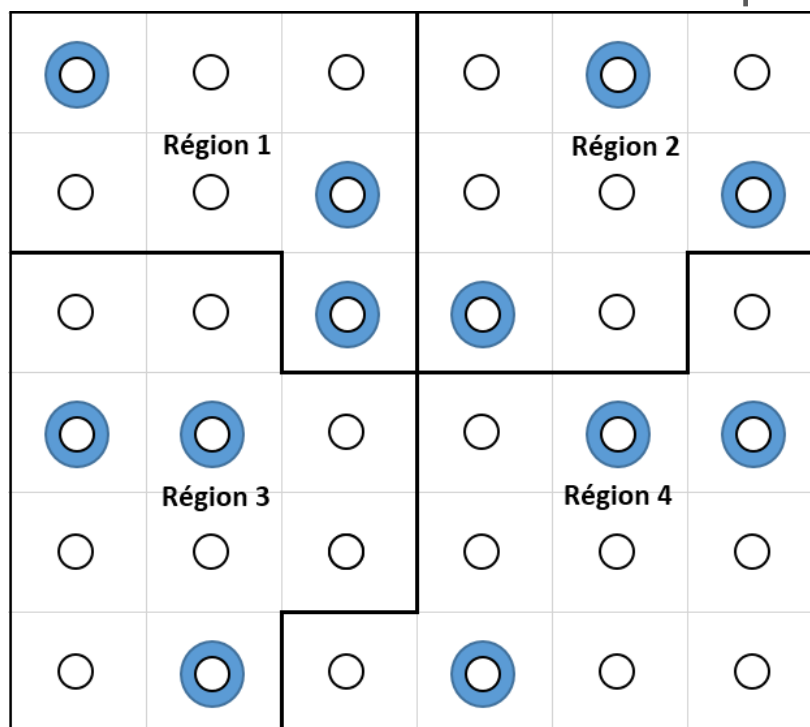


Principales étapes d'une enquête

3. Plan d'échantillonnage

- Échantillonnage probabiliste

- Stratifié : Permet de bien couvrir la population à l'étude : ex. région. Allocation de la taille d'échantillon proportionnelle à la taille.



Principales étapes d'une enquête

3. Plan d'échantillonnage

- Échantillonnage probabiliste
 - Complexe :
 - Plusieurs degrés. Région, quartier, logement, personnes.
Ex.: Enquêtes-ménages.

Principales étapes d'une enquête

3. Plan d'échantillonnage

- Détermination de la taille d'échantillon dépend
 - de la portée de l'étude : provinciale, régionale ;
 - des domaines d'estimation : par sexe, par groupe d'âge, ... ;
 - de la précision désirée autour de l'estimation. Par exemple l'intervalle de confiance ;
 - des taux de réponse attendus ;
 - Possibilité de sur-échantillonner certains sous-groupes de la population
 - Produire des estimations avec une bonne précision pour ces sous-groupes
 - Phénomène à l'étude touche davantage ces sous-groupes
 - Taux de réponse attendus plus faibles dans ces sous-groupes

Principales étapes d'une enquête

4. Instruments de collecte

Qu'est-ce qu'on désire obtenir comme information?

- Questionnaires
 - Définir les concepts à mesurer
 - Validation et compréhension des questions
- Mesures cliniques ou anthropométriques
- Considérations éthiques
 - Préserver la confidentialité des réponses
 - Faire attention aux questions sensibles (santé sexuelle, revenu)

Principales étapes d'une enquête

4. Instruments de collecte

Exemples d'enjeux liés à la formulation d'un questionnaire (tirés du dénombrement des personnes en situation d'itinérance, qui sera abordé plus loin).

9. Est-ce que la perte de votre dernier domicile est liée à la pandémie de la COVID-19?

(ex. : perte d'emploi due à la pandémie, conflit en lien avec la COVID-19 avec des colocataires)

OUI :

a. PERTE D'EMPLOI

b. CONFLITS

c. AUTRE _____

d. NON

e. NE S'APPLIQUE PAS
(SANS DOMICILE AVANT LA
PANDÉMIE)

f. NE SAIT PAS

g. REFUS

Principales étapes d'une enquête

4. Instruments de collecte

1. Où avez-vous passé la nuit de mardi dernier le 11 octobre?

a. HÉBERGEMENT POUR NOUVEAUX IMMIGRANTS, RÉFUGIÉS OU DEMANDEURS D'ASILE [FIN]
b. MAISON DE TRANSITION CORRECTIONNELLE FÉDÉRALE (après prison fédérale, 2 ans et +)

c. CHEZ QUELQU'UN

d. HÔPITAL

e. CENTRE DE THÉRAPIE/DÉSINTOXICATION

f. PRISON PROVINCIALE OU POSTE DE POLICE

HÔTEL/MOTEL

g. Payé soi-même

h. Payé par un organisme ou un programme d'aide

i. MAISON DE CHAMBRES

j. REFUGE D'URGENCE POUR PERSONNES EN SITUATION D'ITINÉRANCE

k. HALTE-CHALEUR

LIEUX EXTÉRIEURS

l. SANS CAMPEMENT (ex. : rue, parc, abribus, forêt, immeuble abandonné [*squat*], viaduc)

m. AVEC CAMPEMENT (ex. : tente, abri, cabane ou autre installation extérieure)

n. VÉHICULE (ex. : voiture, fourgonnette, véhicule récréatif [VR], camion, bateau)

1.1 Si « oui » à « l », « m » ou « n » :

Étiez-vous dans un lieu caché (une cachette) durant cette nuit-là ? a. OUI b. NON c. REFUS

MAISON D'HÉBERGEMENT POUR FEMMES VICTIMES DE VIOLENCE

o. 1^{re} étape (3 mois ou moins)

p. 2^e étape (plus de 3 mois)

q. Ne sait pas



Principales étapes d'une enquête

5. Modes de collecte

Comment allons-nous obtenir les informations?

En lien étroit avec la base de sondage

- Entrevue en face-à-face
- Entrevue téléphonique
- Questionnaire auto-administré (sur papier)
- Questionnaire informatisé : Panel Web
- Prise de mesures cliniques : pèse-personnes, prise de sang, prise de tension artérielle, ...

➔ Mixte: mélange des modes précédents



Principales étapes d'une enquête

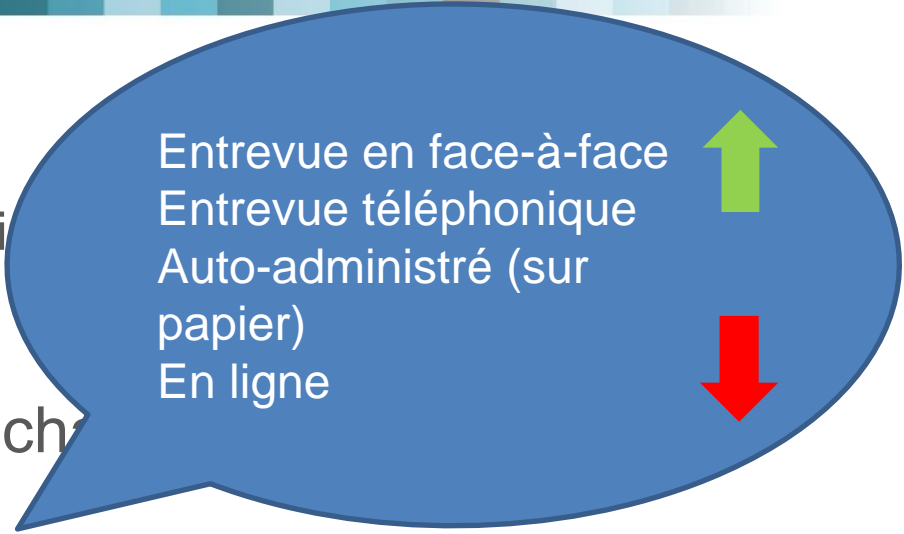
6. Saisie et traitement

- Validation des réponses

- Au moment de la collecte : saisir les données
- Éviter les incohérences

- Mesures de la qualité de l'échantillon

- Taux de non-réponse globale
 - le participant refuse de participer à l'étude
- Taux de non-réponse partielle
 - le participant accepte de participer à l'étude, mais refuse de répondre à certaines questions



Entrevue en face-à-face
Entrevue téléphonique
Auto-administré (sur papier)
En ligne

↑
↓

Principales étapes d'une enquête

6. Saisie et traitement

■ Pondération

- Redresser l'échantillon pour pouvoir inférer les résultats à la population-cible
 - Représentativité
 - L'échantillon final obtenu des répondants peut être loin de la réalité de la population-cible même dans le cas d'un échantillonnage aléatoire simple.
 - Poids d'enquête : déterminer le nombre de personnes que représente chacun des répondants.
 - Tenir compte des probabilités inégales dues en autres au sur-échantillonnage de certains sous-groupes de la population ou de la participation inégale de ceux-ci à l'enquête.
 - Allocation non proportionnelle de l'échantillon à travers les strates
- ➔ Si l'échantillon comprend 500 personnes de chacune des régions de la Côte-Nord et de Montréal, chaque répondant de Montréal devrait représenter alors plus de personnes que ceux de la Côte-Nord.

Principales étapes d'une enquête

6. Saisie et traitement

■ Pondération

- Redresser l'échantillon pour pouvoir inférer les résultats à la population-cible
- Représentativité
 - L'échantillon final obtenu des répondants peut être loin de la réalité de la population-cible même dans le cas d'un échantillonnage aléatoire simple.

Région	Échantillon	Population	Pondération de base
Côte-Nord	500	71 000	142
Montréal	500	1 700 000	3400

→ Si l'échantillon comprend 500 personnes de chacune des régions de la Côte-Nord et de Montréal, chaque répondant de Montréal devrait représenter alors plus de personnes que ceux de la Côte-Nord.

Principales étapes d'une enquête

6. Saisie et traitement

- Principales étapes de pondération
 - Poids de base: l'inverse de la probabilité de sélection à chaque degré du plan d'échantillonnage
 - Tenir compte de la non-réponse à chaque degré
 - Au final, poststratification sur des variables connues de la population (sexe, groupe d'âge, scolarité); appelée aussi étape du calage aux marges.



Principales étapes d'une enquête

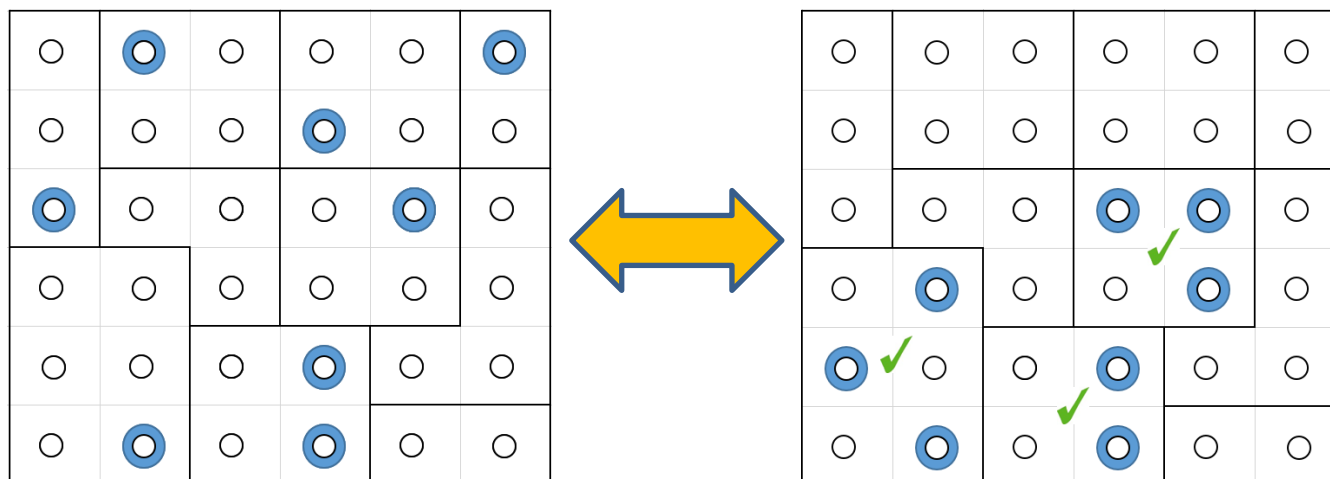
7. Analyses

- Estimations pondérées : proportions, populations estimées. moyennes
 - Réduire le biais global et potentiel induit par le fait de travailler sur un échantillon
- Mesures de précision des estimations
 - Déduites d'un échantillon et non de la population : estimations entachées d'une erreur échantillonnale (appelée aussi erreur-type; en anglais : standard error)
 - L'erreur-type diminue lorsque la taille d'échantillon augmente

Principales étapes d'une enquête

7. Analyses

- Mesures de précision des estimations
 - Plan d'échantillonnage complexe a un impact sur le calcul des mesures de précision : unités d'analyse non indépendantes (grappes ou clusters).



Aléatoire : indépendant
Bonne variabilité

Complexe : dépendant
Variabilité sous-estimée

Institut national
de santé publique

Québec

Principales étapes d'une enquête

7. Analyses

- Mesures de précision des estimations
 - Un effet de plan de 2 sur un échantillon de taille X signifie que la variance du plan complexe est 2 fois plus grande que celle obtenue d'un EAS de même taille.
 - En d'autres mots, il faut une taille d'échantillon 2 fois plus grande pour obtenir une variance similaire à un EAS
 - Estimer l'effet de plan par l'utilisation de méthodes de calcul appropriées :
 - Méthode de rééchantillonnage : Bootstrap, Jackknife
 - Linéarisation de Taylor par des logiciels spécialisés: SUDAAN
 - Corriger les poids d'enquête en divisant par l'effet de plan

Principales étapes d'une enquête

7. Analyses

- Mesures de précision des estimations
 - Accompagner chaque estimation d'une mesure de précision : coefficient de variation (CV), intervalle de confiance, variance.

$$\text{CV en \%} = \frac{\text{Erreur - type}}{\text{Estimation}} \times 100$$

Critères de diffusion de Statistique Canada:

- $16,6\% \leq \text{CV} < 33,3\%$: estimation présentant une forte variabilité; à interpréter avec circonspection;
- $\text{CV} \geq 33,3\%$: estimation imprécise; non diffusée.
- Les erreurs-type sont aussi utiles pour les tests de comparaison

Principales étapes d'une enquête

Illustration à l'aide d'un exemple

- Depuis début 2022, il est difficile de connaître le nombre de nouveaux cas hebdomadaires de COVID-19, car on ne dépiste plus tout le monde (personnel de la santé + population vulnérable).
- Afin d'avoir une estimation du nombre de nouveaux cas hebdomadaires (incidence), nous avons lancé un sondage, dont les résultats sont disponibles ici : <https://www.inspq.qc.ca/covid-19/donnees/estimation-sondage>
- Penchons-nous à nouveau sur les principales d'une enquête.

Principales étapes d'une enquête

Illustration à l'aide d'un exemple

Étape 1 : Définir la population-cible

- Ensemble des québécois âgés de 18 ans et plus.

Étape 2 : Base de sondage

- Utilisation de la base de données Clic Santé, qui sert à prendre des rendez-vous pour la vaccination.
- Cette base contient des adresse courriels.

Principales étapes d'une enquête

Illustration à l'aide d'un exemple

Étape 3 : Plan d'échantillonnage

- Échantillonnage probabiliste : systématique.
- Après la première semaine, on a remarqué que les taux de réponse étaient plus faibles chez les jeunes (18-34 ans) et chez les hommes.
- Solution : Sur-échantillonnage des jeunes et des hommes.
- Taille d'échantillon : 30 000 adultes québécois échantillonnés à chaque semaine.
- Nouveaux échantillons à chaque semaine => étude transversale.

Principales étapes d'une enquête

Illustration à l'aide d'un exemple

Étape 4 : Instrument de collecte

- Questionnaire.
- Voici la question principale permettant de faire le suivi des cas de COVID-19 :

Dans les 7 derniers jours incluant aujourd'hui, avez-vous eu un résultat de test positif à la COVID-19?

Oui

Non

Pas de résultat de test, mais je crois que je l'ai contractée (autodiagnostic en fonction de mes symptômes des 7 derniers jours)

Institut national
de santé publique

Québec 

Principales étapes d'une enquête

Illustration à l'aide d'un exemple

Étape 5 : Modes de collecte

Jour de la semaine	Nombre de questionnaires complétés
Lundi (courriel d'invitation)	2856
Mardi	661
Mercredi	186
Jeudi (courriel de rappel)	1318
Vendredi	292
Samedi	104
Dimanche	94
Total	5511

Principales étapes d'une enquête

Illustration à l'aide d'un exemple

Étape 6 : Saisie et traitement

- VOXCO permet directement d'exporter les réponses en fichiers Excel et SAS.
- Le taux de réponse global est d'environ de 18 % : 5 500 sondages complétés / 30 000 courriels envoyés.
- Pondération
 - Tient compte de la probabilité initiale de sélection
 - Tient compte du sur-échantillonnage et de la non réponse
 - Post stratification faite sur les variables sociodémographiques suivantes : genre, âge, région, scolarité, fait de vivre seul.

Principales étapes d'une enquête

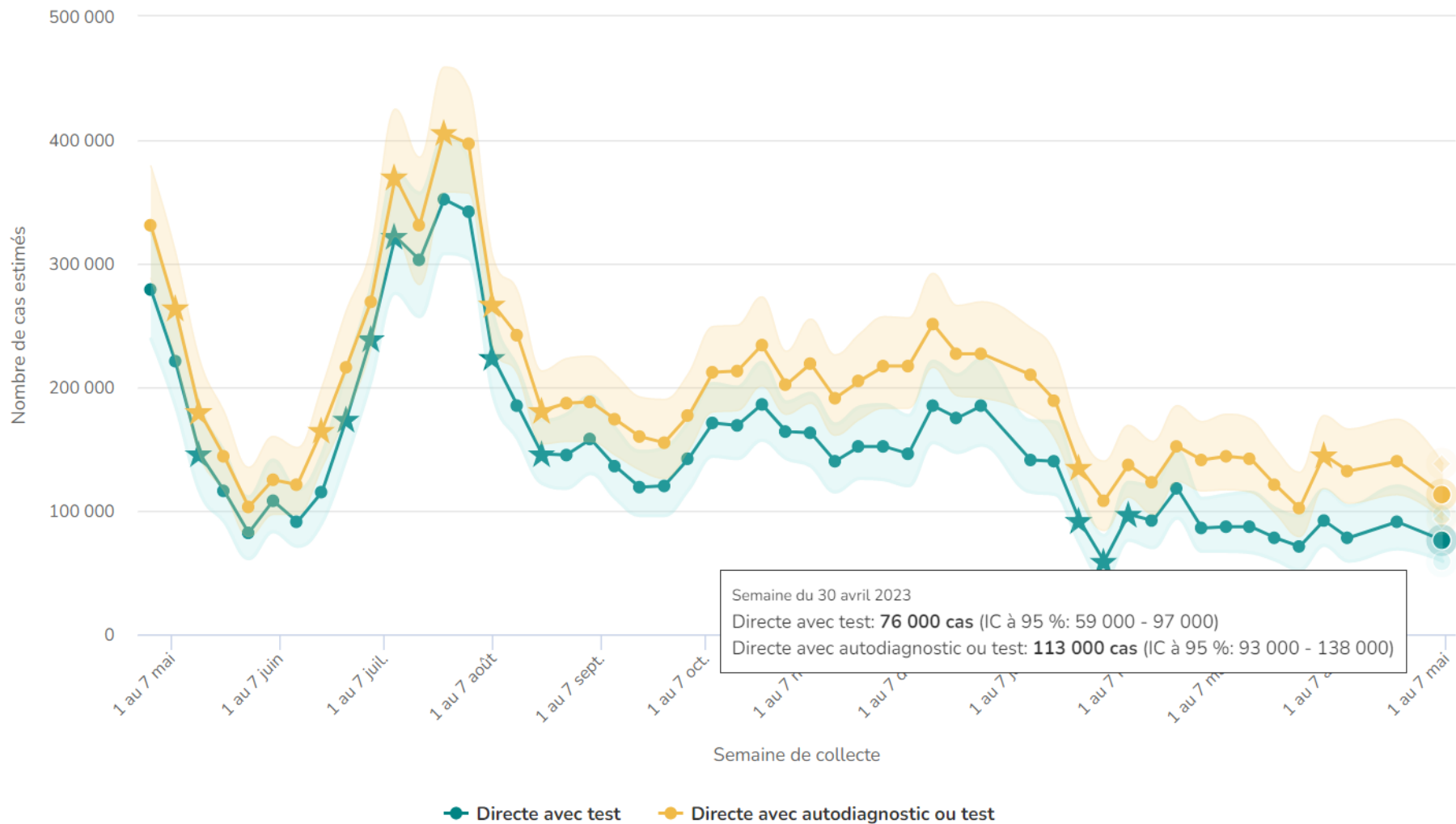
Illustration à l'aide d'un exemple

Étape 7 : Analyses

- Calcul d'estimations pondérées
- Estimation de l'erreur-type par rééchantillonnage (bootstrap)
- Calcul des CV et non-diffusion si $CV \geq 33,3\%$
- Calcul d'intervalles de confiance (IC)
- Tests de comparaison pour suivre l'évolution du nombre de cas.

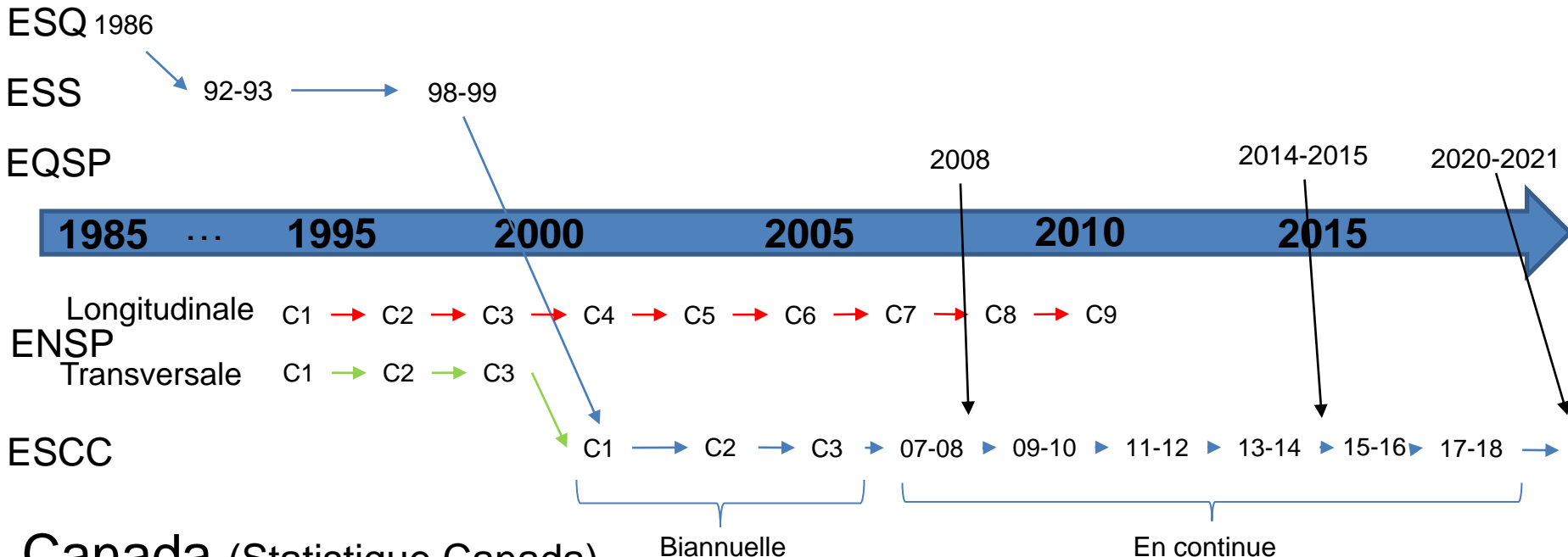
Principales étapes d'une enquête

Illustration à l'aide d'un exemple



Principales enquêtes de santé au Québec et au Canada

Québec (Institut de la statistique du Québec)



Canada (Statistique Canada)

ESQ: Enquête Santé Québec
 ESS: Enquête sociale et de santé
 EQSP: Enquête québécoise sur la santé de la population

ENSP: Enquête nationale sur la santé de la population
 ESCC: Enquête sur la santé dans les collectivités
 canadienne

Institut national
 de santé publique
Québec 

Principales enquêtes de santé

Générales au **Canada**

- Enquête nationale sur la santé de la population (ENSP)
 - Population-cible : personnes de 12 ans et plus vivant en ménages privés dans les provinces et territoires
 - Volet transversale: 1994-1995, 1996-1997 et 1998-1999 ensuite remplacé par l'ESCC
 - Volet longitudinale: de 1994-1995 suivi par 8 vagues : fréquence aux deux ans
 - Mode de collecte: entrevues téléphoniques
 - À portée provinciale uniquement
 - Ancêtre de l'ESCC

Principales enquêtes de santé

Générales au **Canada**

- Enquête sur la santé dans les collectivités canadiennes (ESCC)
 - Population-cible : personnes de 12 ans et plus vivant en ménages privés dans les provinces et territoires
 - Biannuelle pour les trois premiers cycles; en continue à partir de 2007
 - À portée régionale (socio-sanitaires) : données sur deux années de collecte
 - Mode de collecte mixte : entrevues en face-à-face et entrevues téléphoniques et maintenant en ligne
 - Plusieurs thématiques de santé abordées.
 - Contenus thématiques (ex. : cycle 2023 – télétravail).

*Institut national
de santé publique*

Québec 

Principales enquêtes de santé

Générales au Québec

- Enquête Santé Québec (ESQ, 1987), Enquête sociale et de santé (ESS, 1992-1993 et 1998)
 - Premières enquêtes d'envergure sur la santé des québécois
 - Population-cible : québécois de 15 ans et plus vivant en ménages privés
 - Modes de collecte : entrevue en face-à-face (ménage) et questionnaire-papier traitant de sujets sensibles (individus)
 - À portée régionale

Principales enquêtes de santé

Générales au Québec

- Enquête québécoise sur la santé de la population (EQSP, 2008, 2014-2015, 2020-2021)
 - Population-cible : québécois de 15 ans et plus vivant en ménages privés
 - Modes de collecte : entrevues téléphoniques en 2008, mixte en 2014-2015 (entrevues téléphoniques et questionnaire rempli en ligne), principalement en ligne en 2020-2021
 - À portée régionale en 2008 (possibilité de sur-échantillonner pour le local), locale (RLS) en 2014-2015.
 - Contenu : complément aux ESCC : Documenter des indicateurs prévus au plan commun de surveillance
 - Drogues, santé sexuelle, activités physique (2014-2015), santé au travail, blessures, suicides, nutrition, ...

*Institut national
de santé publique*

Québec 

Principales enquêtes de santé

Thématiques au **Canada**

- Enquête sur la santé dans les collectivités canadiennes (ESCC)
 - Volet Santé mentale (2002, 2012)
 - Volet Nutrition (2005, 2015)
 - Vieillesse en santé (2008-2009)
- Enquête de surveillance de l'usage du tabac au Canada (ESUTC)

Principales enquêtes de santé

Thématiques au Québec (ISQ)

- Enquête de santé chez les Cris (1992) ;
- Enquête québécoise sur les activités physiques et de loisir (1998) ;
- Enquête de santé chez les communautés culturelles (1998) ;
- Enquête longitudinale du développement des enfants du Québec (1998-2015) ;
- Enquête auprès de la clientèle des ressources pour personnes itinérantes des régions de Québec et de Montréal (1998-1999) ;
- Enquête sociale et de santé auprès des enfants et des adolescents québécois – Volet Nutrition ;
- Enquête sur le tabac, l'alcool, la drogue et le jeu chez les élèves du secondaire (ETADJES) 2008 et 2013 ;
- Enquête québécoise sur le développement des enfants à la maternelle 2012 (EQDEM) ;
- Enquête québécoise sur la santé des jeunes du secondaire 2010-2011 (EQSJS): nouvelle édition: 2016-2017.

**Institut national
de santé publique**

Québec 

Principales enquêtes de santé

Thématiques au Québec (INSPQ)

- Enquête de santé chez les Inuits en 2004 et en 2017;
- Étude des blessures subies au cours de la pratique d'activités récréatives et sportives (ÉBARS) : 2004, 2009 et 2015-2016;
- Étude clinique de santé buccodentaire du Québec (2012-2013): échantillon d'élèves de 2^{ème} et de 6^{ème} années du primaire;
- Dénombrement des personnes en situation d'itinérance (2022).

Avantages et inconvénients des enquêtes

Avantages

- Couvrent des dimensions de la santé, des caractéristiques des populations, de leurs habitudes de vie et comportements autrement impossibles à mesurer.
- Permettent de cibler certaines populations (jeunes du secondaire, de la maternelle, etc.).
- Permettent des croisements intéressants entre des problématiques et certains déterminants potentiels, facilitant l'identification de facteurs de risques et de populations plus vulnérables.
- Permettent parfois des analyses longitudinales poussées apportant une compréhension nouvelle de l'influence variable de certains déterminants lors d'étapes charnières de la vie.

Inconvénients

- Problèmes de comparabilité entre les cycles d'enquêtes : changements apportés au plan d'échantillonnage et au mode de collecte, à l'univers de la population, à l'ordre et au libellé des questions, etc.
- Faible puissance statistique dans de nombreuses enquêtes ne permettant pas le calcul d'estimations infrarégionales ou le croisement de variables.
- Les données sont sauf exception auto-rapportées, entraînant ainsi des biais (ex. de l'IMC).
- Taux de réponse souvent faibles, pouvant entraîner d'autres biais de représentativité
- Coût souvent élevé, surtout lorsqu'un suréchantillonnage est demandé.

Défis : jumeler les sources de données en surveillance

- Conjuguer « enquête » et « données administratives »
 - Puiser le meilleur de ces deux mondes en tentant de jumeler ces deux sources de données
 - Exemple : Indice de défavorisation (abordé plus loin)
 - Exemple : Dénombrement des personnes en situation d'itinérance (abordé plus loin)

Standardisation des estimations

www.inspq.qc.ca

Retour sur la mise en situation

À la surprise d'un chercheur, à l'inverse de la littérature, la proportion de MPOC chez les habitants d'une réserve indienne québécoise (7,8%) est inférieure à celle de l'agglomération voisine (9,6%), dont la population est majoritairement caucasienne:

- Qu'est-ce qui pourrait expliquer ce résultat surprenant?
- Les deux populations sont-elles vraiment comparables ?
- Si non, comment arrive-t-on à les comparer ?

Standardisation



- Lorsqu'on veut comparer des territoires, ou encore comparer un même territoire à différents moments dans le temps, il faut tenir compte de la structure d'âge potentiellement différente des populations comparées.
- C'est d'autant plus crucial lorsque la caractéristique (la maladie) à l'étude varie selon l'âge.
- La probabilité de développer une MPOC augmente justement avec l'âge.

Standardisation

- Une estimation (proportion, taux, moyenne) brute reflète la situation telle qu'observée dans la population. C'est le fardeau réel.
- Le calcul d'une estimation brute est simple: un numérateur divisé par un dénominateur.
- Pour un taux brut (comme le taux d'incidence):

$$Taux_{brut} = \frac{\text{Nombre de (nouveaux) cas}}{\text{Nombre de sujets à risque sur la période}}$$

Standardisation directe

Cette méthode applique les taux spécifiques (par groupe d'âge) **de la population à l'étude** à une population de référence. C'est une moyenne pondérée des taux spécifiques, les poids étant dérivés de la population de référence:

$$Taux_{ajusté} = \sum_{i=1}^p W_i \times Taux_i$$

où le nombre de strates d'âge i va de 1 à p ,

où W_i est la proportion de la population de référence dans chaque groupe d'âge,

et où $Taux_i$ est le taux du groupe d'âge i dans la population à l'étude.

Standardisation indirecte

Un peu à l'inverse, cette méthode applique les taux spécifiques (par groupe d'âge) **de la population de référence** à la structure d'âge des populations à l'étude.

La méthode consiste à comparer le nombre de cas **observés** au nombre de cas **attendus** selon l'hypothèse voulant que la population à l'étude ait les mêmes taux spécifiques que la population de référence.

Le ratio du nombre de cas observés sur le nombre de cas attendus s'interprète comme un excédant ou un déficit dans la population à l'étude.

Standardisation indirecte

Autrement dit:

Si la population à l'étude avait les mêmes taux qu'une grande population de référence, combien de cas devrait-on s'attendre à obtenir ? Et est-ce que ce nombre est plus faible ou plus élevé que le nombre total de cas observé dans la population à l'étude ?

Standardisation indirecte

$$Attendus = \sum_{i=1}^p n_i \times Taux_{i,réf}$$

où le nombre de strates d'âge i va de 1 à p ,

où n_i est le nombre d'individus de la population à l'étude par groupe d'âge

et où $Taux_{i,réf}$ est le taux du groupe d'âge i dans la

population de référence

Cette méthode ne mène pas directement à une mesure de taux ajusté mais plutôt à un ratio, l'indice comparatif d'incidence, ICI (mieux en anglais sous le vocable de « standardized incidence rate » (SIR))

$$ICI = \frac{Observés}{Attendus}$$

Avantages et inconvénients des deux méthodes

Standardisation directe

- +++ Rend valide les comparaisons de taux entre deux populations différentes.
- + Le calcul d'intervalles de confiance est relativement simple, de sorte qu'il est possible de tester des hypothèses.
- Méthode instable en présence d'un événement ou d'une maladie rare, ou encore lorsqu'une population à l'étude est très petite.
- Nécessite les taux spécifiques (par groupe d'âge) dans la population à l'étude
- La valeur standardisée est une mesure « synthétique » qui n'est utile qu'à des fins de comparaisons.

Standardisation indirecte

- + Il n'est pas nécessaire de connaître les taux spécifiques (par groupe d'âge) dans la population à l'étude.
- + La variabilité est moins grande qu'avec la standardisation directe (la précision est meilleure)
- + L'interprétation est simple et intuitive
- Ne permet pas de comparer directement deux populations, parce que le système de poids (les n_i) est différent pour chacune.

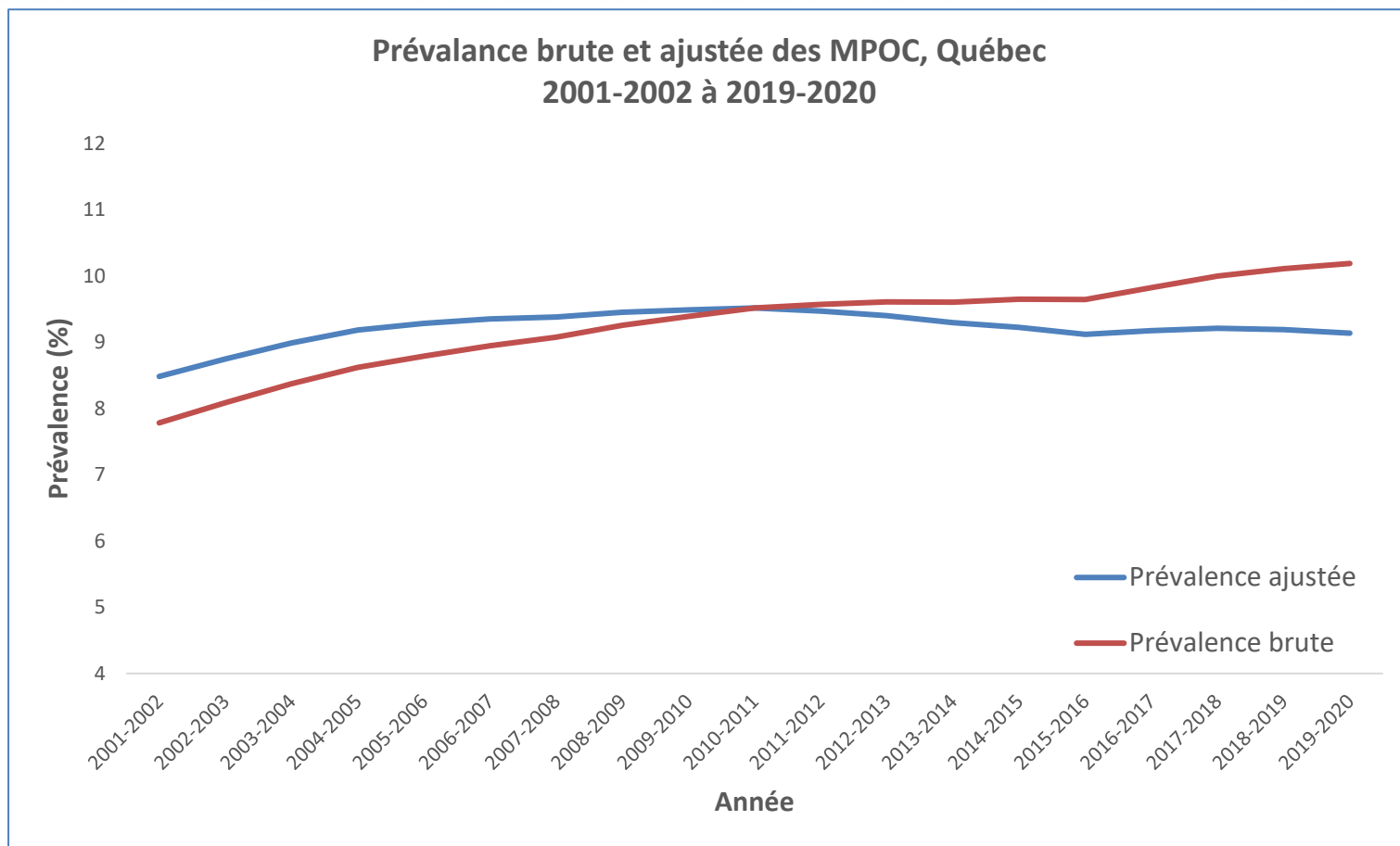
Choix de la standardisation

- En surveillance de la population, la variable d'âge est presque toujours présente, de sorte que les taux spécifiques peuvent être calculés.
- À une échelle populationnelle, les problèmes de précision de la standardisation directe sont fortement atténués.
- Pouvoir comparer directement des populations est très important en surveillance.
- La standardisation directe est donc à privilégier.
- La standardisation indirecte demeure utile, notamment pour calculer l'excès de mortalité des malades par rapport aux non-malades.

Standardisation: autres considérations

- La population de référence doit idéalement être assez grande (exemple: Québec, Canada).
- La population de référence doit être proche dans le temps de la population à l'étude. Une année de recensement est préférable.
- Le choix des groupes d'âge dépend de la maladie à l'étude. En général, plus la maladie est rare (plus le nombre de cas est faible), plus les groupes d'âge sont larges. Vice versa.
- L'estimation brute correspond au fardeau réel dans la population. L'estimation ajustée est une mesure synthétique utile principalement pour comparer.

Retour sur la mise en situation



Tests statistiques et intervalles de confiance

www.inspq.qc.ca

Principes de base

■ Tests d'hypothèse

- Un test statistique permet de vérifier s'il existe une différence significative sur le plan statistique entre des mesures estimées dans une population (ex. : taux, moyennes, proportions).
- Supposer une hypothèse nulle (H_0) souvent vue comme une absence d'effet et observer des données pour vérifier si cette hypothèse peut être fondée.
- Souvent en lien avec une hypothèse alternative (H_1)

Ex: H_0 : prop. fumeurs Femmes = prop. fumeurs Hommes

H_1 : prop. fumeurs Femmes \neq prop. fumeurs Hommes

Principes de base

Tests d'hypothèse

	Vérité : H_1 vraie	Vérité : H_0 vraie
Jugement statistique : rejet de H_0	Aucune erreur: Puissance du test ($1-\beta$)	Erreur de type I : rejet de H_0 alors que H_0 est vraie (erreur α)
Jugement statistique : non-rejet de H_0	Erreur de type II : non-rejet de H_0 alors que H_1 est vraie (erreur β)	Aucune erreur: Jugement concordant avec la vérité

Principes de base

Tests d'hypothèse

- Cas extrême 1 : on utilise un test statistique qui conclut toujours au non-rejet de H_0
 - Aucune erreur de type I, mais +++ erreurs de type II.
- Cas extrême 2 : on utilise un test statistique qui conclut toujours au rejet de H_0
 - +++ erreurs de type I, mais aucune erreur de type II.

Un bon test statistique se situe entre ces deux extrêmes. Il est construit de sorte que les erreurs de type I ET de type II sont minimisées.

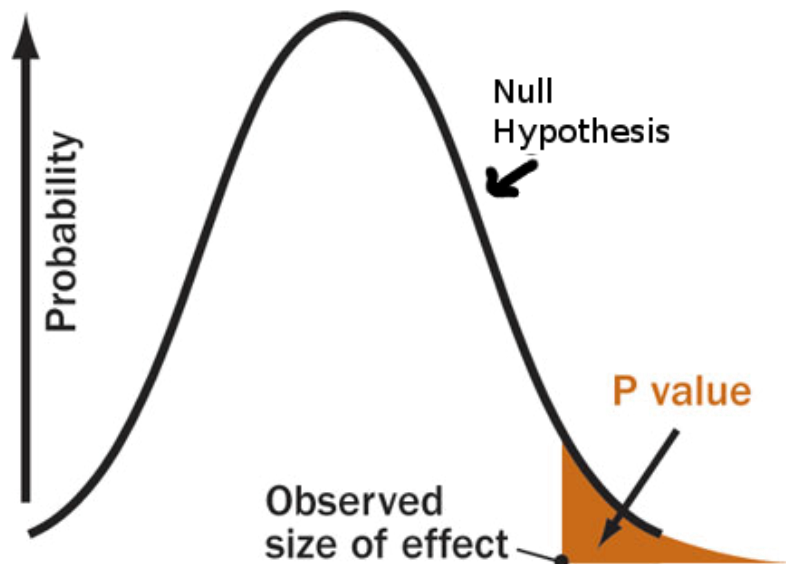
Principes de base

Tests d'hypothèse

- Concept de la valeur p (p-value)

- Définition générale

Probabilité d'observer un résultat aussi extrême que celui observé dans l'échantillon sachant que l'hypothèse nulle est vraie.



Pour le résultat du test, la valeur p est ensuite comparée à une valeur seuil α : le % maximal de l'erreur de type I que nous sommes prêts à commettre.

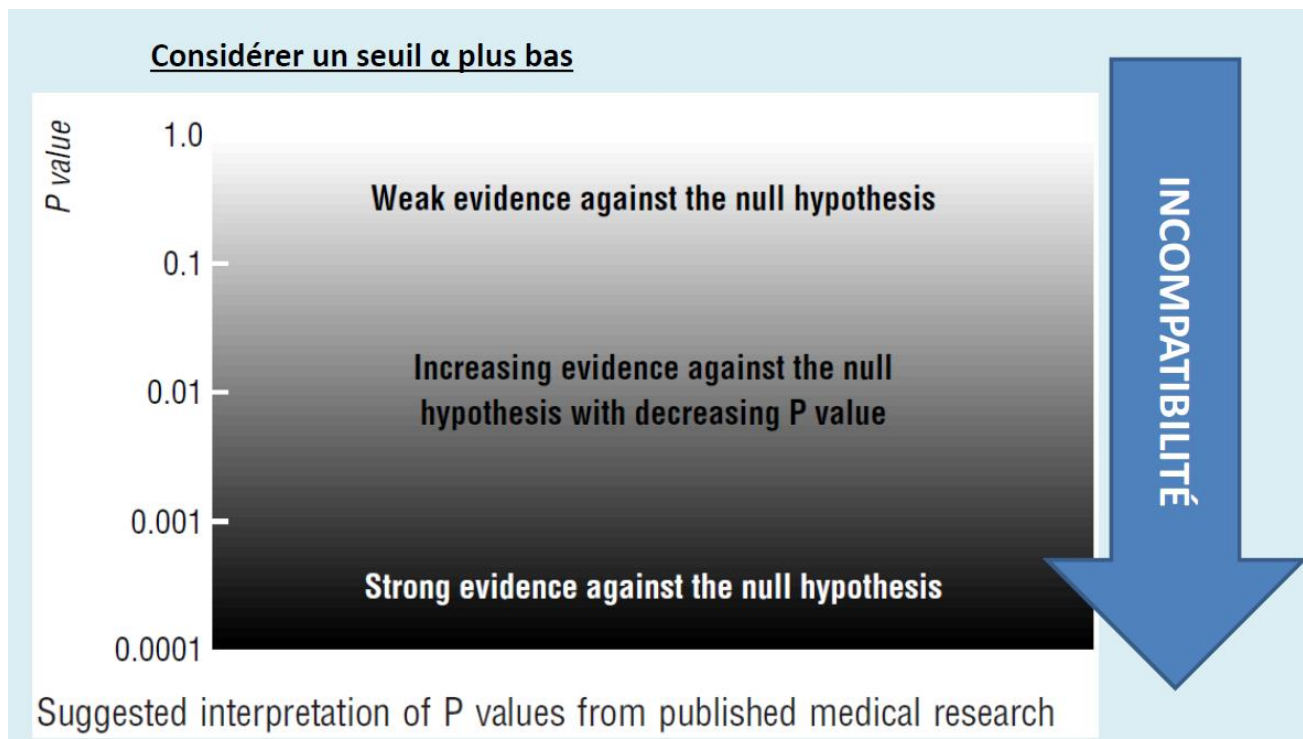
Principes de base

Tests d'hypothèse

- Concept de la valeur p (p-value)

- Autre définition générale

Mesure continue de la compatibilité entre les données observées et l'hypothèse nulle au niveau populationnel



Principes de base

Tests d'hypothèse

■ Mauvaise interprétation

Article de l'ASA: The ASA's Statement on p-Values: Reporting, Interpreting, and Communicating the Results of Hypothesis Testing: Principles and Guidelines for Practitioners

\neq probabilité que l'hypothèse soit vraie ou bien (1 - probabilité) que l'hypothèse soit fautive

Une valeur-p ne devrait pas être présentée seule.

Elle devrait être accompagnée de graphiques, de statistiques descriptives, d'intervalles de confiance, etc.

Une petite valeur p devrait plutôt être interprétée comme

- Les résultats du test fournissent de l'évidence contre l'hypothèse nulle
- Les résultats du test pourraient suggérer qu'il y a un effet

par hasard

résultat



Principes de base

Groupes de comparaison

- Validité des tests statistiques

- Si les groupes ou sous-populations à comparer sont indépendants ou mutuellement exclusifs.

- Dans la plupart des cas, cette règle est respectée.

Ex. Taux d'incidence des MPOC

- Entre les hommes et les femmes
- Entre les niveaux de scolarité

- Si le phénomène étudié varie selon l'âge et que la distribution de l'âge est différente dans les deux groupes à comparer :

- Le test devrait être fait sur les mesures ajustées (standardisation)

Principes de base

Groupes de comparaison

Deux cas particuliers

1. Entités géographiques de type sous-région vs l'ensemble de la région

Ex. Le taux d'incidence de la région de Montréal est-il différent de l'ensemble du Québec ?

- La sous-région fait également partie de l'ensemble de la région → les groupes à tester ne sont pas indépendants
- Problème plus important si la sous-région représente une part relativement importante de la région

Solution : faire des comparaisons de la sous-région vs. l'ensemble de la région sans cette sous-région

Ex. Le taux d'incidence de la région de Montréal est-il différent du **reste du Québec** ?

Principes de base

Groupes de comparaison

2. Pour les données d'enquêtes, groupes à comparer également indépendants selon le plan de sondage

Pour l'ESCC, c'est le cas pour les régions → variables de stratification

Ce n'est pas évident pour toutes les autres variables de croisement, notamment celles portant sur les caractéristiques des individus.

Solution : il faut tenir compte du degré de dépendance (covariance) dans la statistique du test

$$\begin{aligned} \text{Var}(p_1 - p_2) &= \text{Var}(p_1) + \text{Var}(p_2) - \text{Covar}(p_1, p_2) \\ &\neq \text{Var}(p_1) + \text{Var}(p_2) \end{aligned}$$

Il n'existe pas de formule simple pour estimer cette covariance

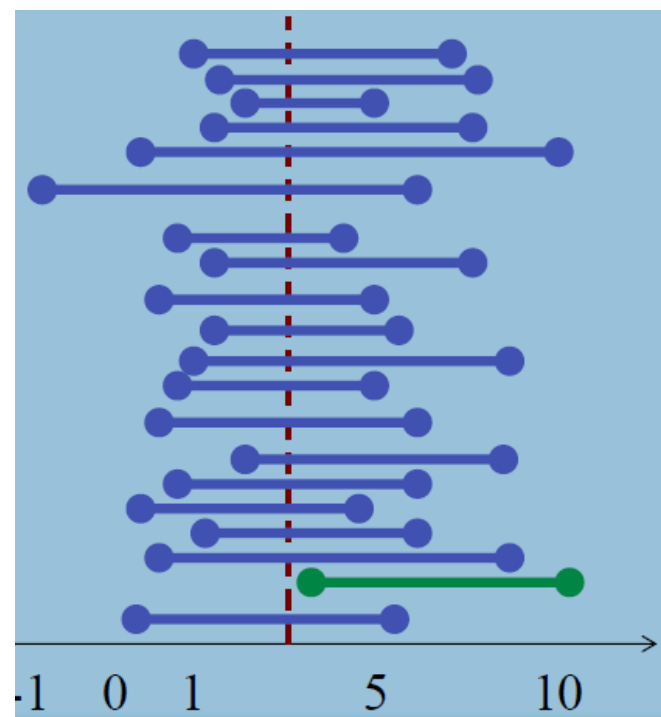
→ Estimer directement $\text{Var}(p_1 - p_2)$ par la méthode de rééchantillonnage bootstrap

Principes de base

Intervalles de confiance (IC)

■ Définition générale

- Si la procédure pour obtenir un intervalle de confiance de niveau 95% est répétée un grand nombre de fois, 95% des intervalles ainsi construits contiendraient la vraie valeur populationnelle.
- Un IC contient ou non la vraie valeur, mais on ne peut pas affirmer qu'il y a 95 % de chances de contenir cette valeur.
- Dans l'exemple, on remarque que 19 IC sur 20 contiennent la vraie valeur populationnelle, soit $19/20 = 95\%$



Principes de base

Intervalles de confiance

■ Autre définition

- Un intervalle de confiance est défini comme un ensemble de valeurs d'intérêt qui sont comprises dans un intervalle de valeurs possibles.
- Un IC à $(1-\alpha)\%$ contient toutes les valeurs possibles telles que la valeur p du test d'hypothèse H_0 vs. H_1 est $> \alpha$.

En plus de nous renseigner sur les résultats d'un test d'hypothèse, l'IC nous indique la **force** et le **sens** d'un effet d'intérêt.

Ce que ne permet pas la valeur-p.

Ex: Le test sur la différence des proportions de fumeurs entre les hommes et les femmes donne une valeur de 4,2% avec un IC à 95% de [2,1%-6,8%]

$$H_0: D = p_H - p_F = 0 \quad H_1: D = p_H - p_F \neq 0$$

Comme la valeur 0 n'est pas incluse dans cet intervalle, la différence de 4,2% est statistiquement significative ($\neq 0$).

Constructions de statistiques de test et d'intervalles de confiance

- Pour les proportions (données d'enquêtes)

Test du type $H_0 : p_1 = p_2$ vs $H_1 : p_1 \neq p_2$

Les proportions suivent une loi binomiale, mais il est difficile de construire une statistique de test à partir de cette loi.

Solution simple : Utiliser directement l'approximation normale.

Contrainte : Si $Np > 5$ et $N(1-p) > 5$, donc p et n assez grands

Cette solution est basée sur le Théorème Limite Central.

Constructions de statistiques de test et d'intervalles de confiance

- Pour les proportions (données d'enquêtes)

(Test du type $H_0 : p_1 = p_2$ vs $H_1 : p_1 \neq p_2$)

Si la contrainte n'est pas respectée (données d'enquête)

Solution alternative: Utiliser l'approximation normale sur la transformation logit (logarithme népérien sur la cote)

$$\text{logit} = \ln\left(\frac{p_i}{1 - p_i}\right)$$

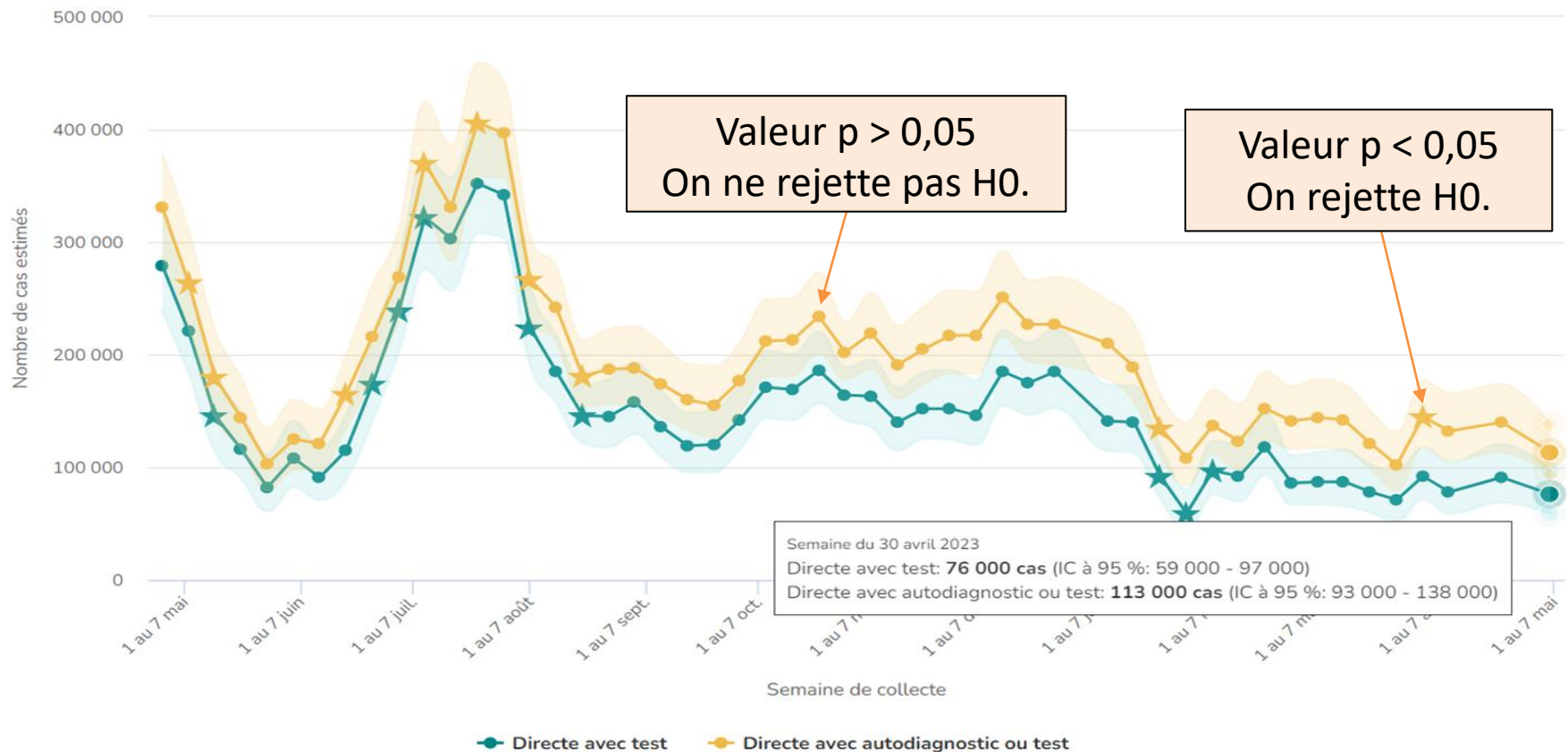
$$\text{Statistique de Wald } W = \frac{\left(\ln\left(\frac{p_1}{1-p_1}\right) - \ln\left(\frac{p_2}{1-p_2}\right)\right)^2}{\text{Var}\left(\ln\left(\frac{p_1}{1-p_1}\right) - \ln\left(\frac{p_2}{1-p_2}\right)\right)} \sim \text{Khi} - \text{deux à 1 d.l.}$$

C'est cette approche qui était utilisée dans la section sur les enquêtes pour comparer l'incidence hebdomadaire de COVID-19.

Constructions de statistiques de test et d'intervalles de confiance

p : proportion de québécois majeurs ayant eu la COVID-19 dans les 7 derniers jours.

$H_0 : p_{sem\ précédente} = p_{sem\ courante}$ vs $H_1 : p_{sem\ précédente} \neq p_{sem\ courante}$



Constructions de statistiques de test et d'intervalles de confiance

- Pour les proportions (données binomiales)
Intervalle de confiance de \hat{p}
 - Approximation normale: $\hat{p} \pm z_{1-\alpha/2} \sqrt{\hat{p}(1-\hat{p})}$
 - Intervalle binomial: très complexe, mais toujours centré.
 - Approximation normale sur la transformation logit.

C'est cette approche (approx. normale sur la transformation logit) qui est utilisée à la dispositive précédente pour construire les IC jaunes et verts.

$$I_l = \ln\left(\frac{\hat{p}}{1-\hat{p}}\right) - z_{1-\alpha/2} \sqrt{\text{Var}\left(\ln\left(\frac{\hat{p}}{1-\hat{p}}\right)\right)} ; U_l = \ln\left(\frac{\hat{p}}{1-\hat{p}}\right) + z_{1-\alpha/2} \sqrt{\text{Var}\left(\ln\left(\frac{\hat{p}}{1-\hat{p}}\right)\right)}$$

$$I_p = \frac{\exp(I_l)}{1 + \exp(I_l)} ; U_p = \frac{\exp(U_l)}{1 + \exp(U_l)}$$

Constructions ...

Pour les proportions (données d'enquêtes)

Exemple 1 : Proportion de personnes ayant fumé 30 paquets-année et plus chez les fumeurs quotidiens actuels et les anciens fumeurs quotidiens de moins de 15 ans âgés de 55 ans et plus ayant atteint le niveau universitaire, Québec, ESCC 2013-2014

Catégorie	p (%)	Selon bootstrap		CV	IC 95%	
		ET(p)	ET (logit(p))		Appr. Normale	Transf. Logit
Universitaire	12,0	0,0107	0,1016	8,9	9,9 - 14,1	10,1 - 14,3

Approximation normale sans transformation

$$\hat{p} \pm z_{1-\alpha/2} \sqrt{\text{Var}(\hat{p})} = 0,120 \pm 1,96 \times 0,0107 \rightarrow [9,9 - 14,1]$$

Approximation normale sur la transformation logit

$$I_l = \ln\left(\frac{0,120}{1-0,120}\right) - 1,96 \times 0,1016 = -2,192 ; I_p = \frac{\exp(-2,192)}{1+\exp(-2,192)} = 10,1$$

$$U_l = \ln\left(\frac{0,120}{1-0,120}\right) + 1,96 \times 0,1016 = -1,793 ; U_p = \frac{\exp(-1,793)}{1+\exp(-1,793)} = 14,3$$



Constructions ...

Pour les proportions (données d'enquêtes)

Exemple 2 : Test $H_0 : p_{\text{universitaire}} = p_{\text{Pas de DES}}$ vs $H_1 : p_{\text{universitaire}} \neq p_{\text{Pas de DES}}$

Validité du test

- Groupes à comparer indépendants ? **NON à cause du plan de sondage**
- Phénomène étudié et la scolarité varient-ils selon l'âge ? **OUI, alors test sur les proportions ajustées (standardisation)**

Catégorie	p (%)	Selon bootstrap		CV	IC 95%	
		ET(p)	ET (logit(p))		Appr. Normale	Transf. Logit
Universitaire	11,6	0,0108	0,1054	9,3	9,5 - 13,7	9,7 - 13,9
Pas de DES	18,7	0,0113	0,0744	6,0	16,5 - 21,0	16,6 - 21,0

Selon la méthode bootstrap, $Var \left(\ln \left(\frac{0,116}{1-0,116} \right) - \ln \left(\frac{0,187}{1-0,187} \right) \right) = 0,0157$

$$\text{Statistique de Wald } W = \frac{\left(\ln \left(\frac{0,116}{1-0,116} \right) - \ln \left(\frac{0,187}{1-0,187} \right) \right)^2}{0,0157} = 20,091$$

Valeur-p < 0,00001

Constructions ...

Pour les proportions (données d'enquêtes)

Remarque : si on n'avait ignoré le non-respect de l'indépendance, on aurait plutôt obtenu une variance de

$$\begin{aligned} & \text{Var} \left(\ln \left(\frac{0,116}{1-0,116} \right) - \ln \left(\frac{0,187}{1-0,187} \right) \right) \\ &= \text{Var} \left(\ln \left(\frac{0,116}{1-0,116} \right) \right) + \text{Var} \left(\ln \left(\frac{0,187}{1-0,187} \right) \right) \\ &= 0,1054^2 + 0,0744^2 = \mathbf{0,0166} \end{aligned}$$

Catégorie	p (%)	Selon bootstrap		CV	IC 95%	
		ET(p)	ET (logit(p))		Appr. Normale	Transf. Logit
Universitaire	11,6	0,0108	0,1054	9,3	9,5 - 13,7	9,7 - 13,9
Pas de DES	18,7	0,0113	0,0744	6,0	16,5 - 21,0	16,6 - 21,0

Selon la méthode bootstrap, $\text{Var} \left(\ln \left(\frac{0,116}{1-0,116} \right) - \ln \left(\frac{0,187}{1-0,187} \right) \right) = \mathbf{0,0157}$

$$\text{Statistique de Wald } W = \frac{\left(\ln \left(\frac{0,116}{1-0,116} \right) - \ln \left(\frac{0,187}{1-0,187} \right) \right)^2}{0,0157} = 20,091$$

Valeur-p < 0,00001

Constructions de statistiques de test et d'intervalles de confiance

- Pour la prévalence

Test du type $H_0 : p_1=p_2$ vs $H_1 : p_1 \neq p_2$

La prévalence suit approximativement une loi normale puisqu'elle est généralement calculée sur un grand nombre d'individus.

Alors la statistique $Z = \frac{|\hat{p}_1 - \hat{p}_2|}{\sqrt{\text{Var}(\hat{p}_1) + \text{Var}(\hat{p}_2)}} \sim N(0,1)$

Avec $\text{Var}(\hat{p}_i) = \frac{\hat{p}_i \times (1 - \hat{p}_i)}{N_i}$

Intervalle de confiance de niveau $1-\alpha$

$$\hat{p} \pm z_{1-\alpha/2} \sqrt{\text{Var}(\hat{p})}$$

$$\hat{p} \pm z_{1-\alpha/2} \sqrt{\text{Var}\left(\frac{\hat{p} \times (1 - \hat{p})}{N}\right)}$$

Constructions de statistiques de test et d'intervalles de confiance

- Pour les taux d'incidence ou de mortalité

Test du type $H_0 : tx_1 = tx_2$ vs $H_1 : tx_1 \neq tx_2$

Les taux sont généralement de faibles mesures et il faut utiliser la transformation logarithme pour qu'ils suivent approximativement une loi normale.

$$\text{La statistique } Z = \frac{|\ln(tx_1) - \ln(tx_2)|}{\sqrt{\text{Var}(\ln(tx_1)) + \text{Var}(\ln(tx_2))}} \sim N(0,1)$$

$$\text{Où } \text{Var}(\ln(tx_i)) = \frac{\text{Var}(tx_i)}{tx_i^2}$$

Constructions ...

Pour les taux d'incidence ou de mortalité

L'Intervalle de confiance du $\ln(tx)$ en tenant compte de l'approximation normale est

$$\ln(tx) \pm z_{1-\alpha/2} \sqrt{\text{Var}(\ln(tx))}$$

d'où l'intervalle pour tx est

$$e^{\ln(tx) \pm z_{1-\alpha/2} \sqrt{\text{Var}(\ln(tx))}}$$

Constructions ...

Pour les taux d'incidence ou de mortalité

Exemple 3 : Taux d'incidence de MPOC ajustés selon l'âge chez les personnes de 35 ans et plus, Région de Capitale-Nationale et de l'Estrie, 2014-2015.

Région	tx	Var(tx)	IC 95%	Reste du Québec			Test		
				tx	Var(tx)	IC 95%	Z	valeur-p	Sig.
Capitale-Nationale	6,271	0,0150	(6,035 - 6,516)	6,731	0,0016	(6,653 - 6,810)	3,465	0,0005	-
Estrie	6,635	0,0219	(6,351 - 6,932)	6,694	0,0016	(6,617 - 6,772)	0,385	0,7004	=

IC (Estrie):

$$\text{Var}(\ln(tx_i)) = \frac{\text{Var}(tx_i)}{tx_i^2} = \frac{0,0219}{6,635^2} = 0,0005$$
$$e^{\ln(6,635) \pm 1,96\sqrt{0,0005}}$$

Constructions ...

Pour les taux d'incidence ou de mortalité

Exemple 3 (suite) : Test $H_0 : tx_{\text{Estrie}} = tx_{\text{Reste Qc}}$ vs $H_1 : tx_{\text{Estrie}} \neq tx_{\text{Reste Qc}}$

Validité du test

- Groupes à comparer indépendants ? **OUI**
- Phénomène étudié et la région varient tous deux **indépendamment** selon l'âge (standardisation)

On peut ici sommer les deux variances pour estimer la variance de la différence, car les groupes à comparer sont indépendants.

Région	tx	Var(tx)	IC 95%
Capitale-Nationale	6,271	0,0150	(6,035 - 6,516)
Estrie	6,635	0,0219	(6,351 - 6,932)

$$Z = \frac{|\ln(tx_1) - \ln(tx_2)|}{\sqrt{\text{Var}(\ln(tx_1)) + \text{Var}(\ln(tx_2))}} = \frac{|\ln(6,635) - \ln(6,271)|}{\sqrt{0,0005 + 3,4819 \times 10^{-4}}} = 0,385$$

Si le seuil du test est 1%, $z_{1-\alpha/2} = 2,56$, alors vu que $Z = 0,385 < 2,56$, aucune évidence au rejet de H_0

Constructions de statistiques de test et d'intervalles de confiance

- Comparaison basée sur les intervalles de confiance
 - Principe ($H_0 : m_1 = m_2$ vs $H_1 : m_1 \neq m_2$)

Comparaison des intervalles de confiance de deux mesures

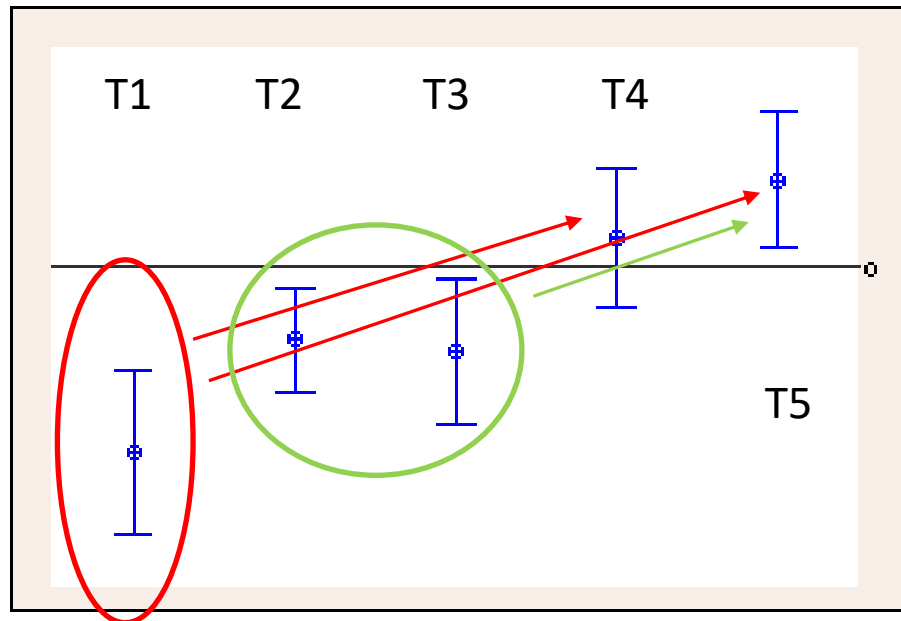
 - Si non chevauchement ou disjoints, tests significatifs
 - Si chevauchement, on ne peut rien conclure
 - Test très conservateur : rejet de l'hypothèse nulle seulement pour les fortes différences. Il y a des différences significatives au seuil α qui ne sont pas détectées...
 - Méthode simple, mais le prix à payer est la perte de puissance (erreurs type II)



Constructions ...

Comparaison basée sur les intervalles de confiance

Exemple:



T1 est statistiquement différent de T4 et T5

T2 et T3 sont statistiquement différents de T5

Pour les autres comparaisons, on ne peut rien conclure !!!

Constructions ...

Comparaison basée sur les intervalles de confiance

Retour à l'exemple 2 :

Test $H_0 : p_{\text{universitaire}} = p_{\text{Pas de DES}}$ vs $H_1 : p_{\text{universitaire}} \neq p_{\text{Pas de DES}}$

Catégorie	p (%)	Selon bootstrap		CV	IC 95%	
		ET(p)	ET (logit(p))		Appr. Normale	Transf. Logit
Universitaire	11,6	0,0108	0,1054	9,3	9,5 - 13,7	9,7 - 13,9
Pas de DES	18,7	0,0113	0,0744	6,0	16,5 - 21,0	16,6 - 21,0

Comme les IC [9,7 – 13,9] et [16,6 – 21,0] ne se chevauchent pas, on peut conclure au rejet de H_0 par le test de comparaison basé sur les intervalles de confiance.

Avec le test de Wald de la différence des deux proportions (valeur-p < 0,0001), l'évidence était également au rejet de l'hypothèse nulle que les deux proportions sont égales.

Problème de comparaisons multiples

Principe de base

Test global H_0 : toutes les mesures des régions sont égales

contre H_1 : il existe au moins une région différente des autres

La multiplicité des tests de comparaison deux à deux augmente la probabilité de trouver un test global significatif alors qu'il ne l'est pas.

Ex. à un seuil de 5%, on accepte de faire l'erreur de rejeter H_0 alors qu'elle est vraie 1 fois sur 20

→ Si 20 tests de ce genre sont faits, en moyenne, un de ces tests sera rejeté par erreur.

Prendre une décision en faisant plusieurs tests de comparaisons multiples à un seuil α ne garantit pas du tout que le test global est fait à ce même seuil. En fait, dans ce cas, $\alpha_{\text{global}} > \alpha$.

→ Ne contrôle pas le seuil du test global

Problème de comparaisons multiples

Test global

- Ajustement par la correction de Bonferroni

Une manière simple de contourner le problème est d'ajuster à la baisse le seuil des différents tests de comparaisons multiples

- La correction simple de Bonferroni est de diviser le seuil du test global par le nombre de comparaisons (soit r): $\alpha_{\text{mult}} = \alpha_{\text{global}}/r$

Point fort: S'assure de contrôler le seuil du test global

Point faible: méthode trop conservatrice → risque de passer à côté d'intéressantes différences significatives

- Il existe dans la littérature d'autres corrections moins conservatrices, mais ça demande plus de programmation.

Problème de comparaisons multiples

Test global

- Ajustement par la correction de Bonferroni

Test global H_0 : toutes les mesures des régions sont égales

contre H_1 : il existe au moins une région différente des autres

Cette procédure est souvent utilisée en deux étapes.

1. Faire les r tests statistiques au seuil de α_{global}/r

Si au moins une valeur- $p < \alpha_{\text{global}}/r$, on rejette H_0 du test global et on passe à l'étape 2.

Si toutes les valeurs- $p \geq \alpha_{\text{global}}/r$, rien ne va contre l'évidence de H_0 .

2. Poursuivre avec les r tests de comparaisons multiples à un seuil $> \alpha_{\text{global}}/r$ (généralement, α_{global} , donc sans correction de Bonferroni) sinon tests trop conservateurs.

Problème de comparaisons multiples

Test global

- Pour les proportions (données d'enquête): test du Khi-deux → Test global H_0 : toutes les proportions des régions sont égales contre H_1 : il existe au moins une région différente des autres

Test d'indépendance du khi-deux mesurant l'association entre une variable d'intérêt et une variable de croisement (ex., RSS).

Comme pour l'estimation de la variance, il faut tenir compte du plan de sondage (ajustement de Satterthwaite, méthode Rao-Scott de deuxième degré)

Si ce test est significatif au seuil α_{global} , on peut procéder ensuite aux comparaisons multiples au même seuil.

Problème de comparaisons multiples

Exemple 4 (suite)

Test	Domaine	Statistique	Deg. de liberté	Valeur du Chi-carré	Valeur p
1: Variables: paquet_annee_cat2 VS region	Région	Chi-2 ajusté (ajust. R.-S. du sec. degré)(*)	10,213	35,95	0,0001

Région	Prop.	Reste du Qc	Diff.	Z	Valeur-p	Sig
Bas-Saint-Laurent	11,1	14,7	-3,6	2,09	0,0366	(-)
Sag./Lac Saint-Jean	11,3	14,7	-3,4	1,96	0,0501	NS
Région de Québec	15,8	14,5	1,3	0,67	0,5031	NS
Mauricie-CDQ	15,1	14,5	0,5	0,26	0,7921	NS
Estrie	14,2	14,6	-0,4	0,19	0,8516	NS
Montréal-Centre	12,3	15,2	-2,9	2,06	0,0398	(-)
Outaouais	19,5	14,4	5,2	2,54	0,0110	(+)
Abitibi-Témiscamingue	13,4	14,6	-1,2	0,70	0,4823	NS
Côte-Nord	13,0	14,6	-1,6	0,83	0,4072	NS
Nord-du-Québec	14,4	14,6	-0,2	0,09	0,9247	NS
Gaspésie-Îles-Madeleine	14,7	14,6	0,1	0,06	0,9508	NS
Chaudières-Appalaches	12,5	14,7	-2,2	1,11	0,2656	NS
Laval	12,9	14,7	-1,8	0,93	0,3541	NS
Lanaudière	19,0	14,3	4,7	2,31	0,0211	(+)
Laurentides	18,6	14,3	4,3	1,97	0,0492	(+)
Montérégie	14,9	14,6	0,3	0,19	0,8482	NS



Association et modélisation statistique

www.inspq.qc.ca

Corrélation

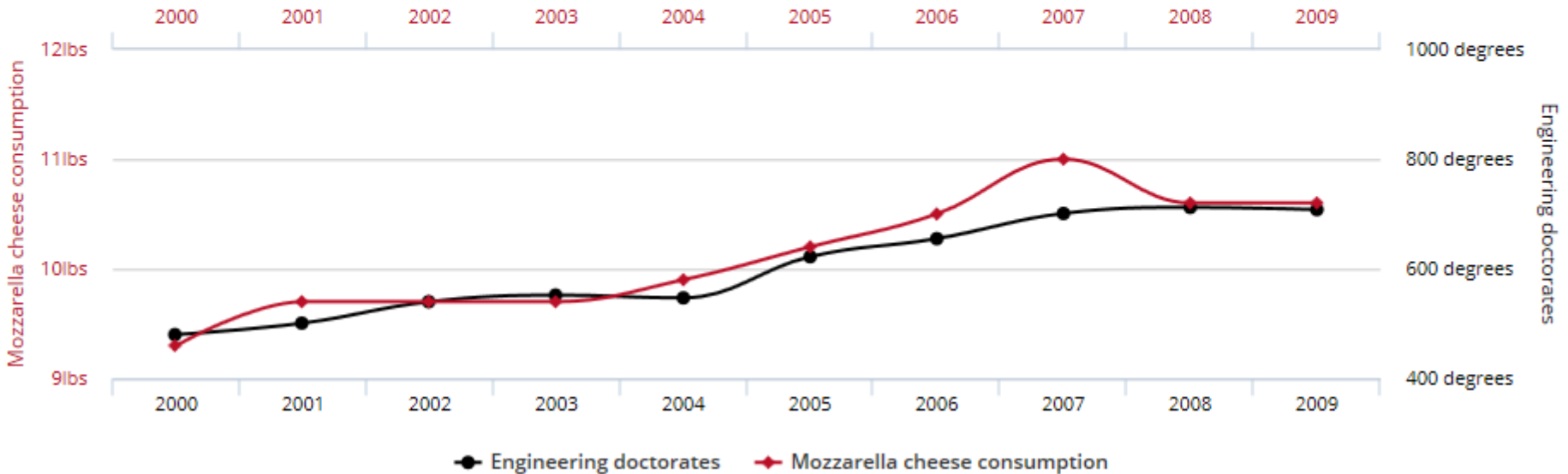
La corrélation est l'intensité de l'association ou de la liaison entre plusieurs variables numériques. Autrement dit, les variables étudiées varient simultanément.

- Exemple précédent: les MPOC et l'âge de l'individu. Mais quelle est l'intensité de cette relation ?
- Autres exemples: Taille et poids, tabagisme et VO2max, consommation de crème glacée et noyade
- L'association peut être positive ou négative, elle peut être forte, faible ou nulle.
- Corrélation \neq causalité

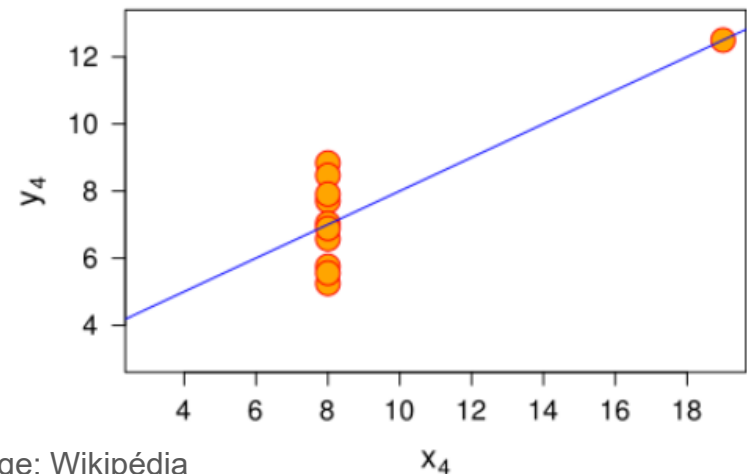
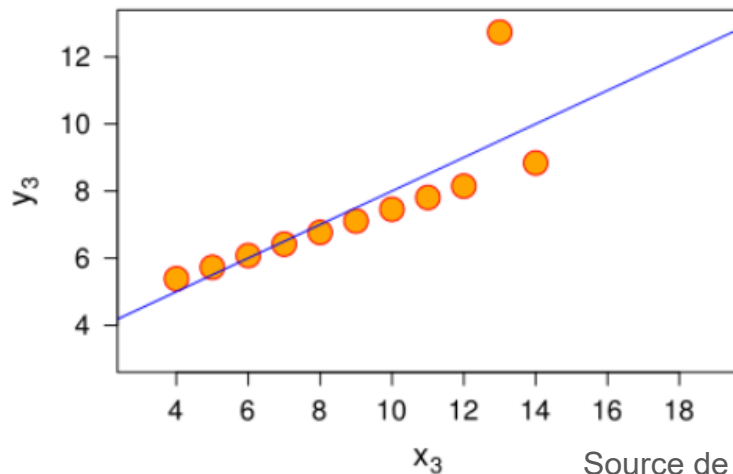
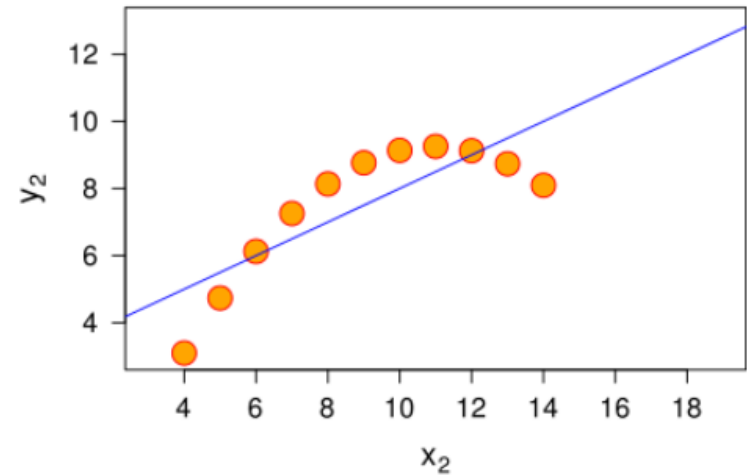
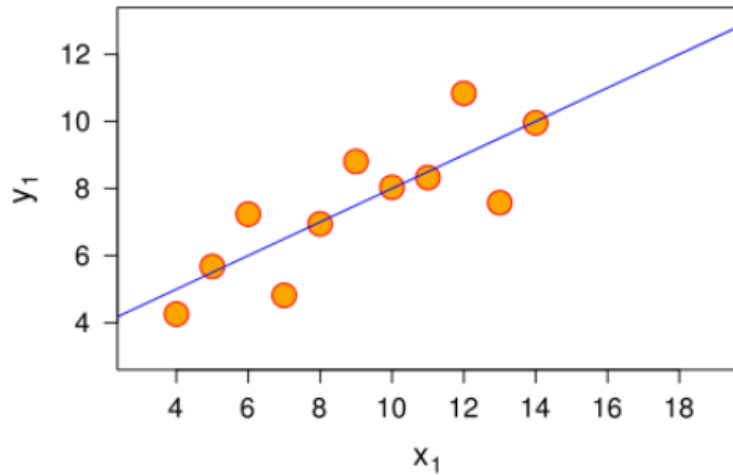
Corrélation

Per capita consumption of mozzarella cheese correlates with Civil engineering doctorates awarded

Correlation: 95.86% (r=0.958648)



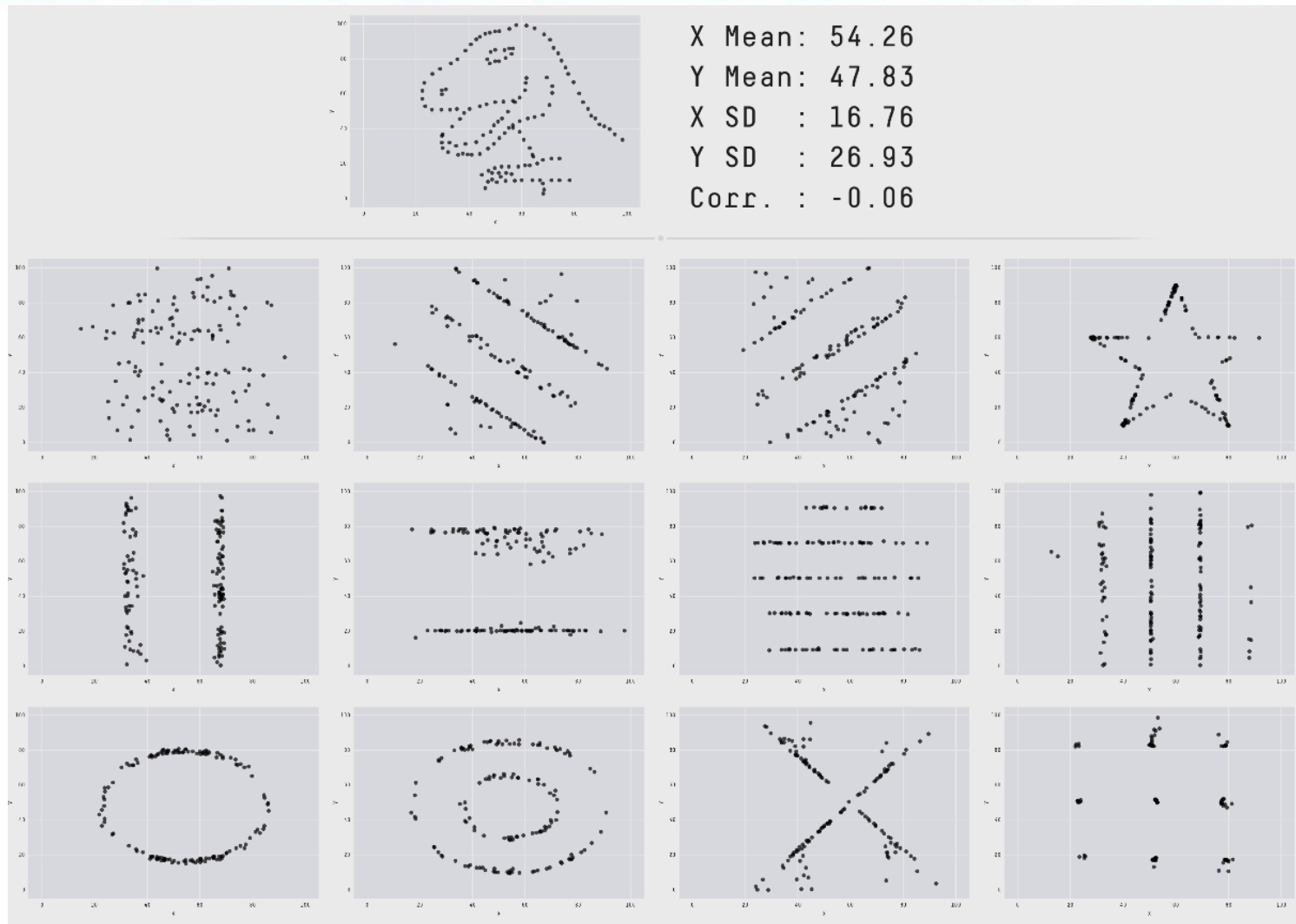
Corrélation (quartet d'Anscombe)



Source de l'image: Wikipédia



Corrélation (Datasaurus)



Modélisation statistique

- Le calcul des prévalences, des taux d'incidence et des taux de mortalité permet de **décrire** la population.
- Afin de mieux **comprendre** une maladie, la surveillance implique aussi d'**identifier** les facteurs qui expliquent les différences dans la population.
- Exemple précédent: l'âge explique en partie les différences qu'on constate chez les MPOC.
- La standardisation permet d'ajuster pour une seule variable. La modélisation permet d'ajuster pour plusieurs variables en même temps, en plus d'identifier celles ayant le plus d'importance et celles ne jouant aucun rôle.

Modélisation statistique



- La régression est la forme de modélisation la plus connue permettant d'estimer la relation entre des variables.
- Il existe de nombreux types de régression, les plus communs et les plus utilisés en santé étant: linéaire, logistique, Poisson, survie ou durée de vie (Cox).

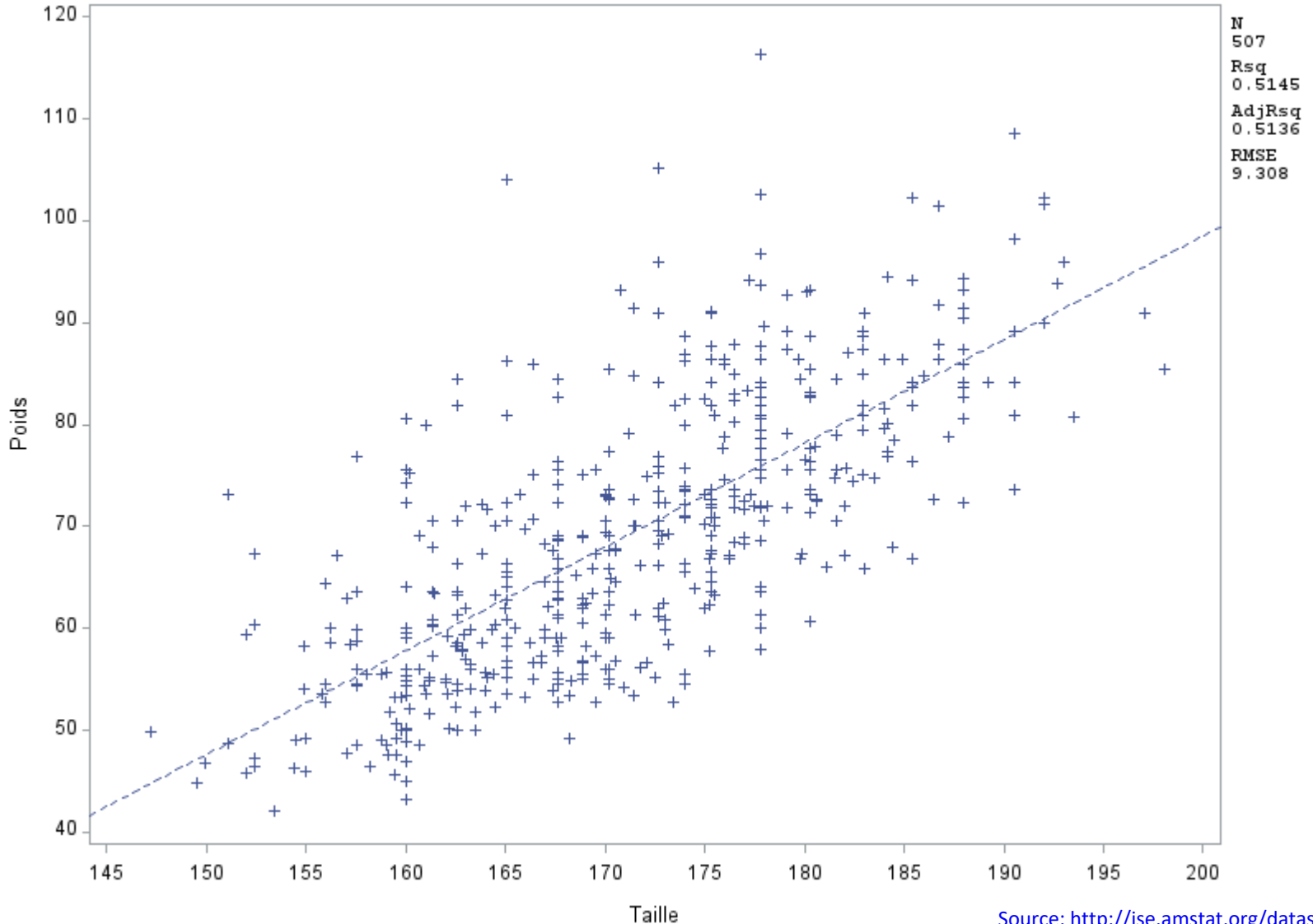
Régression linéaire

À utiliser dans quel contexte ?

- Lorsque la variable dépendante Y est continue (taille d'un individu, tension artérielle, notes des étudiants).
- Si on anticipe que la relation entre Y et les X_i est linéaire et que certains postulats sont respectés.
- Quand on désire identifier quelles variables indépendantes X_i (continue ou catégorique) prédisent de manière statistiquement significative la variable dépendante Y .
- Pour obtenir une équation de régression qui permet de réaliser des prédictions.

Régression linéaire

$$\text{Poids} = -105.01 + 1.0176 \text{ Taille}$$



Régression logistique

À utiliser dans quel contexte ?

- Lorsque la variable dépendante Y est catégorique (généralement dichotomique, 1 ou 0, mort ou vivant).
- Même si la relation entre Y et les X_i n'est pas linéaire.
- Quand on veut estimer la probabilité d'un événement en fonction de plusieurs prédicteurs.
- Pour éventuellement, calculer des rapports de cotes (*odds ratio*)
 - Cependant, difficile à interpréter
 - Alternative: ajuster un modèle de régression log-binomiale mais peut présenter des problèmes de convergence

Régression logistique

Tableau 5 Variables démographiques et socioéconomiques associées à l'usage de cannabis au cours des 12 derniers mois selon les régressions logistiques multivariées, population de 15 à 34 ans, Québec, 2012

Caractéristiques	RC ¹	IC ²	Valeur p ³
Sexe			
Être un homme (c. femme)	1,97	1,38-2,82	< 0,001
Âge			
Être âgé de 25 à 34 ans (c. 15 à 24 ans)	0,65	0,39-1,07	0,090
Statut d'immigrant			
Non immigrant (c. immigrant)	2,41	1,29-4,52	0,006
Modalité de vie			
Vivre seul (c. en couple)	2,91	1,83-4,62	< 0,001
Autres modalités de vie (c. en couple)	1,90	1,14-3,16	0,013
Niveau de scolarité du ménage			
Diplôme d'études secondaires non complété (c. diplôme postsecondaire)	1,83	0,96-3,47	0,065
Diplôme secondaire ou professionnel (c. diplôme postsecondaire)	1,32	0,66-2,66	0,433

¹ RC : rapport de cotes.

² IC : intervalles de confiance à un niveau de 95 %.

³ Valeur p du test de Wald.

Régression de Poisson

À utiliser dans quel contexte ?

- Lorsque la variable dépendante Y est un nombre d'événements, une fréquence ou un taux (nombre de clients dans un commerce, taux d'incidence des MPOC).
- Pour calculer des risques relatifs tout en ajustant pour différentes variables de confusion.
- Si la distribution de ce nombre d'événements suit une loi de Poisson (il existe des alternatives si ce n'est pas le cas).
 - Souvent présence d'un problème de surdispersion (en anglais: overdispersion). Ajustement d'un modèle de régression binomiale négative ou modèle de Poisson robuste.

Régression de Poisson

Tableau A.1 Identification des comorbidités associées au risque de décès chez les cas confirmés de la COVID-19 (n = 51 880)

Condition médicale	Ensemble des cas						Par milieu de vie					
	RR brute		RR ajusté âge		RR ajusté		CHSLD		RPA		Domicile	
	RR	IC 95%	RR	IC 95%	RRa	IC 95%	RRa	IC 95%	RRa	IC 95%	RRa	IC 95%
Maladies respiratoires	2,69	(2,55-2,83)	1,35	(1,28 - 1,41)	1,15	(1,09-1,20)	1,07	(1,01-1,13)	1,24	(1,10-1,39)	1,44	(1,24 - 1,69)
Maladies cardiovasculaires	5,62	(5,36-5,90)	1,45	(1,38 - 1,52)	1,12	(1,07-1,18)	0,99	(0,91-1,09)	1,00	(0,82-1,21)	1,36	(1,11 - 1,67)
Diabète	3,31	(3,15-3,47)	1,47	(1,40 - 1,54)	1,13	(1,08-1,19)	1,08	(1,03-1,14)	1,19	(1,05-1,33)	1,29	(1,11 - 1,50)
Cancer	2,68	(2,52-2,86)	1,29	(1,21 - 1,37)	1,12	(1,05-1,18)	1,02	(0,95-1,09)	1,16	(1,02-1,33)	1,36	(1,16 - 1,60)

(Source: Simard et al, 2020)

Régression de Cox (de survie)

À utiliser dans quel contexte ?

- Lorsque la variable dépendante Y est catégorique (généralement dichotomique, 1 ou 0, mort ou vivant).
- Quand on veut aussi tenir compte du temps nécessaire avant que l'événement survienne.
- Si censurer des sujets est nécessaire (parce qu'ils quittent la cohorte).
- Pour obtenir des rapports de taux (hazard ratio)
- Afin de tenir compte d'événements compétitifs.

Régression de Cox

Table 5

Associations between asthma onset and time-varying air pollutant levels, per interquartile range increase in pollutant levels at the residential address.^a

Pollutant	Sample size	Interquartile range	Hazard ratios (95% CI)		
			Crude	Model 1 ^b	Model 2 ^c
NO ₂ ^d	216,746	5.27 ppb	1.10 (1.08–1.12)*	1.07 (1.05–1.09)*	1.04 (1.03–1.06)*
O ₃ ^e	829,277	3.26 ppb	1.10 (1.09–1.11)*	1.13 (1.11–1.14)*	1.07 (1.06–1.08)*
PM _{2.5} ^f	1,133,938	6.53 µg/m ³	1.31 (1.30–1.33)*	1.32 (1.31–1.33)*	1.33 (1.31–1.34)*

^aExposures based on annual levels. ^bAssociations adjusted for sex and indexes of social and material deprivation.

^cAssociations adjusted for year of birth, sex, indices of social and material deprivation. ^dRestricted to the Montreal subcohort, 1996 to 2006. ^eFor the years 1999–2010. ^fFor the years 1996–2011. **p* < 0.001.

Rapport de cotes vs. risque relatif

- Les deux vont presque toujours dans le même sens.
- Les deux sont presque identiques lorsque l'événement est rare (<5%)
- Le rapport de cotes tend à surestimer l'association lorsque l'événement est commun.
- Le risque relatif est plus facile à interpréter.
- Le risque relatif ne peut pas être calculé dans la plupart des études de type cas-témoin.
- Les rapports de cotes sont plus facile à ajuster par des covariables (via une régression logistique)

Rapport de taux vs. risque relatif

- Les deux vont généralement dans le même sens.
- Le risque relatif ne tient pas compte du temps.
- Et donc l'interprétation diffère légèrement (ici on suppose la même estimation):
 - Le risque relatif indique que le risque de développer une MPOC est trois fois plus élevé chez les fumeurs par rapport aux non-fumeurs sur l'ensemble de la période.
 - Le rapport de taux indique que le risque de développer une MPOC est trois fois plus élevé chez les fumeurs à tout moment pendant la période.

Sujets spéciaux



www.inspq.qc.ca

Indice de défavorisation

www.inspq.qc.ca

Indice de défavorisation



- Rappel: une des principales faiblesses des bases de données administratives est l'absence d'information socio-économique sur les individus.
- Sans une telle information, il devient très difficile d'étudier les inégalités sociales de santé.
- Qui plus est, en santé, un modèle statistique qui ne tient pas compte du profil socio-économique des individus est un modèle moins fiable (variable confondante)

Indice de défavorisation

- L'indice de défavorisation vient pallier (partiellement) ce manque d'information socio-économique.
- Il utilise des données du recensement, qui sont produites à une échelle géographique très fine (aire de diffusion), qui couvrent l'ensemble du pays et qui sont très fiables.
- Une aire de diffusion contient environ de 400 à 800 personnes.

Indice de défavorisation

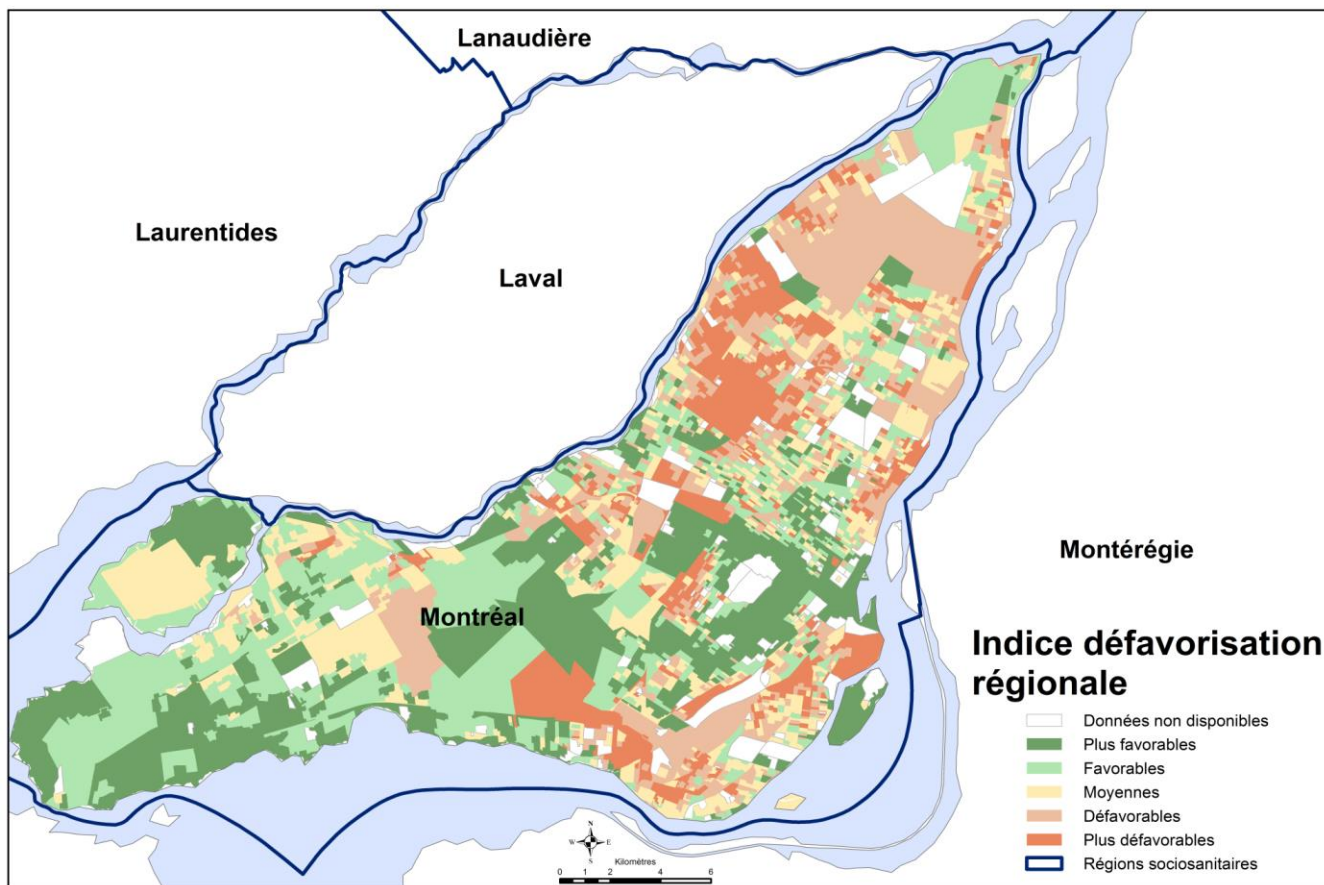
- L'indice de défavorisation a deux composantes:
 - Matérielle : formée principalement par le revenu moyen, la proportion de personnes sans DES et la proportion de personnes avec un emploi.
 - Sociale : formée principalement par les proportions de familles monoparentales, de personnes vivant seules et de personnes séparées, divorcées ou veuves.

Indice de défavorisation

- Chaque composante est représentée par un quintile de défavorisation, chaque quintile regroupant 20% de la population.
- Ces quintiles de défavorisation matérielle et sociale peuvent être assignées à n'importe quelle base de données comprenant le code postal de l'individu.

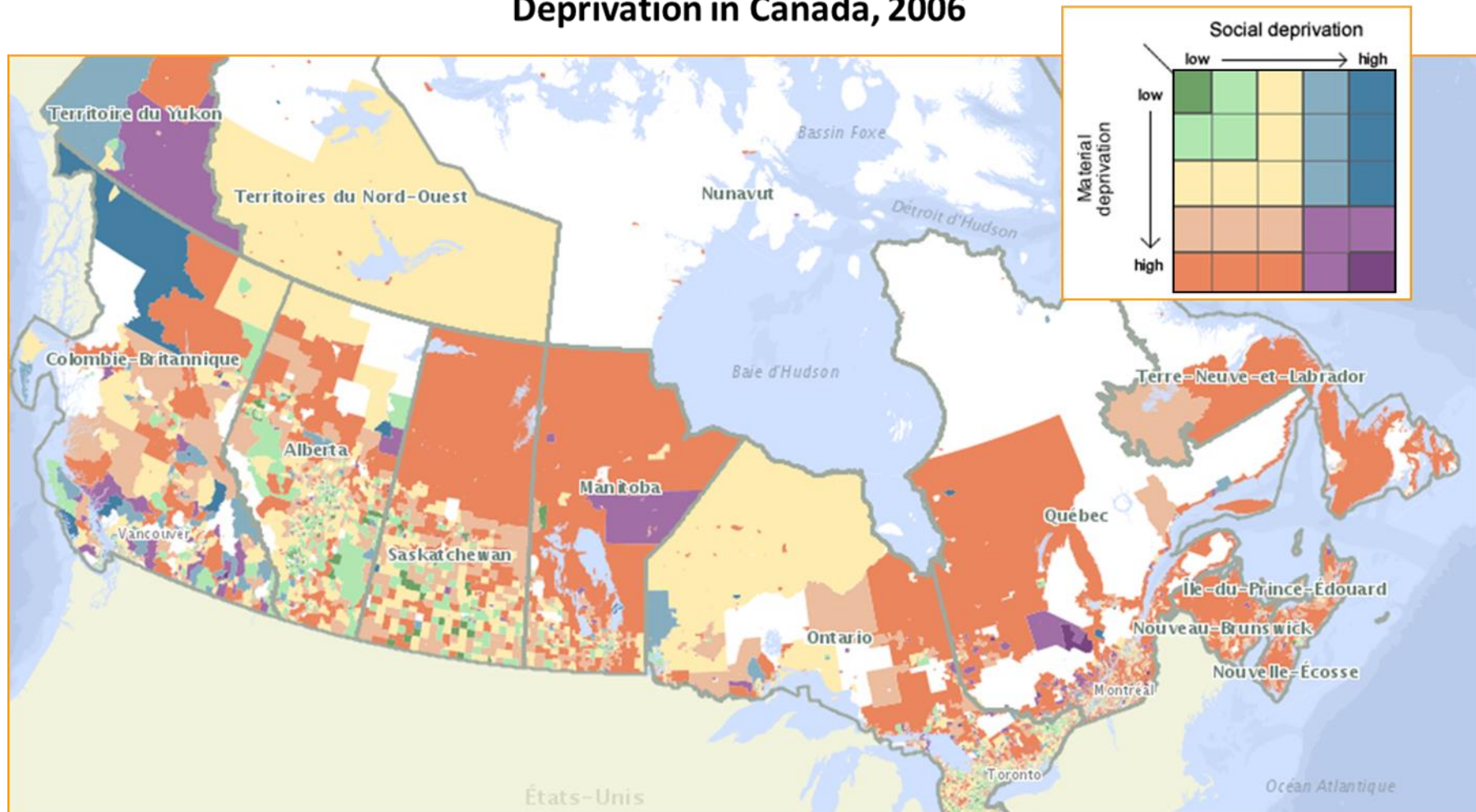
Indice de défavorisation

Indice de la défavorisation matérielle régionale pour la région de Montréal, 2011

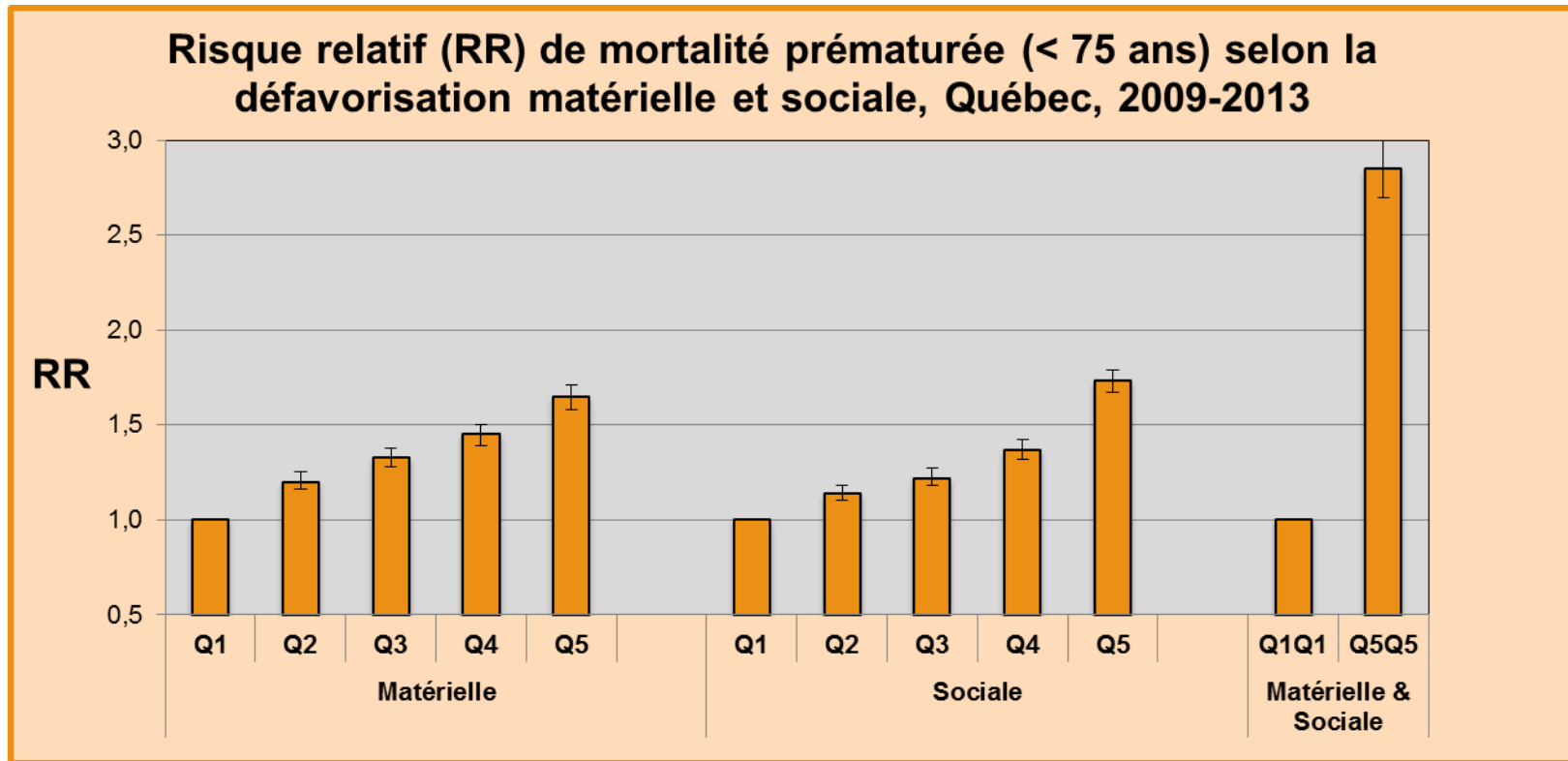


Indice de défavorisation

Deprivation in Canada, 2006



Exemple d'inégalité sociale de santé



NOTE: Les risques relatifs sont ajustés selon l'âge, le sexe, la zone de résidence et l'autre forme de défavorisation (matérielle ou sociale) chez la population.



Indice de défavorisation

Quelques limites à mentionner :

- Ne tient pas compte d'autres facteurs pouvant être associés à la santé tels que l'origine ethnique, l'immigration, l'identité autochtone et l'environnement bâti.
- On suppose que les individus à l'intérieur d'une même aire de diffusion sont homogènes, ce qui n'est pas nécessairement le cas.
- L'indice sous-estime les inégalités sociales de santé par rapport à l'utilisation de données individuelles.

Indice de multimorbidité / comorbidité

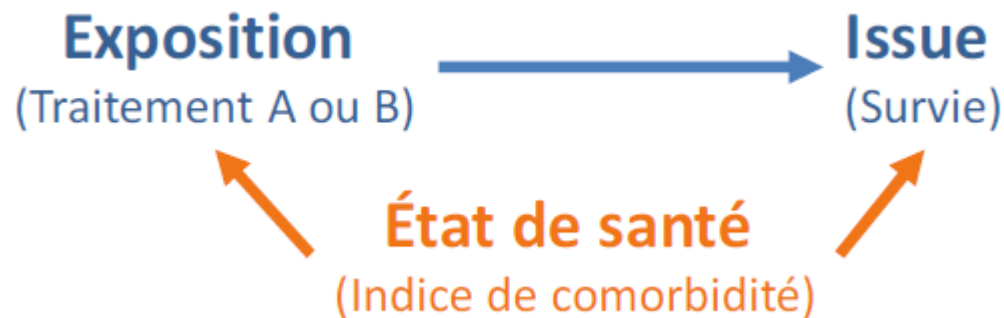
www.inspq.qc.ca

Indice de multimorbidité / comorbidité

- La **multimorbidité** est la coexistence de deux maladies/conditions ou plus chez un même individu.
→ Se rapportent donc aux individus
- La **comorbidité** est la présence de maladies/conditions pour les personnes ayant une maladie donnée appelée maladie index.
- Un indice de multimorbidité ou de comorbidité est une valeur numérique qui résume le fardeau médical chez un individu dans une population. C'est un proxy de son état de santé.
- L'indice de Charlson, qui combine 17 maladies chroniques, est probablement le plus connu.

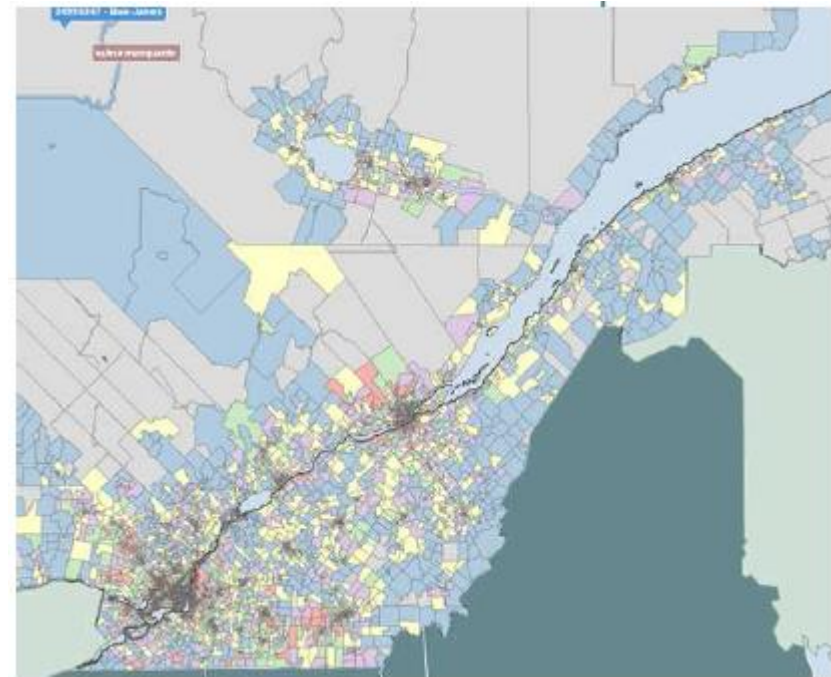
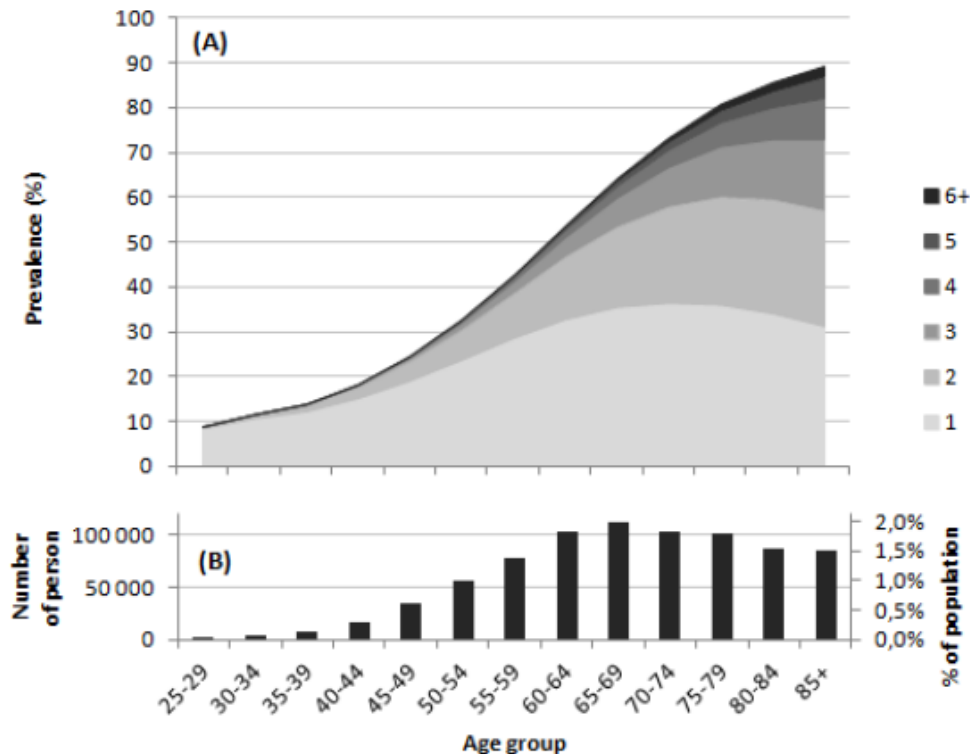
Indice de multimorbidité

- Tout comme l'indice de défavorisation, l'indice de multimorbidité sert principalement comme:
 - Variable de stratification
 - Covariable d'ajustement dans la modélisation statistique.



Indice de multimorbidité

- L'indice peut aussi permettre de dresser un portrait de la multimorbidité:



Indice de multimorbidité

- Le consensus scientifique est qu'un tel indice doit inclure au moins 12 maladies.
- L'indice de multimorbidité peut être:
 - Un cumul de maladies
 - Une liste de maladies (Oui/Non pour chaque maladie)
 - Un score pondéré qui tient compte de la sévérité des maladies. Par exemple, le cancer métastatique a un poids plus élevé que l'arthrite rhumatoïde.

Indice de multimorbidité

- L'indice québécois récemment développé couvre 32 maladies.

Hypertension
Arythmie cardiaque
Troubles vasculaires périphériques
Infarctus du myocarde
Insuffisance cardiaque congestive
Valvulopathie
Maladies cérébro-vasculaires
Diabète sans complication
Diabète avec complication
Anémie avec perte de sang
Anémie carencielle
Coagulopathie
Tumeurs sans métastase
Cancer métastatique
Hypothyroïdie
Maladies rénales

Troubles hydro-électrolytiques
Maladies de la circulation pulmonaire
Maladies pulmonaires chroniques
Obésité
Démence
Dépression
Troubles neurologiques
Abus d'alcool
Abus de drogue
Psychose
Maladies hépatiques
Arthrite rhumatoïde / collagénose avec manifestation vasculaire
Perte de poids
Paralysie
Ulcères gastriques
SIDA/VIH



Indice de multimorbidité

TABLE 2. Combined List of Medical Conditions, Prevalence, and Weights According to the VW, SCH, and CH Derivation Weight Methods Obtained in ICD-9 (2005/2006) and ICD-10 (2011/2012) Derivation Sample to Predict 30-day Mortality

Comorbidities	Prevalence (%)		Adjusted Odds Ratio (95% CI)*		CH Weights		SCH Weights		VW Weights	
	ICD-9	ICD-10	ICD-9	ICD-10	ICD-9	ICD-10	ICD-9	ICD-10	ICD-9	ICD-10
Hypertension [†]	36.1	41.7	0.70 (0.68–0.72)	0.75 (0.73–0.77)	0	0	–1	–1	–2	–3
Chronic pulmonary disease [‡]	16.1	17.4	1.26 (1.23–1.30)	1.24 (1.21–1.28)	1	1	1	1	2	2
Cardiac arrhythmias [†]	13.4	16.5	1.17 (1.13–1.21)	1.22 (1.18–1.25)	0	1	1	1	1	2
Diabetes, uncomplicated [‡]	13.9	15.7	1.04 (1.01–1.08)	1.06 (1.03–1.09)	0	0	0	0	0	0
Deficiency anemia [†]	9.1	13.2	0.85 (0.82–0.88)	0.83 (0.81–0.86)	0	0	–1	–1	–1	–2
Any tumor without metastasis [§]	12.0	12.8	2.44 (2.35–2.53)	2.57 (2.47–2.66)	2	3	3	3	6	11
Hypothyroidism [†]	10.6	12.5	0.83 (0.79–0.86)	0.92 (0.89–0.95)	0	0	–1	0	–1	–1
Renal disease [§]	8.6	10.3	1.42 (1.37–1.47)	1.30 (1.26–1.35)	1	1	1	1	2	3
Fluid and electrolyte disorders [†]	5.3	8.6	1.93 (1.86–2.01)	1.75 (1.69–1.81)	2	2	2	2	5	6
Peripheral vascular disorders [†]	6.6	8.0	1.48 (1.42–1.54)	1.17 (1.13–1.21)	1	0	1	1	3	2
Myocardial infarction [§]	8.4	7.8	2.03 (1.96–2.10)	1.89 (1.82–1.96)	2	2	2	2	5	7
Congestive heart failure [‡]	6.5	6.7	2.03 (1.95–2.11)	1.87 (1.81–1.95)	2	2	2	2	5	7
Obesity [†]	5.2	6.2	0.99 (0.92–1.05)	0.88 (0.83–0.94)	0	0	0	0	0	0
Valvular disease [‡]	5.1	5.9	0.71 (0.67–0.75)	0.80 (0.76–0.83)	0	0	–1	–1	–2	–3
Metastatic cancer [‡]	5.0	5.4	5.38 (5.15–5.63)	5.15 (4.93–5.37)	5	5	6	5	12	19
Dementia [§]	3.7	5.4	1.69 (1.62–1.76)	1.89 (1.82–1.96)	2	2	2	2	4	7
Cerebrovascular disease [§]	5.9	5.3	2.12 (2.03–2.20)	2.35 (2.26–2.45)	2	2	2	3	5	10
Depression [†]	5.4	5.2	0.65 (0.60–0.70)	0.66 (0.62–0.71)	0	0	–1	–1	–3	–5
Neurological disorders [†]	4.3	4.8	2.00 (1.91–2.10)	1.72 (1.64–1.79)	2	2	2	2	5	6
Alcohol abuse [†]	4.1	4.0	1.18 (1.10–1.27)	1.29 (1.21–1.37)	0	1	1	1	1	3
Liver disease [†]	3.1	3.9	1.65 (1.54–1.76)	2.28 (2.17–2.39)	2	2	2	3	3	9
Diabetes, complicated [‡]	2.7	3.8	0.87 (0.81–0.94)	0.83 (0.78–0.88)	0	0	0	–1	–1	–2
Psychoses [†]	3.3	2.7	1.34 (1.25–1.44)	0.94 (0.84–1.05)	1	0	1	0	2	0
Pulmonary circulation disorders [†]	1.9	2.6	2.21 (1.99–2.25)	1.47 (1.39–1.55)	2	1	2	1	5	4
Rheumatoid arthritis/collagen vascular	2.1	2.4	1.21 (1.21–1.41)	0.98 (0.91–1.06)	1	0	1	0	2	0

Tiré de l'article de

Simard M, Sirois C, Candas C. Validation of the Combined Comorbidity Index of Charlson and Elixhauser to Predict 30 Day Mortality Across ICD-9 and ICD-10. *Medical Care*. 2018;56(5): 441-447.

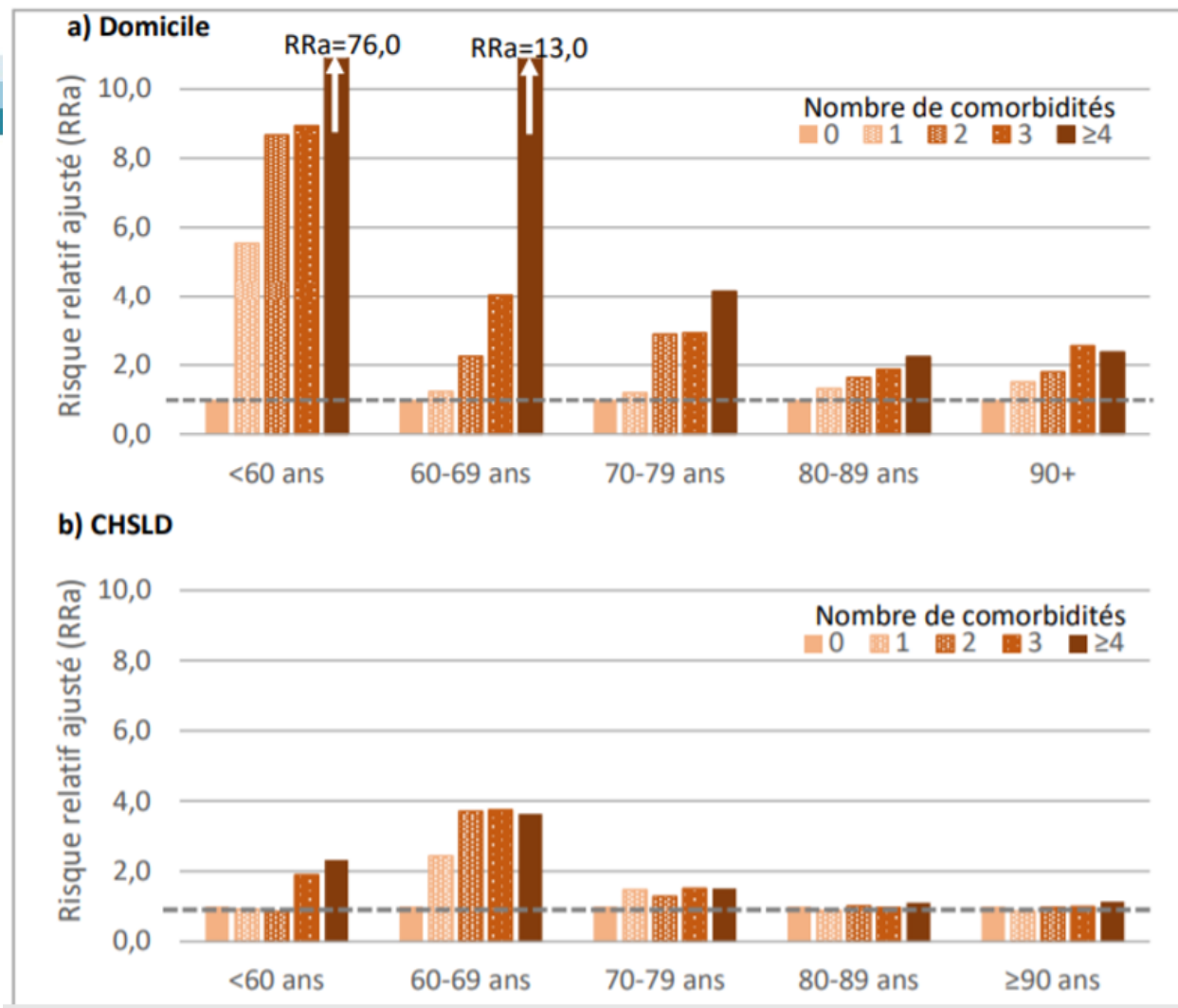


Indice de multimorbidité

- Ayant été démontré qu'il est plus pertinent que le cumul et tout aussi performant que de considérer une liste de maladies, le score pondéré a été privilégié.
- Les poids ont été recalculés pour qu'ils reflètent mieux la réalité québécoise d'aujourd'hui.
 - Par exemple, le poids associé au VIH/SIDA au Québec (3) est beaucoup moins élevé que pour Charlson (6).

Indice de multimorbidité / comorbidité

Cumul des comorbidités et risque relatif ajusté de décès



14 décembre
2020

Impact des comorbidités sur les risques de décès et d'hospitalisation chez les cas confirmés de la COVID-19 durant les premiers mois de la pandémie au Québec

Marc Simard, Catherine de Montigny, Sonia Jean, Élise Fortin, Claudia Blais, Isabelle Théberge, Louis Rochette, Marjolaine Dubé, Pierre-Luc Trépanier, Rodica Gilca

Analyse de survie: risques de traumatismes selon l'usage de médicaments parmi des personnes atteintes de TDAH

www.inspq.qc.ca

Questions de recherche



- Le TDAH accroît le risque de traumatismes non-intentionnels.
- Le but de l'étude est de déterminer s'il existe une relation entre l'usage de médicaments pour le TDAH et le risque de traumatismes impliquant une visite à l'urgence ou une hospitalisation.

Méthodologie

- Étude de cohorte rétrospective entre 2000 et 2021 d'individus de moins de 25 ans à l'aide de bases de données administratives.
- Les bases sont:
 - Services pharmaceutiques parmi les adhérents au régime public d'assurance-médicaments.
 - Hospitalisations
 - Services médicaux rémunérés à l'acte
 - Fichier d'inscription des personnes assurées à l'assurance-maladie

Méthodologie



- L'exposition est définie en trois catégories selon des épisodes de médication et/ou de couverture au régime public d'assurance-médicament
 - Couvert par le régime public sans prise de médicaments
 - Couvert par le régime public avec prise de médicaments
 - Non couvert par le régime public

Analyses

- Analyse de survie classique: suivi des individus dans le temps selon une exposition jusqu'à ce qu'on observe l'évènement ou la fin du suivi.
- Complexité de l'analyse:
 - Plusieurs événements traumatiques peuvent être observés lors de la période à l'étude.
 - L'exposition varie dans le temps: prise ou non prise de médicaments, entrée et sortie du régime public d'assurance-médicaments

Analyses

- Ajustement d'un modèle de Prentice-Williams-Perterson.
 - Ce modèle est une adaptation de modèle de régression de Cox qui permet des évènements récurrents.
- Pour l'exposition qui varie au cours du suivi (*time-dependent variable*) , nous devons adapter la façon de définir l'exposition par une variable de comptage : de t1 à t2: exposition A, de t2 à t3: exposition B etc.
 - Ces périodes de comptages sont définies pour chaque individu séparément et sont appelées épisodes.
 - Tous les épisodes sont ensuite entrés dans le modèle final.

Analyses



- L'ajustement du modèle permet d'obtenir des rapport de taux (*Hazard Ratios*) comparant l'exposition vs non exposition aux évènements traumatiques ajustés pour l'âge, le sexe, les comorbidités, indice de défavorisation.

Dénombrement des personnes en situation d'itinérance (PSI)

www.inspq.qc.ca

Dénombrement des PSI

- L'itinérance est une problématique complexe et difficile à décrire, car les personnes en situation d'itinérance sont difficiles à joindre.
- Un dénombrement des personnes en situation d'itinérance a eu lieu au Québec en 2018 : <https://publications.msss.gouv.qc.ca/msss/document-002292/>
- Le dénombrement de 2022 a trois objectifs :
 - Estimer le nombre de personnes en situation d'itinérance visible au Québec, au cours de la nuit du 11 au 12 octobre 2022;
 - Décrire le profil des personnes en situation d'itinérance, grâce à la passation d'un questionnaire;
 - Comparer les résultats de l'exercice de 2022 à celui de 2018.

Dénombrement des PSI

- Penchons-nous sur l'objectif d'estimer le nombre de personnes en situation d'itinérance visible.
- Lieux principaux où les personnes en situation d'itinérance visible peuvent avoir passé la nuit du 11 au 12 octobre :
 - Lieux extérieurs
 - Organismes communautaires
 - Hôpitaux
 - Centres de détention

Dénombrement des PSI

- Hôpitaux : depuis 2018, les données administratives de MED-ECHO ont un code permettant de savoir si la personne est en situation d'itinérance ou non.
- Centres de détention : il existe aussi un code administratif, mais sa fiabilité n'est pas validée.
Nous avons donc tenté de demander à tous les détenus s'ils avaient accès à un domicile permanent et sécuritaire au moment de leur incarcération.
- Organismes : nous avons dressé une liste exhaustive des organismes et ils nous ont transmis le nombre de personnes hébergées la nuit du 11 octobre 2022 par un sondage en ligne.

Dénombrement des PSI

- Lieux extérieurs : cela pose plusieurs défis !
- On procède par sillonnage :
 - Chaque région identifie les secteurs susceptibles de contenir des personnes qui passeront la nuit à l'extérieur.
 - Le soir du 11 octobre, des équipes parcourent les secteurs de sillonnage, et abordent tout le monde sur leur chemin.
 - Parmi les personnes interrogées, certaines personnes mentionnent qu'elles passeront la nuit à l'extérieur.

Dénombrement des PSI

- Comment s'assurer que le nombre soit le plus exact possible ? En faisant notamment les ajustements suivants :
 - Ajustement pour les secteurs non couverts
 - Ajustement pour les personnes leurres

Dénombrement des PSI

Ajustement pour les secteurs non couverts

- À Montréal, certains secteurs de sillonnage n'ont pas pu être couverts, car il manquait de bénévoles.
- Pour ces secteurs, on suppose que le nombre moyen de personnes en situation d'itinérance qu'ils contiennent est égal au nombre moyen de personnes dénombrées dans les secteurs couverts du même arrondissement.
- On suppose donc que les secteurs à l'intérieur d'un même arrondissement sont homogènes.
- Pour que cette approche soit rigoureuse, il faut choisir aléatoirement les secteurs qui seront sillonnés le soir du dénombrement.

Dénombrement des PSI

Ajustement pour les personnes leurres

- Même à l'intérieur des secteurs sillonnés, on ne peut pas être certain d'avoir abordé tout le monde, surtout lorsque la densité de population est grande. Rappelons aussi que la population est mobile.
- Il faut donc avoir un moyen pour évaluer la qualité du sillonnage. L'outil utilisé est le déploiement de personnes leurres dans certains secteurs de sillonnage.

Dénombrement des PSI

Ajustement pour les personnes leurres

- Les personnes leurres sont des bénévoles qui ont comme consigne de se positionner dans un secteur de sillonnage. Lorsqu'elles sont abordées par une équipe de bénévole, elles mentionnent qu'elles sont des personnes leurres.
- En utilisant la proportion de personnes leurres trouvées, on peut corriger l'estimé du nombre de personnes ayant passé la nuit à l'extérieur, pour qu'il soit plus représentatif de la réalité.
- Par exemple, si 50% des personnes leurres ont été abordées, on double les estimations obtenues.

Dénombrement des PSI

En résumé...

- Chaque problématique de santé peut nécessiter le développement de méthodes statistiques spécifiques à cette problématique.
- Il faut parfois user d'ingéniosité et combiner à la fois des données administratives, des données issues de sondages en ligne et des données issues de questionnaires papiers.

Merci de votre attention

Coordonnées:

- louis.rochette@inspq.qc.ca

- jeremie.sylvain-morneau@inspq.qc.ca

www.inspq.qc.ca