Utilité des mégadonnées en surveillance des maladies infectieuses et contribution possible aux enquêtes sur les maladies d'origine alimentaire au Canada : aperçu et document de travail

May 2017

Cheryl Waldner, D.M.V., Ph. D.
Professeure, Université de la Saskatchewan

Avec la contribution de :

Nathaniel Osgood, Ph. D.
Professeur agrégé, Université de la Saskatchewan

Patrick Seitzinger
Candidat à la maîtrise en santé publique, École de santé publique

National Collaborating Centre for Infectious Diseases
Centre de collaboration nationale des maladies infectieuses

Utilité des mégadonnées en surveillance des maladies infectieuses et contribution possible aux enquêtes sur les maladies d'origine alimentaire au Canada : aperçu et document de travail

Préparé à l'intention du Centre de collaboration nationale des maladies infectieuses (CCNMI)

May, 2017

Cheryl Waldner DVM PhD
Professeure, Université de la Saskatchewan

Avec la contribution de :

Nathaniel Osgood, Ph. D.
Professeur agrégé, Université de la Saskatchewan

Patrick Seitzinger
Candidat à la maîtrise en santé publique, École de santé publique

Centre de collaboration nationale des maladies infectieuses
Rady Faculty of Health Sciences
Université du Manitoba
Tél.: 204-318-2591
Courriel: nccid@umanitoba.ca
www.ccnmi.ca
Projet no 332 du CCNMI

La production de ce document a été rendue possible grâce à la contribution financière de l'Agence de santé publique du Canada dans le cadre du financement des activités du Centre de collaboration nationale des maladies infectieuses. Les opinions exprimées ici ne représentent pas nécessairement le point de vue de l'Agence de la santé publique du Canada.

Utilité des mégadonnées en surveillance des maladies infectieuses et contribution possible aux enquêtes sur les maladies d'origine alimentaire au Canada : aperçu et document de travail

Sommaire

Devant le potentiel des « mégadonnées » de constituer une « ressource inexploitée de faits probants qui peuvent servir à orienter l'élaboration des politiques et la prise de décisions », le Centre de collaboration nationale des maladies infectieuses (CCNMI) a commandé un document de travail pour lancer les discussions parmi les professionnels de la santé publique du Canada. Le document avait pour principal objet d'examiner si les mégadonnées peuvent contribuer à la politique de santé publique en ce qui a trait à la gestion des maladies infectieuses au Canada. La première partie du rapport présente les mégadonnées et diverses sources de ces données pour la surveillance des maladies infectieuses, décrites dans la littérature évaluée par des comités de lecture et des rapports techniques d'accès public jusqu'au 31 décembre 2016. L'auteure et ses collaborateurs examinent également certaines des options et des difficultés liées à la visualisation et à l'analyse des résultats pour que ces derniers puissent plus efficacement servir la prise de décisions. Ils décrivent enfin les types de mégadonnées qui ont été utilisés en surveillance et dans les enquêtes sur les maladies d'origine alimentaire; citent un chercheur qui applique les mégadonnées aux maladies d'origine alimentaire; et résument les résultats d'un sondage officieux portant sur les mégadonnées auprès de professionnels de la santé publique dont le travail a trait aux maladies infectieuses et aux enquêtes sur les éclosions au Canada.

Les mégadonnées sont très souvent décrites par les difficultés de gestion et d'analyse qu'engendrent leur taille ou leur quantité (volume); par le rythme auquel elles sont colligées, transmises ou reçues (vélocité); et par l'éventail possible des sources, des types de fichiers et des structures de données possibles (variété). Nous avons examiné divers exemples de mégadonnées pour en déterminer l'utilité en surveillance des maladies infectieuses et dans les enquêtes sur les éclosions. Les exemples les plus souvent donnés étaient les suivants : séquençage pangénomique, résumés analytiques de nouvelles provenant d'Internet, statistiques des moteurs de recherche sur les recherches faites dans Internet ou l'utilisation de pages Web, récits trouvés dans des forums d'Internet, information provenant de sites de médias sociaux, données recueillies de manière active et passive au moyen de téléphones intelligents, dossiers des ventes au détail dans les pharmacies et les épiceries, dossiers de santé électroniques, surveillance active externalisée dans Internet, registres des lignes téléphoniques d'info santé, dossiers d'absentéisme et données d'observation de la Terre (y compris la télédétection).

Divers documents de recherche et articles de revue originaux ont porté sur l'application de ces divers types de mégadonnées en santé publique. Nous avons dégagé de ces travaux plusieurs avantages possibles qui pourraient justifier l'inclusion de nouvelles sources de mégadonnées dans les programmes existants pour améliorer à la fois la surveillance des maladies infectieuses et les enquêtes sur les éclosions de maladies. Ces avantages sont les suivants : 1.) améliorer l'opportunité, la

résolution géographique et l'intégralité de l'information; 2.) combler les lacunes de couverture dans les programmes existants de surveillance; 3.) accroître la précision de la surveillance dans le cas des maladies émergentes et réémergentes; 4.) trouver des faits probants pour mieux orienter les modèles prédictifs et la planification pour la gestion des maladies. Nous avons cité l'exemple de la croissance inédite des possibilités offertes par les téléphones intelligents et les portables pour la compréhension du comportement humain, les répercussions de ce dernier sur la transmission des maladies et les mesures de lutte contre cette transmission. Les téléphones intelligents et autres appareils portables semblables peuvent fournir des données de haute résolution et en temps réel sur l'emplacement, l'activité et les contacts; ils offrent en outre la possibilité de limiter les enquêtes à certains contextes ou de les mener à la suite de l'intervention d'utilisateurs; ils fournissent en outre des outils pour la collecte de données biométriques, visuelles, vidéographiques et sonores.

Le présent rapport avait entre autres objectifs précis d'explorer les domaines dans lesquels les mégadonnées ont été utilisées pour améliorer la surveillance des maladies entériques et les enquêtes sur les éclosions de maladies d'origine alimentaire. La première – et l'une des répercussions les plus évidentes des mégadonnées sur les maladies d'origine alimentaire – se trouve dans le domaine des diagnostics. Le séquençage pangénomique n'est plus réservé aux laboratoires de recherche et au Laboratoire national de microbiologie, car maintenant la plupart des laboratoires provinciaux de diagnostic ont accès à la technologie nécessaire. L'électrophorèse en champ pulsé (ECP) a longtemps été la norme pour l'identification de souches étroitement apparentées, mais le séquençage pangénomique devient de plus en plus une solution de rechange abordable et efficace à l'ECP, car il offre, par comparaison, un stockage des données plus efficient et la possibilité d'obtenir une résolution plus fine pour différencier les organismes, de même que la prédiction du type de souche, de la virulence et de la résistance antimicrobienne.

Le deuxième sujet d'étude le plus important des mégadonnées et des enquêtes sur les maladies d'origine alimentaire est l'application des données obtenues par externalisation ouverte et extraites des forums d'Internet, dont les sites d'évaluation en ligne des restaurants tels que Yelp, et des médias sociaux, Twitter étant le plus souvent cité. Plusieurs grandes villes des États-Unis ont utilisé des données provenant de Yelp ou de Twitter pour cibler des inspections de restaurants et elles ont obtenu, dans certains cas, des résultats très encourageants pour ce qui est de repérer les lieux qui enfreignent gravement le code sanitaire. On cherchait en outre, dans ces initiatives, à diminuer le laps de temps entre le début d'une éclosion et le moment où les services sanitaires locaux en prenaient connaissance et menaient une enquête pour ainsi diminuer le nombre total de personnes atteintes. Toutefois, compte tenu de la nature non structurée, diversifiée et toujours changeante des données brutes, il faut des algorithmes d'apprentissage automatique perfectionnés pour reconnaître les cas possibles, assez sensibles et spécifiques pour qu'il ne faille pas trop de temps aux experts pour les évaluer.

Les auteurs ont également observé un intérêt constant pour la collecte et l'analyse de données sur les ventes au détail afin de détecter les éclosions de maladie et les sources possibles de maladies d'origine alimentaire. La surveillance dans les pharmacies des médicaments d'ordonnance et des médicaments en vente libre a fait l'objet d'un examen tant au Canada qu'ailleurs dans le monde et les résultats ont été plus ou moins probants dans le cas des maladies entériques. Les données des cartes de fidélité des consommateurs ont également été utilisées dans quelques cas au Canada pour

déterminer les causes les plus probables d'éclosion de maladies d'origine alimentaire. IBM a récemment exploré l'utilisation d'un système d'analyse spatiotemporelle pour comparer les données de lecture obtenues par balayage à la vente au détail à des cas de maladies d'origine alimentaire et dressé ensuite une liste des aliments suspects les plus probables. On cherchait ainsi à réduire le temps nécessaire pour déterminer la cause et diminuer le nombre total de cas.

Plusieurs auteurs ont insisté sur la nécessité d'élaborer et de mettre en œuvre des modèles dynamiques de simulation et de risque pour mieux tirer profit de la variété et du volume croissants des données et ainsi mieux soutenir le processus décisionnel. La compréhension de l'influenza, de la dengue et du virus Zika sont des exemples d'utilisation la plus efficace des modèles élaborés à partir des mégadonnées jusqu'à maintenant. Les chercheurs ont toutefois incité à la prudence en raison du volume et de la variété des intrants des modèles constamment mis à jour à mesure de la publication de nouvelles données et en raison du volume et de la vélocité des extrants des modèles qui en résultent. Par conséquent, pour que les modèles prédictifs et dynamiques soient efficients et efficaces dans la surveillance des maladies et la gestion des éclosions, il faut prévoir ses propres ressources de mégadonnées pour programmer et gérer les données.

Il faut en outre des compétences et des ressources particulières pour l'analyse et la visualisation afin d'intégrer et de tenir à jour efficacement les mégadonnées dans la prise de décisions en santé publique. Divers obstacles à l'utilisation et à la visualisation efficaces des données ont été observés aux égards suivants : 1) les limites des outils actuels de visualisation pour l'affichage des flux complexes de données interreliées qui changent constamment au fil du temps, 2) les limites des ressources humaines et les besoins en formation et 3) les difficultés d'accès à la technologie et aux logiciels d'information dans de nombreux organismes publics.

L'auteure et ses collaborateurs font également état d'autres mises en garde concernant l'application des mégadonnées à la surveillance des maladies infectieuses et aux enquêtes sur les éclosions. La première et peut-être la plus importante est le risque de faux positifs associés non seulement au grand volume de points de données, mais également au très grand nombre de variables dont il faudrait tenir compte pour établir d'éventuelles associations, en particulier avec l'utilisation grandissante des outils d'apprentissage automatique peu ou pas supervisés. La deuxième restriction, comme il a été dit ci-dessus, a trait à la nécessité d'acquérir de nouvelles compétences en gestion des données et, dans certains cas, un vocabulaire tout à fait nouveau pour appliquer ou du moins évaluer de manière critique les produits des mégadonnées. La troisième limite a trait aux problèmes constants en santé publique de sécurité des données, de gouvernance et de respect de la vie privée qui deviennent de plus en plus difficiles lorsqu'il faut accéder aux données à partir de différents lieux, de l'informatique nuagique, des outils nouveaux d'exploration des données et la possibilité de trouver de nouveaux moyens d'établir des liens entre des données anonymisées. Finalement, la qualité des données demeure un enjeu fondamental. Une grande partie de l'information utilisée est orientée vers d'autres fins et a été à l'origine colligée pour des raisons très différentes, ce qui en limite l'exactitude, la précision ou l'intégralité pour résoudre des questions liées à la santé. De nombreuses sources de mégadonnées se fient également à des renseignements déclarés volontairement, ce qui limite l'exactitude des descriptions et des interprétations, l'attribution des causes, le moment et le lieu d'exposition.

Les mégadonnées offrent de nombreuses possibilités d'accroître l'efficacité de la surveillance des maladies et des enquêtes sur les éclosions, mais elles s'ajoutent aux méthodes les plus récentes, elles ne les remplacent pas. Les professionnels de la santé publique qui ont répondu à notre sondage ont dit qu'ils appuyaient l'application des mégadonnées aux enquêtes sur les éclosions de maladies d'origine alimentaire. Les exemples d'histoires de réussite fournis par les répondants reflétaient dans une grande mesure les cas décrits dans la littérature. Les répondants ont toutefois signalé aussi les difficultés déjà connues de la suffisance des ressources et de la formation; des obstacles à la gestion, au stockage et à la visualisation des données, de même que les limites imposées par le respect de la vie privée et les considérations de nature juridique.

Table des matières

Contents

PREAMBLE

En septembre 2015, le *Relevé des maladies transmissibles au Canada* a consacré un numéro spécial aux mégadonnées et présenté des exemples de la façon dont elles « transforment […] notre manière de détecter les maladies infectieuses, de les contrôler et d'en effectuer le suivi » au Canada. De même, *The Journal of Infectious Diseases* a également publié un numéro spécial en 2016 dans lequel sont résumés les progrès récents dans l'application des mégadonnées dans divers domaines, dont la surveillance des maladies, les modèles de transmission des infections et le suivi des attitudes et des mouvements. L'intérêt pour ce domaine grandit rapidement si l'on considère le nombre de publications portant sur les mégadonnées et l'augmentation exponentielle des maladies infectieuses depuis 2001 (1).

Plusieurs facteurs motivent l'intérêt grandissant pour l'utilisation des mégadonnées pour améliorer la surveillance des maladies infectieuses et les enquêtes sur les éclosions de maladies, notamment les possibilités suivantes :

1. améliorer l'opportunité, la résolution géographique et l'intégralité de l'information;
2. combler les lacunes de couverture dans les programmes existants de surveillance;
3. accroître la sensibilité de la surveillance aux maladies émergentes ou réémergentes;
4. recueillir des faits probants pour mieux orienter les modèles prédictifs et la gestion des maladies.

L'action de nombreux systèmes de surveillance est entravée par le laps de temps qui s'écoule entre la survenue de la maladie et sa déclaration, de même que par les limites de la résolution spatiale des données (1). La diminution de l'intervalle entre le début de la maladie et la déclaration des données de surveillance, résumées à une échelle géographique pertinente, accélérait l'identification des éclosions à la fois locales et nationales et orienterait une gestion de la maladie plus opportune et efficace. Le temps de dépistage peut être un facteur déterminant important de la gravité de l'éclosion. La durée totale des éclosions de maladies d'origine alimentaire a, par exemple, été associée au nombre de jours qui se sont écoulés entre la déclaration des premiers symptômes et le début d'une enquête, selon une étude sur des éclosions régionales et nationales dans lesquelles sont intervenus les Centres for Disease Control de la Colombie-Britannique (2).

Les mégadonnées produites par différentes technologies et plateformes, entre autres les téléphones intelligents, les médias sociaux et Internet, offrent également la possibilité de compléter la surveillance existante (3). Les données obtenues par externalisation ouverte peuvent fournir de l'information à l'échelle

individuelle et en temps quasi réel (1). Même si l'on ne peut pas vérifier directement le nombre de cas établis qui en résultent, les empreintes numériques passives et l'information fournie par des volontaires actifs peuvent servir à compléter la surveillance des médecins et des laboratoires. Les données déclarées volontairement peuvent attirer l'attention sur des maladies que les personnes atteintes n'auraient pas jugées assez graves pour consulter un médecin et auraient, par conséquent, échappé aux systèmes de surveillance actuels (1,3). Même si les cas qui échappent aux systèmes de surveillance traditionnels peuvent être moins graves d'un point de vue clinique, ils peuvent être très importants pour comprendre la transmission de la maladie et établir des modèles prédictifs, organiser les efforts de lutte contre la maladie et mesurer le fardeau total de la maladie en raison de la perte de productivité.

En deuxième lieu, l'exploration des mégadonnées est importante aussi pour obtenir de l'information additionnelle dans des régions qui ne sont pas visées par les initiatives actuelles de surveillance et d'autres « populations cachées » (1). C'est le cas, par exemple, du système de surveillance FoodNet Canada pour obtenir un meilleur échantillonnage des agents pathogènes humains et l'échantillonnage actif des sources d'alimentation au détail dans trois sites sentinelles (http://www.phac-aspc.gc.ca/foodnetcanada/necessity-importance-fra.php) : 1) Bureau de santé de Middlesex-London en Ontario; 2) région sanitaire de Fraser en Colombie-Britannique; et 3) Services de santé de l'Alberta : zones de Calgary et Alberta centrale en Alberta. Ce programme fournit une information exceptionnelle et très précieuse, mais il faut aussi des options efficientes et efficaces pour rehausser la surveillance et mieux comprendre les facteurs de risque des maladies entériques dans des régions qui ne font pas partie directement du programme des sites sentinelles et parmi les populations aux difficultés et aux besoins d'information uniques, par exemple les collectivités autochtones éloignées et les nouveaux arrivants au Canada. Internet et les médias sociaux ont permis d'obtenir de l'information de populations « difficiles à joindre » et de mieux comprendre leurs expériences, les facteurs de risque, les contacts et les mouvements à un niveau de détail et à une ampleur qui n'étaient pas possibles auparavant (4).

D'autres options de collecte et de nouvelles options peuvent également être mises à profit pour améliorer la surveillance et lutter contre les difficultés des maladies infectieuses dans les groupes démographiques moins susceptibles de participer à des sondages téléphoniques traditionnels, y compris ceux qui n'ont pas de téléphone relié à une ligne terrestre (5,6). Pour la première fois en 2015, le Conseil de la radiodiffusion et des télécommunications canadiennes a indiqué qu'il y avait plus de ménages canadiens qui utilisaient exclusivement des services sans fil mobiles (20,4 %) qu'exclusivement des téléphones filaires (14,4 %) et

que plus de ménages avaient au moins un téléphone cellulaire (84,9 %) qu'un téléphone filaire (78,9 %) (http://www.crtc.gc.ca/fra/publications/reports/policymonitoring/2015/cmr2.htm).

En troisième lieu, nous voulons accroître l'intérêt pour les mégadonnées parce que nous craignons que les systèmes traditionnels de surveillance ne soient dépassés par les maladies émergentes et réémergentes telles que le SRMO, le SRAS, le virus Zika, le virus Ebola, la tuberculose et la résistance antimicrobienne pour ne nommer que celles-là (1). Pour répondre à la nécessité de dépister les maladies émergentes, y compris celles qui pourraient être associées au bioterrorisme, de nombreuses compétences d'Amérique du Nord et d'ailleurs ont investi d'importantes ressources pour mettre au point des systèmes de surveillance syndromique. Celle-ci [traduction] « utilise des caractéristiques cliniques qui peuvent être discernées avant que le diagnostic ne soit confirmé ou des activités mises de l'avant dès le début des symptômes pour avertir de changements dans l'activité de la maladie » (7). On pensait que la surveillance syndromique convenait tout particulièrement bien au dépistage des nouvelles maladies en raison de l'importance accordée au dépistage et à la déclaration des symptômes qui peuvent accroître la sensibilité du système pour y inclure des maladies autrefois inconnues. Un grand nombre des outils envisagés sous le nom des mégadonnées peut fournir l'externalisation à grande échelle de l'information sur les symptômes en temps réel, avant le diagnostic clinique, et pour des régions géographiques précises.

En plus du potentiel de multiplication des possibilités de trouver de nouvelles sources d'information grâce aux plateformes de mégadonnées, on reconnaît que l'analyse de ces dernières et les moteurs d'inférence d'apprentissage peuvent offrir des avantages considérables quand on les compare aux processus statistiques utilisés d'ordinaire pour détecter une activité anormale dans la plupart des systèmes de surveillance syndromique existants (8). Par exemple, les méthodes de type bayésien d'élaboration de modèles comprennent des algorithmes qui évalueront la probabilité qu'une maladie survienne vraiment, compte tenu à la fois des données observées et de la probabilité antérieure de la maladie dans le contexte épidémiologique actuel.

En dernier lieu, les mégadonnées pourraient avoir une influence substantielle sur l'efficacité des modèles prédictifs en santé publique. Ces derniers sont des outils indispensables à la compréhension de la transmission des maladies et à la planification de leur gestion. Toutefois, l'application de mégadonnées à la modélisation dynamique dans le cas des maladies infectieuses en est encore à ses balbutiements, si l'on compare à d'autres domaines tels que le marketing et les prévisions météorologiques (1). La qualité et la quantité de l'information disponible pour établir les paramètres nécessaires aux modèles font partie des différences les plus marquées entre les modèles de prédiction des maladies infectieuses et les modèles de

prédiction des événements météorologiques (9). Les modèles des maladies infectieuses sont particulièrement sensibles aux paramètres des taux de contact et à l'hétérogénéité dans les contacts entre les groupes et à l'intérieur même des groupes. Alors que depuis toujours, la plupart des modèles ont été limités par l'hypothèse d'un mélange aléatoire dans les groupes et axés uniquement sur les mélanges entre les grands centres démographiques (10), il existe de nouveaux outils de mégadonnées qui peuvent fournir des données détaillées sur les mouvements aux modèles de maladies infectieuses et tenir compte de l'hétérogénéité des contacts à l'intérieur des populations. Ces données précises sur les mouvements permettront de mesurer les taux variables de contacts entre des personnes infectées et susceptibles de l'être, en réponse à l'attention médiatique et aux changements qui en découlent dans les préoccupations du public (10).

Objectifs et stratégie d'élaboration du rapport

Devant le potentiel des « mégadonnées » de constituer une « ressource inexploitée de faits probants qui peuvent servir à orienter l'élaboration des politiques et la prise de décisions », le Centre de collaboration nationale des maladies infectieuses (CCNMI) a commandé un document de travail pour lancer les discussions entre les professionnels de la santé publique au Canada. Le document avait pour principal objet d'examiner si les mégadonnées peuvent contribuer à la politique de santé publique en ce qui a trait à la gestion des maladies infectieuses au Canada. La première partie du rapport présente les mégadonnées et diverses sources de ces données pour la surveillance des maladies infectieuses, décrites dans la littérature évaluée par des comités de lecture et des rapports techniques d'accès public jusqu'au 31 décembre 2016. La plupart des documents examinés portent sur l'Amérique du Nord. Les exemples concernant un contexte canadien ont été inclus et mis en lumière lorsqu'il était possible de le faire. Le rapport porte également sur les options et les difficultés à visualiser et à analyser l'information qui en découle pour qu'elle puisse être utilisée plus efficacement dans la prise de décisions.

Bien que de nombreuses publications à ce jour sur l'utilisation des mégadonnées dans le contexte des maladies infectieuses portent sur l'influenza, le virus Ebola ou les maladies à transmission vectorielle, l'utilisation possible des mégadonnées pour faciliter la surveillance des maladies d'origine alimentaire et les enquêtes sur les éclosions a néanmoins été ciblée comme thème possible des discussions futures. Le présent rapport souligne les types de mégadonnées appliquées à la surveillance des maladies d'origine alimentaire et aux enquêtes les concernant, selon les descriptions contenues dans la littérature, présente le commentaire d'un chercheur qui travaille activement à l'application des mégadonnées aux maladies d'origine alimentaire

et, finalement, résume les résultats d'un sondage non officiel sur les mégadonnées auprès des professionnels dont le travail a trait aux maladies infectieuses et aux enquêtes sur les éclosions au Canada.

Le présent rapport constitue une introduction et un aperçu de base. Même si l'auteure et ses collaborateurs ont fait tous les efforts possibles pour effectuer une recherche systématique de la littérature existante, il ne s'agit pas d'une exploration exhaustive ni d'un résumé de tous les aspects des mégadonnées et des maladies infectieuses ou des mégadonnées et des maladies d'origine alimentaire. Le rapport ne vise pas non plus à fournir des recommandations précises sur les étapes suivantes à mener pour adopter les mégadonnées dans les enquêtes sur les éclosions. Toutes les opinions exprimées sont celles de l'auteure et de ses collaborateurs, à moins d'indication contraire.

Les documents résumés dans le présent rapport proviennent d'une recherche systématique dans PubMed, Scopus et ProQuest Public Health au moyen des termes « mégadonnées » et (« maladies infectieuses » ou « enquête sur les éclosions » et leurs synonymes). Nous avons fait suivre la recherche dans les bases de données d'une sélection séquentielle des titres, des résumés et, finalement, des rapports complets au contenu pertinent. Nous avons également examiné les 200 premières mentions dans Google avec les mêmes termes de recherche pour s'assurer de l'inclusion des travaux les plus récents et peut-être inédits, de même que de la littérature grise pertinente. Nous avons ensuite mené un examen des bibliographies des documents principaux pour déterminer l'information importante qui était passée inaperçue. Nous avons également utilisé les options dans les bases de données PubMed, Scopus et ProQuest pour repérer les documents plus récents qui avaient ensuite cité les documents choisis dans la recherche initiale ou qui ont été jugés semblables à ceux que nous connaissions déjà. Nous avons effectué d'autres recherches avec les termes « mégadonnées » et « visualisation » et « maladie », de même que « mégadonnées » et « maladie d'origine alimentaire » et leurs synonymes.

UNE INTRODUCTION AUX MÉGADONNÉES

En affaires, les mégadonnées ont été très fructueusement utilisées pour proposer la vente de produits spécifiques à des clients individuels d'entreprises comme Amazon et Netflix et pour orienter la publicité et l'ordonnancement des résultats de recherche pour Google. Les mégadonnées ont également servi dans un certain nombre de domaines scientifiques, par exemple pour accroître l'exactitude des prévisions météorologiques à court et à moyen terme. Les soins de santé ont mis plus de temps à profiter de la qualité sans cesse croissante des données produites par les nouvelles technologies. L'exception la plus remarquable, en ce qui concerne la gestion des maladies infectieuses, a été l'utilisation du séquençage pangénomique et de la bio-informatique en virologie et en microbiologie. On incite les professionnels de la santé publique à mieux intégrer les mégadonnées à la surveillance des maladies et aux enquêtes sur les éclosions de façon à utiliser les faits probants de plus en plus connus pour orienter les décisions et améliorer l'efficacité et l'efficience des programmes de santé publique. Par ailleurs, comme les épidémiologistes et les biostatisticiens travaillent déjà avec de grands ensembles de données, de nombreux praticiens de la santé publique peuvent se demander comment les mégadonnées se distinguent des sources d'information traditionnellement indispensables à la surveillance des maladies et aux enquêtes connexes.

Un certain nombre de facteurs caractérisent les mégadonnées, mais les trois caractéristiques les plus souvent citées sont les suivantes : le volume, la vélocité et la variété des données (1,11). Le volume a trait à la taille réelle des données, compte tenu du fait que la taille absolue désignant les mégadonnées peut varier considérablement d'une discipline à une autre (1). Links a indiqué qu'on pouvait considérer que les données étaient « méga » lorsque leur taille devient si considérable qu'elle entrave la transformation des données en information qui peut être traitée au moyen des outils logiciels épidémiologiques et statistiques traditionnels (11). Le volume pourrait refléter un grand nombre de sujets ou d'observations par sujet ou les deux. Les grandes bases de données en épidémiologie récemment mesurées en gigaoctets (10^9 octets) en sont un exemple et il y a de nombreux types de mégadonnées qui sont le plus souvent indiqués en téraoctets (10^{12} octets), en pétaoctets (10^{15} octets) ou dans certains cas, en zettaoctets (10^{21} octets ou 1 billion de gigaoctets) et plus. Les bases de données dans certaines de ces catégories dépassent la capacité de stockage des disques rigides même les plus grands pour les ordinateurs personnels.

En plus du volume, les mégadonnées se caractérisent également par leur vélocité et leur variété. La vélocité peut désigner la vitesse à laquelle les données sont colligées ou transmises et reçues. De nombreux types de mégadonnées sont colligés en temps quasi réel, et elles peuvent être consultées en continu au lieu d'être

mises en lots avant un transfert périodique. En dernier lieu, les mégadonnées peuvent avoir diverses sources et structures organisationnelles (12). Par exemple, les mégadonnées pourraient contenir des données numériques structurées, du texte non structuré, des fichiers PDF, des courriels, des enregistrements vocaux, des images et des données d'appareils mobiles (13). En raison de la taille, du rythme de croissance et de leur complexité, les mégadonnées peuvent exiger un effort substantiel et des ressources informatiques pour nettoyer, fusionner, relier, jumeler, transformer, analyser, explorer et stocker (13). Voici d'autres qualificatifs des mégadonnées : véracité (reflet de l'imperfection, de l'incomplétude ou de l'absence de fiabilité des données); validité (exactitude des données); volatilité (à quel point la donnée devient obsolète); et variabilité (problèmes d'incohérence et d'imprévisibilité des flux de données) (13,14).

Un aperçu des sources possibles de mégadonnées donne une idée de ce que chacune de ces sources pourrait faire en santé publique et en particulier dans les enquêtes sur les éclosions de maladies d'origine alimentaire. Cinnamon et coll. décrivent trois types de mégadonnées (15). Les résultats de données dirigés, lorsque la technologie est capable d'enregistrer des données (p. ex. des caméras de surveillance, la technologie de télédétection par satellite), sont axés sur une personne ou un endroit. Les données automatisées sont colligées de manière passive par la production d'enregistrements électroniques découlant du fonctionnement normal de systèmes ou de technologies tels que les téléphones cellulaires, les navigateurs Web, les transactions de cartes de crédit ou les programmes de fidélité des consommateurs. En dernier lieu, les données spontanées sont produites de manière active ou passive par les citoyens par le truchement de plateformes comme les médias sociaux et les applications d'externalisation ouverte (15). Les données actives découlent des contributions interactives et intentionnelles à l'infrastructure numérique pour les plateformes de participation et les sciences citoyennes (1). D'autres classifications proposées pour les mégadonnées en santé comprennent les suivantes : les données des consultations médicales, les données participatives sur les syndromes et les données numériques sans lien avec la santé telles que les réseaux de contacts, les modèles de voyage, l'acceptation des vaccins et les choix alimentaires (1).

LES SOURCES POSSIBLES DE MÉGADONNÉES POUR LES MALADIES INFECTIEUSES

i. Séquençage pangénomique et bio-informatique

L'analyse des données du séquençage pangénomique est l'une des applications les plus largement reconnues des mégadonnées en surveillance des maladies infectieuses et dans les enquêtes connexes. Le séquençage pangénomique peut être utilisé pour inclure ou exclure des cas des enquêtes particulières sur des éclosions ou

des sources suspectes pour orienter les mesures de lutte contre les infections et les messages de santé publique (16). Le séquençage pangénomique, accompagné des analyses bio-informatiques pertinentes, peut aussi être utilisé pour prévoir l'existence de cas non diagnostiqués et intermédiaires dans les chaînes de transmission, et pour en déduire la direction que prendra la transmission (16). Par exemple, l'analyse phylogénétique des données du séquençage pangénomique de *Salmonella typhimurium* DT104 dans les populations humaines et animales au Royaume-Uni a montré que la transmission entre les espèces était limitée (16,17).

La bio-informatique implique l'analyse des variations dans des systèmes biologiques à l'échelle moléculaire (18). L'analyse phylogénétique et les méthodes semblables d'analyse évolutive peuvent être utilisées dans certains cas pour déduire l'origine et l'émergence de certains agents pathogènes, évaluer des sources possibles de maladie et déterminer les chaînes de transmission les plus probables (16). L'analyse de « l'horloge moléculaire » peut parfois être utilisée pour estimer l'époque de l'ancêtre commun le plus récent et les dates possibles des transmissions. L'analyse de l'horloge moléculaire des données du séquençage pangénomique est fondée sur l'hypothèse que les substitutions de bases s'accumuleront à un rythme constant. On s'est servi de cette stratégie pour faire des inférences sur les incidences de transmission dans les éclosions de SARM (16).

Le séquençage pangénomique est de plus en plus utilisé dans la surveillance des maladies d'origine alimentaire et il a donné d'excellents résultats pour ce qui a été d'identifier la source d'éclosions de listériose et pour en limiter l'ampleur; cet organisme a toutefois un génome relativement petit et plus conservé que ceux des bactéries *Salmonella*, *E. coli* et *Campylobacter*. Même si l'interprétation des données du séquençage pangénomique de grands organismes est plus difficile, cette technologie a réussi à distinguer des souches de salmonelle qu'on ne parvenait pas à séparer au moyen de l'électrophorèse en champ pulsé (EPCP) (19). En plus de son application dans les enquêtes sur les éclosions, le séquençage pangénomique a également identifié des sources de cas sporadiques de maladie d'origine alimentaire (20).

Options de gestion des données du séquençage pangénomique et outils d'analyse

Aestrup et coll. ont résumé les plateformes servant actuellement au partage des données du séquençage pangénomique, y compris GenomeTrakr et COMPARE (21). La Food and Drug administration des États-Unis a mis au point GenomeTrakr. Ce programme comprend le séquençage et la comparaison en temps réel des isolats bactériens d'origine alimentaire avec le nombre moyen de séquences ajoutées mensuellement, et

leur nombre est rapidement passé de 169 en 2013 à 4 529 en 2016 (www.fda.gov/Food/FoodScienceResearch/WholeGenomeSequencingProgramWGS). Les données de séquençage sont assemblées, analysées et stockées au National Center for Biotechnology Information (NCBI). Les outils à la disposition du public au NCBI permettent de produire les arbres phylogénétiques de base, mais les analyses plus poussées doivent être faites localement. Les 50 États devraient participer à GenomeTrakr d'ici 2019 (19).

COMPARE Europe est un site unique de partage et d'analyse des données provenant de bactéries, de parasites et de virus qui comprend des génomes simples et des projets métagénomiques (21). Les concepteurs ont reconnu qu'il fallait des outils de bio-informatique faciles à utiliser et qu'il était difficile de télécharger de gros volumes de données pour l'analyse locale. COMPARE permettra aux chercheurs d'utiliser de nouveaux outils pour les données et il a été configuré pour gérer le partage sécurisé des données (www.compare-europe.eu). Des outils d'analyse plus perfectionnés sont constamment en cours d'élaboration, entre autres des outils qui permettraient de combiner les données du séquençage pangénomique et des données épidémiologiques spatiotemporelles dans des modèles bayésiens d'inférences uniques et de construire des arbres plus efficaces de la transmission. Jombart et coll. ont relaté ce type d'analyse à l'aide de données provenant de l'éclosion de SRAS en 2003 à Singapour (22). Le Centre d'épidémiologie génomique du Danemark a également mis au point un certain nombre d'outils bio-informatiques grâce auxquels identifier des espèces, le typage par séquençage multilocus (méthode MLST (multilocus sequence typing)), les plasmides, la virulence, la résistance antimicrobienne (RAM), les sérotypes et des outils phylogénétiques (23).

IRIDA (Integrated Rapid Infectious Disease Analysis – analyse intégrée rapide des maladies infectieuses) (http://www.irida.ca/), une plateforme de source ouverte et gratuite, est à la disposition des chercheurs canadiens qui doivent gérer des données génomiques (https://www.slideshare.net/IRIDA_community/irida-immemxi-hsiao). IRIDA utilise Galaxy pour gérer l'ordonnancement des tâches. Les participants en sont l'Agence de la santé publique du Canada (ASPC), les organismes provinciaux de santé publique et certaines universités. IRIDA propose des liens vers d'autres outils qui facilitent l'identification des îlots génomiques qui peuvent coder la RAM et les facteurs de virulence (IslandViewer) et la cartographie (GenGIS). On peut voir un aperçu de la phylogéographie avec GenGIS, élaborée à l'Université Dalhousie, à https://www.slideshare.net/beiko/gengis-presentation-at-vizbi-2016. D'autres outils sont nécessaires pour mieux identifier les grappes basées sur le séquençage pangénomique et les métadonnées épidémiologiques (source, lieu, moment). EpiQuant est un exemple d'outil mis au point par l'ASPC pour améliorer

l'application des données génomiques en épidémiologie (https://www.slideshare.net/IRIDA_community/hetman-immem-xi-final-march-2016).

Difficultés de l'analyse et de l'interprétation des données du séquençage pangénomique

En avril 2016, une réunion mixte de deux jours a été organisée pour discuter de l'utilisation du séquençage pangénomique pour protéger la santé publique et améliorer la sécurité alimentaire aux États-Unis. Cette rencontre a réuni des représentants de l'industrie, des organismes de réglementation et des chercheurs (www.uspoultry.org/foodsafety/docs/WGS_Meeting_Summary_072916-02.pdf). Cette réunion a donné un aperçu de la façon dont le séquençage pangénomique est actuellement utilisé et les limites de cette technologie. Ni la technologie du séquençage pangénomique ni aucune technique individuelle ne parvient, à elle seule, à établir la cause d'une éclosion; bon nombre des participants ont plutôt insisté sur le fait qu'une enquête doit s'appuyer sur une combinaison d'épidémiologie, de données du traçage en amont, d'évaluations environnementales et de microbiologie. Les participants ont aussi prévenu que même si le séquençage pangénomique peut fournir de l'information actuellement obtenue par d'autres méthodes telles que la spéciation, le sérotypage, la présence de facteurs de virulence et des déterminants de la résistance, on n'observera pas dans toutes les espèces et tous les environnements la même virulence phénotypique ni les mêmes gènes de RAM (20).

La similarité entre les génomes d'organismes se mesure à l'aide de l'analyse des SNP (polymorphisme nucléotidique) et le wgMLST (typage par séquençage multilocus pangénomique) (20). Il est toutefois difficile de normaliser et de consigner ces techniques. Par exemple, le dénombrement des SNP peut se faire de plusieurs façons. Pour que le compte des différences de SNP soit correctement interprété, il faut savoir comment ils ont été dénombrés (20). La parenté est souvent illustrée à l'aide d'un arbre phylogénétique. Les inférences sur les voies de la parenté entre les organismes peuvent toutefois être fondées sur un certain nombre de méthodes sous-jacentes (parcimonie, probabilité maximale, méthodes bayésiennes et distance) et la confiance à accorder aux arbres qui en résultent doit être indiquée à l'aide des statistiques pertinentes (20). Les algorithmes fondés sur la parcimonie identifient l'arbre qui exige le plus petit nombre de SNP (23). Les méthodes fondées sur la probabilité maximale donnent l'arbre le plus susceptible de ressortir des données observées. Les méthodes bayésiennes identifient l'arbre possédant la plus forte probabilité postérieure, compte tenu des données observées. Finalement, les méthodes fondées sur la distance donnent l'arbre qui représente le mieux les différences dans les SNP ou celui dont les branches sont les plus courtes (23). Différentes méthodes mèneront souvent à des différences d'interprétation.

Les détails sur les façons de faire des distinctions utiles entre les isolats dépendent des espèces (20). L'évaluation des différences peut être influencée par les changements évolutifs survenus dans l'organisme au fil du temps (SNP prédits toutes les >2 000 générations). Les modifications des SNP peuvent être influencées par le passage dans différents hôtes ou médias dans un laboratoire, la taille du génome et le degré de conservation dans ce dernier, de même que par le fait que des éléments autres que le génome de base sont pris en compte (c.-à-d. des plasmides, des bactériophages) (20).

Pour que le séquençage pangénomique soit uniformément utile, les chercheurs doivent pouvoir interroger rapidement les bases de données du séquençage et déterminer avec exactitude les distances génétiques entre les échantillons. Différentes mesures de la distance, fondées sur des profils k-mer ou des sites nucléotidiques, ont été évaluées dans le cas d'isolats de la salmonelle (24). Les auteurs de l'étude sur la salmonelle ont conclu que les distances fondées sur les sites étaient supérieures (NUCmer et MLST allongée), mais que les ressources informatiques nécessaires pourraient être un facteur limitatif. Il reste cependant des questions à résoudre concernant la façon dont les différents isolats de différentes espèces doivent être évalués pour être considérés comme des souches différentes.

Finalement, il faut poursuivre les recherches pour déterminer la probabilité que la même souche puisse se retrouver dans plusieurs environnements et la variabilité de cette probabilité selon les espèces. Par exemple, si l'on trouve la même souche d'un organisme dans un échantillon d'une source soupçonnée et dans un échantillon diagnostique, les données sont très limitées sur le degré de probabilité qui permet aux chercheurs d'écarter d'autres sources. Il est possible que des isolats identiques sur le plan génétique puissent se trouver à différents endroits (20).

Autres difficultés et limites de l'application du séquençage pangénomique

Il faut normaliser tous les aspects de la méthodologie du séquençage pangénomique et des outils bio-informatiques pour la pratique en santé publique (23). Des chercheurs en bio-informatique ont soulevé d'importantes préoccupations et dit que la méthodologie de traitement des données et d'assemblage du séquençage pangénomique devait être déclarée en toute transparence (19). Un grand nombre des pipelines utilisés pour transformer les données brutes en interprétation biologique ne sont pas du domaine public (p. ex., NCBI). Les mêmes chercheurs ont souligné qu'il fallait connaître le potentiel de faux positifs pour évaluer la parenté et la nécessité d'indiquer le taux de confiance dans les SNP indiqués (19). En plus des

risques de faux positifs attribuables à des facteurs qui influencent le rendement des outils d'analyse, la contamination de l'appareil de séquençage peut aussi engendrer des faux positifs (20).

Nous avons aussi observé, parmi les limites du séquençage pangénomique, que la qualité des données produites à l'aide des plateformes de lecture courte les plus souvent utilisées parce qu'elles coûtent moins cher et sont plus rapides est moindre que la qualité obtenue avec les technologies à lecture longue (19). Les plateformes de lecture longue sont plus susceptibles de donner des séquences génomiques fermées sans « trous » dans le génome. Les technologies de lecture longue produisent des données qui peuvent être assemblées *de novo* sans souche de référence et elles peuvent être utilisées pour détecter les réarrangements génomiques et calculer directement les modèles d'EPCP (20). La qualité de l'identification des SNP et des assemblages génomiques varieront donc selon la méthode initiale de traitement. La qualité de l'identification des SNP varie également selon la méthode d'assemblage, *de novo* ou fondée sur l'alignement sur un génome de référence. Les SNP sont identifiés après l'alignement sur une souche de référence et la qualité de toute inférence dépend alors de la qualité de cette dernière (20). Une autre mise en garde a trait au fait que certaines séquences de référence anciennes, fondées sur la méthode Sanger, peuvent contenir des erreurs d'assemblage (25).

Le séquençage pangénomique exige encore, le plus souvent, la culture des organismes suspects, mais la demande de diagnostics indépendants des cultures augmente (23). L'une des options actuellement envisagées implique la caractérisation directe de tout le matériel génétique à partir du spécimen ou une approche métagénomique. Les difficultés sont multiples lorsqu'on essaie d'extraire de l'information sur des agents pathogènes particuliers du matériel génétique entier. La métagénomique nécessite des bases de données de référence exhaustives pour interpréter les données produites. Des questions éthiques entrent aussi en jeu lorsqu'on gère de l'information recueillie dans l'ADN du patient qui pourrait être présent dans les échantillons diagnostiques (23).

Comme les données du séquençage pangénomique versées dans GenomeTrakr, une base de données publique, deviennent publiques, on craint pour la confidentialité de la source de l'échantillon. Dans les situations d'éclosions très publicisées, même si la source était confidentielle, elle pourrait être facilement prévisible si l'on sait la date, le type d'échantillon, l'organisme et le lieu (19). PulseNet, par comparaison, n'est pas accessible au public. Les problèmes de partage des données liées au versement rapide des données du séquençage pangénomique dans les bases de données publiques sont en outre compliqués par des questions pratiques, des considérations politiques, juridiques et éthiques, des droits de propriété intellectuelle et la volonté de protéger les droits de publication (21). On a proposé d'étiqueter les données pendant de

courtes périodes pour qu'elles puissent être utilisées en santé publique, mais pas par d'autres chercheurs en vue de publications. On a aussi beaucoup discuté d'associer les métadonnées à l'information du séquençage pangénomique (21). On a suggéré que les ensembles de données comprennent à tout le moins le pays, l'année, l'origine, la pathogénicité ou l'indication de la provenance de l'échantillon s'il s'agit d'une infection (21).

Il faut contrebalancer les risques potentiels pour la protection des renseignements personnels des patients (clients, participants) avec les avantages pour le bien public. Depuis toujours, on a eu pour approche de retirer ce qui est perçu comme des données d'identité. Cette façon de faire peut gravement limiter l'utilité des données et ne garantit pas toujours la protection des renseignements personnels, vu la possibilité de faire des liens entre des données provenant de différentes sources. Certains ont proposé, pour protéger la vie privée des patients, de protéger les renseignements personnels en fonction des circonstances (26). Par exemple, les données sociodémographiques pourraient être légèrement modifiées avant leur divulgation pour réduire à un seuil prédéfini le risque pour la protection de la vie privée. D'autres encore ont proposé une approche progressive pour le partage des données génomiques et des métadonnées connexes pour optimiser l'utilité de l'information tout en protégeant la confidentialité (27).

ii. Résumés analytiques généraux des nouvelles

Le Canada est reconnu à l'échelle internationale pour le Réseau mondial d'information en santé publique (RMISP) (28). Ce réseau est un programme mis en œuvre dans le Web qui balaie et examine plus de 30 000 sources mondiales de nouvelles en neuf langues différentes. Plus de 20 000 rapports sont évalués quotidiennement pour y détecter des menaces pour la santé publique. Le programme utilise des agrégateurs de nouvelles qui sont reliés à des journaux nationaux et locaux et à certains bulletins. Les nouvelles examinées ne se limitent pas aux nouvelles sur la santé, elles englobent également les sports, les voyages et les finances. Les balayages sont répétés toutes les 15 minutes et les articles sont traduits et traités en moins d'une minute. Lorsque le système détecte un signal, l'article est revu par une équipe multilingue et multidisciplinaire de professionnels de la santé qui déclenchent une alerte, si les preuves sont suffisantes. Le RMISP est reconnu pour être le premier à avoir détecté l'éclosion de MERS-CoV, de même que les premières activités du SRAS en Chine, en raison de mentions dans les pages financières d'une augmentation des ventes de médicaments antiviraux (28).

L'un des systèmes de surveillance les plus anciens basés sur des incidents est ProMED-mail (Program for Monitoring Emerging Diseases – programme de surveillance des maladies émergentes). ProMED-mail a commencé en 1994 (29) et est affilié à la Société internationale pour les maladies infectieuses. ProMED obtient de l'information des articles dans les médias, des rapports gouvernementaux officiels, des sommaires en ligne et d'observateurs locaux (www.promedmail.org/). Des évaluateurs experts examinent et sondent cette information avant de la diffuser par courriel et dans le site Web. En plus de l'anglais, ProMED est offert en portugais, en espagnol, en russe et en français. ProMED collabore avec HealthMap.

HealthMap est hébergée à l'Université Harvard et est fondée sur les données de ProMED-mail, en plus de l'information de l'Organisation mondiale de la Santé, de GeoSentinel, de l'Organisation mondiale de la santé animale (OIE), de l'Organisation des Nations Unies pour l'alimentation et l'agriculture (FAO), d'EuroSurveillance, de Google News, de Moreover, de Wildlife Data Integration Network, de Baidu News et de SOSO Info (http://www.healthmap.org). Le logiciel s'appuie sur des programmes de source ouverte tels que Google Maps, GoogleMapAPI for PHP, Google Translate API et la bibliothèque xajax PHP AJAX. HealthMap surveille, résume, filtre et cartographie l'information recueillie en neuf langues, 24 heures sur 24. Dans une étude sur 111 éclosions survenues en six mois en 2012, HealthMap a relaté des éclosions à une moyenne de 1,26 jour ($p=0,002$) avant le premier rapport de source officielle. Dans la plupart des cas, l'information relatée était la même dans les rapports officiels et les rapports de HealthMap (30).

En plus de l'utilisation des résumés analytiques des nouvelles comme intrants des systèmes établis de surveillance des maladies tels que le RMISP, ProMED-mail et HealthMap, des requêtes personnalisées ont été élaborées pour extraire de l'information sur des questions précises. À l'aide des comptes rendus médiatiques sur l'épidémie d'Ebola en 2014-2015 et l'éclosion du SRMO en 2015, Chowell et coll. ont pu extraire de l'information des rapports décrivant les grappes épidémiologiques et estimer des chiffres de reproduction et les tendances temporelles des éclosions très semblables à ceux qui avaient été calculés au moyen des données officielles de surveillance (31). Les auteurs ont reconnu que la déclaration de grappes de cas dans les médias était sujette à plusieurs types possibles de biais.

Parmi les limites reconnues aux données des résumés analytiques des nouvelles, citons la possibilité que les reportages soient axés sur des histoires sensationnelles de survivants, sur la population américaine et les grappes de cas importants (31). Les chercheurs ont également signalé que l'âge et le sexe étaient absents de nombreux résumés. Compte tenu des limites possibles de l'information publiée dans les médias, de nouvelles sources d'information pour les plateformes telles que le RMISP ont été envisagées, notamment les stratégies

de recherche dans Internet, les médias sociaux, les téléphones intelligents et d'autres systèmes actifs d'externalisation ouverte à grande échelle et volontaire (28).

D'autres types de plateformes dans le Web, semblables à celles qui sont utilisées dans les résumés analytiques des nouvelles, peuvent aussi servir à regrouper certaines données établies par des laboratoires et des établissements de soins de santé. ResistanceOpen est une application Web mobile qui sert à cumuler, à analyser et à publier des données sur la résistance antimicrobienne (www.resistanceopen.org) (32). Les données publiques sont identifiées à l'aide de requêtes Web et de la consultation d'adresses URL qui ont précédemment fourni des données pertinentes. Il peut falloir, cependant, extraire les données manuellement, une fois que des conservateurs experts les ont identifiées. À l'époque où MacFadden et coll. ont décrit ce système, il y avait des données provenant de 340 endroits de 41 pays, dont huit provinces canadiennes (32). Les résultats du système comprennent une carte de navigation permettant des comparaisons entre les régions. Les auteurs mettent toutefois en garde les lecteurs que les comparaisons sont limitées par les différences dans les pratiques de rapport d'un endroit à un autre et dans les normes utilisées pour classifier le degré de résistance.

La plupart des processus de recherche et de rapports décrits pour les données des médias et des laboratoires exigent actuellement un nombre considérable d'heures-personne d'examen pour évaluer si les incidents et les données recueillis par les moteurs de recherche peuvent être ajoutés au répertoire en ligne. Dans le cas de ResistanceOpen, une nouvelle saisie manuelle des données est également nécessaire. Le RMISP envisage d'examiner de plus près le potentiel des capacités plus perfectionnées de l'intelligence artificielle pour diminuer le nombre d'heures-personne nécessaires à l'examen des résultats avant de cibler un incident (28). Même si de nombreuses nouvelles initiatives viennent des milieux universitaires, étant donné les exigences actuelles en personnel pour le fonctionnement quotidien, en plus de la nécessité des mises à jour permanentes des algorithmes de recherche et de la maintenance des systèmes, les milieux universitaires ne sont peut-être pas le lieu idéal à long terme pour ces types d'activités. Il est difficile pour les chercheurs de justifier un investissement à long terme après la fin du projet initial de recherche (33).

iii. Interrogations dans Internet, historiques de recherche et données de forums

Les personnes qui utilisent Internet laissent des traces de leur activité. Les traces peuvent être classifiées soit comme comportements de recherche sur la santé, ce qui comprend les interrogations en ligne sur un sujet en

particulier, soit comme information que l'utilisateur a intentionnellement partagée dans un blogue ou un forum en ligne (34). En plus des données des forums en ligne traitant précisément de questions de santé, d'autres forums généraux peuvent également contenir de l'information utile. Par exemple, une éclosion de *Campybacter jejuni* à la suite d'un concours de vélo a été liée à la boue sur les participants, grâce aux messages et à des photos dans un forum social en ligne (35).

Comportement de recherche sur la santé et utilisation des moteurs de recherche ou de l'historique de l'accès à des pages Web

Dans un rapport, les auteurs ont utilisé des données de moteurs de recherche pour prévoir la saison du norovirus en Suède (36), d'après les interrogations adressées à un portail gouvernemental officiel de la santé. Le moteur de recherche décrit dans l'étude était Websök, un système mis au point pour analyser les recherches publiques auprès du portail officiel de la santé dans le comté de Stockholm, en Suède. On ne sait pas très bien, par contre, quelle proportion de la population aurait été suffisamment au courant du site Web officiel pour utiliser le portail pour mener ses recherches. Même s'il peut être intéressant de résumer des données de recherche pour des sites Web très précis gérés par les services de santé gouvernementaux, il y a eu beaucoup d'intérêt pour l'examen des données provenant d'outils Internet plus largement utilisés.

L'une des applications des mégadonnées les plus souvent reconnues pour la surveillance des maladies infectieuses a été Google Flu Trends qui a surveillé les modèles de recherches dans Internet, selon le principe qu'un grand nombre des personnes qui utilisent le moteur de recherche peuvent vouloir se renseigner sur les symptômes. Les résultats initiaux de Google Flu Trends ont été prometteurs, mais en février 2013, le logiciel a prévu deux fois plus de visites chez les médecins liées à la grippe que les cliniques et les hôpitaux sentinelles du Centre for Disease Control (CDC) en ont déclarées (37,38). Même si la corrélation de Google Flu Trends et des cas confirmés en laboratoire a été constamment inférieure en ce qui concerne le syndrome grippal (37), les prévisions se sont considérablement améliorées lorsqu'elles ont été ajoutées aux données de surveillance historiques traditionnelles (37). En raison des questions entourant le rendement de l'outil, Google ne publie plus les estimations mises à jour de Google Flu Trends ou de son pendant Google Dengue Trends (www.google.org/flutrends/about/).

L'un des problèmes évoqués avait trait au fait que même si l'algorithme de recherche semblait être régulièrement mis à jour, il ne paraissait pas tenir compte du modèle d'affaires Google pour le classement des recherches recommandées à ses utilisateurs, en fonction du comportement de recherche des autres utilisateurs (38). Les changements permanents sur la plateforme de service Internet limiteraient également la reproductibilité des résultats au fil du temps (38). L'une des principales autres limites des modèles

d'interrogation en ligne était la capacité limitée de tenir compte des différences particulières à l'âge en épidémiologie (39). Les recherches dans Internet peuvent toutefois indiquer l'endroit en raison de l'adresse IP (12).

La surestimation de Google Flu Trends concernant la période de pointe en 2012-2013, déclarée aux États-Unis, n'a pas été observée au Canada (40). Les estimations de Google Flu Trends pour le Canada étaient en corrélation exacte avec les taux de consultation pour un syndrome grippal dans le système sentinelle des médecins, déclarés à ASPC et la positivité à l'influenza A et au rhinovirus établie par le système de surveillance et de détection de virus des voies respiratoires pour toutes les saisons, de 2010 à 2014 (40). Les données de recherche dans Google avaient également auparavant prévu les éclosions de gastroentérite et du norovirus (41,42), de même que l'éclosion de rotavirus chez les jeunes enfants (42).

En plus des données provenant des moteurs de recherche eux-mêmes, certaines pages Web pouvaient également être interrogées pour connaître des statistiques d'utilisation. Les pages de Wikipédia sont souvent indiquées dans les premiers choix par les moteurs de recherche et elles constituent une source d'information souvent utilisée par de nombreuses personnes en Amérique du Nord. Wikipédia publie des données historiques de recherches toutes les heures aux parties intéressées dans divers sites dont http://stats.grok.se/ (34). Les données sont résumées en fonction du nombre de vues des articles, de vues des pages, de fichiers de dénombrement des pages, selon la source. Dans un exemple, les registres d'accès de Wikipédia ont été utilisés pour prévoir avec exactitude dans le futur et l'immédiat, l'influenza et la fièvre de dengue (34). L'information sur les lieux était toutefois d'une exactitude limitée dans l'analyse parce que l'information sur le pays d'origine était déduite en se fondant uniquement sur la langue de recherche.

Information partagée dans les forums d'Internet

Il existe dans la littérature plusieurs exemples de situations dans lesquelles les forums d'évaluation en ligne des restaurants ont été utilisés pour estimer la probabilité de faibles notes aux inspections, le risque de maladies d'origine alimentaire ou d'information supplémentaire pour les enquêtes sur les éclosions. L'un des sites d'évaluation des restaurants les plus souvent cités est Yelp (www.yelp.ca).

Kang et coll. ont comparé l'analyse textuelle des évaluations de Yelp à Seattle à des rapports d'inspection. Les auteurs ont indiqué un taux d'exactitude de 82 % pour ce qui est de distinguer les restaurants où des infractions graves étaient commises des restaurants sans infraction. En plus des indices prédictifs dans le texte des évaluations, les chercheurs ont examiné le nombre d'évaluations, la longueur de ces dernières, le

score moyen, le nombre d'évaluations négatives et les preuves de fausses évaluations. Ils ont conclu que la meilleure information se trouvait dans le contenu textuel des évaluations (43).

Harrison et coll. ont utilisé les évaluations en ligne de Yelp par les clients des restaurants pour identifier les cas non déclarés de maladies d'origine alimentaire à New York en 2012-2013. Les auteurs de l'étude ont, entre autres résultats importants, signalé que seulement 3 % des évaluations décrivant une maladie étaient déclarées au service 311 de la ville (44). Cette étude visait à déterminer les éclosions de maladies infectieuses. Les maladies à très courte période d'incubation étaient exclues des critères de notation. Par exemple, les auteurs de l'étude n'auraient pas retenu la plupart des maladies découlant de l'exposition à des toxines *Staphylococcus aureus* dans la nourriture. L'analyse n'a relevé que trois éclosions non déclarées liées à des restaurants au cours d'une période de neuf mois (44). De nombreuses infractions ont été relevées durant les inspections dans les trois restaurants.

Lorsque le projet de New York a commencé, le personnel des services sanitaires craignait devoir passer beaucoup de temps à enquêter les cas qui en résulteraient (44). Il fallait du personnel possédant des compétences spécialisées en programmation pour lire les évaluations, envoyer les courriels, interviewer les personnes qui avaient fait les évaluations et effectuer des inspections de suivi. Dans ce projet, le personnel a évalué manuellement les scores les plus élevés aux critères. Les données de Yelp sont publiques dans son site Web, mais Yelp a fourni aux chercheurs des données en format xml concernant New York pour améliorer l'efficacité du processus. Les cas ont été définis d'après la présence dans l'évaluation des mots « malade » ou « vomissement » ou « diarrhée » ou « empoisonnement alimentaire », lorsque deux personnes ou plus avaient été malades et que le temps d'incubation était ≥10 heures (44). Les chercheurs ont ensuite créé des comptes pour envoyer des messages privés dans Yelp aux personnes dont les évaluations avaient été retenues. Rien ne garantissait cependant que la personne qui avait initialement fait l'évaluation allait voir message parce qu'elle devait de nouveau ouvrir une session dans Yelp pour le voir. Cette nécessité a accru le délai de suivi et peut avoir fait diminuer le taux de réponse (44).

Les auteurs de l'étude ont conclu que l'analyse des évaluations de restaurants pouvait faire ressortir des éclosions de petite envergure et ponctuelles que les outils traditionnels de surveillance ne relèveraient pas (44). Pour améliorer encore plus la sensibilité du système, ils ont proposé de demander à Yelp d'inclure un lien vers le site Web de déclaration des services sanitaires locaux, d'ajouter des données d'autres sites Web d'évaluation et de faire passer les mises à jour d'hebdomadaires à quotidiennes (44).

D'autres cas probants indiquent que la déclaration de maladies d'origine alimentaire dans les évaluations de restaurants révèle effectivement les aliments visés dans des rapports officiels d'éclosions de maladies d'origine alimentaire. Nsoesie et coll. ont examiné des rapports de maladies et d'aliments visés dans les évaluations de Yelp de 5 824 restaurants de 29 villes (45). La distribution des aliments visés dans Yelp était très semblable aux aliments indiqués dans les données du CDC. Les données avaient toutefois d'importantes limitations. Seulement 17 % des évaluations indiquant une maladie en précisaient la date. La source d'infection pouvait aussi être mal attribuée par son déclarant parce que le public ne connaît pas les différentes périodes d'incubation des différents types d'agents pathogènes (45).

L'examen des évaluations en ligne a pour objet général d'aider les systèmes traditionnels de surveillance par une déclaration en temps quasi réel des maladies d'origine alimentaire et d'améliorer la reconnaissance des éclosions de ce type de maladie. Par exemple, selon une étude réalisée récemment en Colombie-Britannique, la durée médiane d'une enquête sur une éclosion était de 36 jours (2). Les évaluations de Yelp indiquaient bien l'emplacement du restaurant et l'analyse des évaluations a fait ressortir des cas qui ne seraient pas ressortis dans la surveillance courante. Seulement 1,6 % des personnes qui avaient déclaré une maladie avaient consulté leur médecin, et 11 % des rapports de maladies visaient plus d'une personne (45). Même si l'accent mis sur les évaluations en ligne des restaurants comme outil pour atténuer les maladies d'origine alimentaire n'englobe pas directement les repas préparés à la maison, ces efforts se justifient en partie par la constatation qu'environ 44 % des éclosions dont il est fait état dans les ensembles de données sur les aliments du CDC étaient des cas soupçonnés ou confirmés d'aliments consommés dans des restaurants (45).

Dans une étude ultérieure, Schomberg et coll. ont élaboré un modèle fondé sur 71 360 évaluations de restaurants de San Francisco dans Yelp et qui, au cours de la phase pilote, avait prévu des infractions au code sanitaire dans 78 % des 440 restaurants qui ont reçu des avis d'infraction grave au code sanitaire (46). Ce modèle différait des travaux de Kang et coll. parce qu'il visait à déterminer préférablement les infractions au code sanitaire qui augmentaient les risques de transmission de maladies d'origine alimentaire (43). Lorsque ce même modèle a plus tard été appliqué à 1 542 autres restaurants de San Francisco, le modèle a donné une sensibilité de 91 %, une spécificité de 74 %, une surface sous la courbe de 98 % et une valeur de prédiction positive de 29 % (prévalence de 10 %) pour les infractions graves au code sanitaire observées. Le même modèle appliqué aux données de New York a donné une surface sous la courbe de 77 % et une valeur de prédiction positive de 25 % (prévalence de 12 %). L'exactitude moindre peut s'expliquer par les différences géographiques dans la langue utilisée dans les évaluations des restaurants. Le modèle était le plus exact lorsqu'on utilisait les meilleures évaluations de Yelp. Les « vedettes » de Yelp étaient étroitement associées

aux notations du code sanitaire. On a constaté sans surprise que le mot « vomissement » était le plus fortement associé à une faible notation du code sanitaire.

Même si les modèles ont fourni de l'information utile, d'autres études ont semé le doute sur l'utilité des évaluations de Yelp à fournir des renseignements comparables pour tous les types d'établissements alimentaires de détail. Par exemple, dans un autre rapport récent, les évaluations de Yelp pour New York ont semblé être en corrélation avec la propreté des lieux dans les établissements appartenant à des chaînes, mais pas dans les établissements qui n'étaient pas des franchisés (47). Schomberg et coll. ont dit qu'on s'attendrait à de meilleurs résultats des évaluations en ligne dans les grandes villes où la participation aux plateformes d'évaluation en ligne serait supérieure (46). Les auteurs ont reconnu qu'il était possible que des symptômes de maladies entériques découlant d'autres causes puissent être attribués à tort à des maladies d'origine alimentaire. Finalement, les auteurs se sont demandé si les clients qui faisaient des évaluations pouvaient détecter des problèmes que les inspecteurs de santé publique ne voyaient pas parce que dans de nombreux cas, ils fréquentent le restaurant plus souvent que ne le font les inspecteurs, une ou deux fois par année. Les auteurs ont également souligné que les personnes qui font des évaluations consommaient dans les faits les produits offerts par les restaurants, comparativement aux inspecteurs qui font des observations de l'installation et de ces méthodes de travail (46)

iv. Médias sociaux

Les médias sociaux offrent une autre occasion intéressante de mobiliser « l'intelligence collective » du public pour améliorer le dépistage précoce des maladies infectieuses et la lutte contre ces maladies (35). La plupart des études des plateformes de médias sociaux dont nous parlons s'appuient sur l'exploration de textes non structurés pour y relever des données de surveillance passive qui reflètent des symptômes ou les opinions publiques concernant les vaccins ou la sécurité alimentaire (48). Notre examen systématique de l'utilisation des médias sociaux pour la surveillance et la gestion des éclosions a montré que les deux plateformes les plus souvent étudiées ont été Twitter et Facebook (48). Pour utiliser efficacement les médias sociaux en surveillance, il serait important de comprendre les préférences des plateformes pour ce qui est des données démographiques qui présentent un intérêt pour un objectif particulier. Charles-Smith et ses collaborateurs ont noté, dans leur examen systématique, que dans les articles datant de février 2013, les jeunes étaient les plus susceptibles d'utiliser Facebook, alors que les adultes utilisaient davantage Twitter (48).

Facebook

Les études qui ont utilisé les données de Facebook étaient moins courantes que celles qui ont utilisé les données de Twitter (48). La plupart des études fondées sur Facebook avaient axées sur les comportements à risque de maladies chroniques. Les « j'aime » de Facebook prédisent, selon les constatations, de nombreux résultats et comportements en santé, avec une précision analogue à celle des données du Behavioral Risk Factor Surveillance System – un sondage téléphonique permanent fait à l'aide de numéros choisis aléatoirement (49). Étant donné que les « j'aime » de Facebook pourraient être moins sujets aux fluctuations sur de courtes périodes que les sentiments exprimés dans les gazouillis, il n'est pas étonnant que Facebook puisse mieux convenir aux recherches sur les attitudes et les comportements plus stables qui pourraient être liés à des maladies chroniques ou à des expositions permanentes. Gittleman et coll. se sont en outre questionnés sur la transparence des données extraites et partagées, compte tenu de la structure que donne Facebook à ses catégories de « j'aime », et ils se sont demandés si cette catégorisation pouvait influencer les résultats de l'analyse des données (49). Aucune étude précise ne semble avoir été faite sur la question de savoir si les « j'aime » de Facebook étaient en corrélation avec le risque d'exposition à des maladies d'origine alimentaire.

Twitter

Twitter est une plateforme en ligne de réseautage qui permet à des usagers inscrits de publier de brefs messages et de réagir aux messages ou gazouillis publiés par d'autres. N'importe qui peut lire ces messages, inscrit ou non au service de Twitter. Les données de Twitter peuvent être consultées par une interface de programmation applicative ouverte (API) permettant à des tiers de consulter les données de Twitter en temps réel sur leur propre application (33). Les données de Twitter sont uniques lorsqu'on les compare à de nombreuses autres sources parce que les gazouillis sont limités à 140 caractères. D'autres limites possibles comprennent des biais pour les personnes jeunes plus orientées vers les technologies et le fait que l'information sur l'emplacement est limitée aux endroits où les usagers de Twitter choisissent d'envoyer leurs gazouillis (50). L'emplacement n'est inscrit que si la personne publie un gazouillis et quand elle le fait. On s'est servi de Twitter pour étudier l'influenza dans diverses régions, ainsi que des éclosions de choléra, d'*E. coli* et la dengue (48).

Certaines études fondées sur Twitter sont concentrées sur la détection de l'activité d'une maladie en particulier. D'autres sont axées sur la détection des symptômes, puis l'interprétation de ces derniers par rapport à l'activité de la maladie (51). En plus de la mesure de l'intensité de l'activité de la maladie, des éclosions peuvent être identifiées d'après les données de Twitter par la détection de grappes spatiales, des

méthodes d'analyse des réseaux sociaux, l'analyse des modèles de communication et l'identification de mots clés importants par leur signature spatiale (51).

Twitter a aussi servi à comprendre les modèles de mobilité humaine (50) et comme approximation pour estimer les taux de contact et les modifications dans ces derniers pendant une éclosion de maladie. Les gazouillis géomarqués offrent des données de meilleure résolution sur l'emplacement que les enregistrements détaillés des appels (EDA) obtenus des fournisseurs de cellulaires (50). La résolution peut être à moins de 10 mètres comparativement à la résolution de kilomètres pour la plupart des EDA. Le contenu textuel des gazouillis pourrait être utilisé pour déduire l'emplacement des gazouillis qui ne sont pas géomarqués (50). Par exemple, dans une étude, le champ « emplacement » a été déterminé d'après le profil d'utilisation de Twitter, de même que les interfaces de programmation d'applications (API) de GoogleMaps (52).

Pour que les données de Twitter soient les plus utiles en surveillance, il faut pouvoir cartographier les résultats et de les normaliser sur le plan géographique en reliant les gazouillis à la population sous-jacente. Pour réaliser une recherche géographique, la base de données de Twitter exige un point médian et un rayon de latitude et de longitude pour définir un tampon à inclure dans la recherche. Ce tampon peut être superposé et rattaché spatialement à une couche contenant des données des districts de recensement pour obtenir des renseignements sur la population à risque (53).

Il faut également des algorithmes perfectionnés d'apprentissage automatique qui peuvent apprendre à reconnaître les cas réels de la maladie cible (mesurée selon le rappel de l'algorithme ou la sensibilité dans un contexte épidémiologique) (p. ex., venir à bout de cette grippe, médicaments pour la grippe) et classifier correctement les gazouillis qui contiennent des mentions du mot original de la recherche, mais qui ne reflètent pas la maladie de la personne (mesurée selon la précision de l'algorithme ou la spécificité dans un contexte épidémiologique) (p. ex., vaccin contre la grippe, grippe intestinale) (53). Un exemple de ce type d'outil d'apprentissage automatique est une machine à vecteurs de support (MVS). Le langage utilisé dans les gazouillis est dynamique et peut avoir des anomalies géographiques distinctes qui exigent un examen manuel des résultats imprévus et un processus de mise à jour constante des algorithmes de classification (53). Flahault et coll. et Dinov ont, dans une étude, examiné certains des principes de l'apprentissage automatique appliqués à l'utilisation des mégadonnées dans les soins de santé et nous reviendrons ultérieurement dans une section sur les mises en garder de l'utilisation des mégadonnées en surveillance des maladies infectieuses (9,54).

L'utilisation de Twitter suscite de plus en plus d'intérêt en ce qui concerne l'identification des maladies d'origine alimentaire. Dans une étude de délimitation de l'étendue réalisée en 2013, seulement quatre articles de recherches originales portaient sur les médias sociaux et les maladies d'origine alimentaire ou la gastroentérite (35). Par exemple, des gazouillis ont été liés à des éclosions de salmonelle et de norovirus en Allemagne (55). Il y a eu d'autres exemples depuis cette année-là. Dans une étude de cas, les données de Twitter et les algorithmes existants de détection des éclosions ont reconnu une éclosion d'*E. Coli* entérohémorragiques (EHEC) en Allemagne avant MedISys (fondé sur les données des médias) et d'autres systèmes d'alerte rapide (51).

L'absence d'exemples bien documentés d'applications dans la pratique quotidienne des services sanitaires était, jusqu'à tout récemment, l'une des critiques énoncées à l'égard de l'intérêt porté à l'utilisation des médias sociaux pour surveiller les maladies infectieuses. En 2013, le service sanitaire de Chicago et ses partenaires ont lancé Foodborne Chicago (www.foodbornechicago.org/). Le programme visait à cibler les plaintes de maladies d'origine alimentaire dans Twitter, par la recherche des mots « empoisonnement alimentaire » (56). Le programme a relevé 270 gazouillis dont 3 % des auteurs ont consulté un médecin ou les urgences dans les 10 premiers mois. Le personnel a répondu aux gazouillis et inscrit des liens au formulaire de plainte de Foodborne Chicago; 193 plaintes ont été faites à Foodborne Chicago, dont 10 % ont eu besoin de soins médicaux. Les plaintes ont déclenché 133 inspections non annoncées qui ont représenté 7 % de toutes les inspections fondées sur des plaintes pendant cette période. Ces inspections ont permis d'identifier 20 % des infractions très graves et 22 % additionnels des infractions graves. Selon les enquêteurs, un grand nombre de ces cas n'auraient pas été inclus dans les chiffres de la surveillance normale et n'auraient pas suscité d'inspection du service sanitaire. Les inspecteurs craignaient initialement d'être surchargés d'inspections, mais ils ont reconnu les avantages une fois le programme lancé. Le logiciel de source libre est offert à GitHub (56).

Dans le prolongement de ce type de travail, on a élaboré et axé un tableau de bord Web (HealthMap Foodborne Dashboard mis au point à l'Hôpital pour enfants de Boston) pour identifier les gazouillis concernant les maladies d'origine alimentaire partout dans le monde. Ce tableau de bord peut aussi être personnalisé et axé sur un endroit précis. Une étude a décrit l'expérience du service sanitaire de St. Louis (57). Les gazouillis contenant les mots « empoisonnement alimentaire » avec ou sans espace entre les mots ont été retenus. On a répondu aux gazouillis pertinents par le truchement du tableau de bord et augmenté le nombre de déclarations déposées, en plus de celles que créent les mécanismes existants. Sept pour cent des réponses aux gazouillis par le truchement du tableau de bord ont engendré des rapports. Les réponses des

enquêteurs aux personnes qui avaient publié un gazouillis sur un empoisonnement alimentaire comprenaient un message d'empathie, une confirmation des pouvoirs et une demande à la personne de porter officiellement plainte. Les restaurants inspectés à la suite de ces rapports n'étaient pas plus susceptibles de commettre une infraction que les restaurants inspectés à la suite de plaintes du public par d'autres mécanismes. Les chercheurs ont fait ressortir la possibilité d'une interaction en temps opportun avec les personnes qui utilisent la plateforme Twitter.

Le peu de preuves issues de recherches contrôlées qui montrent que ces méthodes fonctionnent dans la pratique fait aussi partie des critiques de l'utilisation des médias sociaux pour la détection des maladies (48). En 2015, à Los Angeles, Sadelik et coll. ont déployé nEmesis, un système d'inspection adaptatif doté d'un algorithme d'apprentissage automatique appliqué aux gazouillis, puis ils ont évalué le système au moyen d'un essai à double insu (58). Chaque endroit signalé par un gazouillis a été jumelé à des sites de contrôle établis d'après le calendrier annuel d'inspections; les contrôles ont été jumelés aussi étroitement que possible à l'emplacement, à la taille, à la cuisine et au type de permis. Dans cet essai, les inspections adaptatives déclenchées par les gazouillis ciblés par nEmesis ont reconnu plus de points d'inaptitude par inspection, de même qu'un plus grand nombre d'établissements qui présentaient un risque important pour la santé, avec une notation de C ou moins, que les restaurants comparables prévus au calendrier des inspections courantes. La capacité du système d'identifier des problèmes dans les exploitations sans permis est l'un des avantages mis en lumière dans le rapport parce qu'ils seraient passés inaperçus avec les inspections ordinaires (58).

v. Téléphones intelligents

Données passives provenant de l'utilisation des téléphones intelligents

L'un des moyens les plus souvent indiqués lorsque les téléphones intelligents sont utilisés pour mieux comprendre la transmission des maladies est la consultation des enregistrements de données d'appel (EDA). Ces données ont été utilisées pour comprendre la transmission spatiale d'un certain nombre de maladies infectieuses ou à transmission vectorielle. Par exemple, les EDA ont été utilisés pour voir les mouvements et les risques de transmission du paludisme au Kenya (59). Un code pour la tour et l'abonné sont inscrits pour chaque appel ou messagerie texte (10). Ces EDA comprennent généralement l'heure à laquelle une communication a été établie, un identificateur unique pour l'appelant, le numéro de téléphone du récepteur, la durée de l'appel, la taille des données transmises et l'emplacement géographique de la tour cellulaire par laquelle l'appel a été acheminé ou a été reçu pour chaque communication (appel ou messagerie texte) fait par

les usagers de téléphones cellulaires (15). En reliant les emplacements des tours sur une carte, il est possible d'examiner les mouvements des téléphones entre les appels.

Les EDA ne sont généralement pas rendus publics. La nécessité de garantir l'anonymat des appelants et d'obtenir l'approbation des organismes de réglementation est très réelle; l'accès aux données n'est fourni que dans le cadre d'ententes négociées (10). De plus, il n'est pas possible d'effectuer un jumelage biunivoque de la maladie et des données sur les mouvements pour des raisons de respect de la vie privée (10). L'agrégation des données aide à protéger la confidentialité dans les régions urbaines et les données agrégées permettent toujours d'évaluer les mouvements de plus longue étendue.

Ces données ont aussi d'autres limites. L'exactitude spatiale et temporelle est liée à la densité des tours et au comportement d'appel. Par conséquent, l'information qui en découle est utile pour les modèles à grande échelle et les déplacements régionaux, mais moins pour les mouvements qui se rattachent une transmission très locale de la maladie. Les EDA ne retiennent pas les mouvements moins étendus que ce que peut fournir la densité de la tour cellulaire dans une région donnée. La qualité des données variera donc selon la densité des tours et, par conséquent, entre les régions urbaines et les régions rurales (10). Finalement, il peut y avoir d'autres erreurs en raison du partage des téléphones ou des cas où une personne possède plus d'un téléphone ou plus d'une carte SIM (15). Par exemple, la personne pourrait avoir une carte SIM dans son téléphone, de même que dans sa tablette ou son portable.

En plus de l'analyse des mouvements entre différents endroits, les méthodes d'analyse des réseaux sociaux sont également souvent appliquées aux EDA pour étudier les liens entre les usagers des téléphones cellulaires (60). Pour modéliser la propagation des maladies infectieuses, il faut comprendre le réseau des contacts. Chaque personne a un certain nombre de contacts qu'on peut décrire comme le degré des nœuds d'une personne dans un réseau. L'hétérogénéité de la distribution du degré dans le réseau peut être un facteur déterminant important de la dynamique de la maladie. Dans une étude, les auteurs ont utilisé les données de surveillance des téléphones cellulaires pour construire un réseau de contacts réalistes et examiner les changements dans ce réseau pendant l'éclosion d'Ebola (60).

Téléphones intelligents, outils de surveillance active

Les téléphones intelligents peuvent aussi être utilisés comme simples outils d'autosurveillance. Par exemple, EbolaTracks, une plateforme basée sur la messagerie texte, a été créée pour l'autosurveillance active des personnes qui ont visité une région où il y a eu une éclosion d'Ebola. On a remis aux participants un

thermomètre et un téléphone cellulaire. On leur a demandé d'envoyer deux fois par jour pendant 21 jours après leur visite dans la région un message pour faire état de leurs symptômes et de leur température (15).

Les téléphones intelligents possèdent un grand nombre de caractéristiques plus perfectionnées encore inexploitées à leur plein potentiel pour la surveillance des maladies dans les recherches publiées jusqu'à maintenant. Les téléphones cellulaires recueillent, stockent et peuvent transmettre des coordonnées GPS qui peuvent ensuite être enregistrées et cartographiées (12). Des données détaillées du GPS et du sans-fil donnent des possibilités inexploitées de connaître les mouvements individuels à petite échelle (10). De plus, les capteurs Bluetooth sur les téléphones peuvent également servir à suivre une proximité physique encore plus précise d'un équipement doté d'un signal Bluetooth ou d'autres téléphones à l'intérieur pour produire des réseaux de contacts perfectionnés. Les historiques des communications des téléphones cellulaires peuvent également servir pour compléter le suivi des contacts épidémiques (61).

Ethica, une application pour téléphones intelligents mise au point à l'Université de la Saskatchewan (www.ethicadata.com) a été utilisée pour acquérir, stocker et analyser des données sur le comportement humain (62,63). Le système peut recueillir des données issues d'une grande variété de capteurs téléphoniques, dont le GPS, le sans-fil, l'accéléromètre, le gyroscope, la température ambiante et l'éclairage, entre autres, en plus d'options de sondage contextuelles et déclenchées par l'utilisateur. Les auteurs d'une étude récente ont évalué la faisabilité d'utiliser l'application Ethica pour recueillir des données sur l'occurrence de maladies entériques et les facteurs de risque de maladies d'origine alimentaire (64). En utilisant des minisondages déclenchés par l'heure et les utilisateurs, ces derniers pouvant entrer des descriptions de repas et tenir des journaux photographiques de leur alimentation, on a recueilli des données pendant 10 semaines sur l'occurrence de maladies entériques et les comportements alimentaires de 96 étudiants universitaires volontaires. Des comptes rendus alimentaires en temps réel recueillis au moyen des caractéristiques de l'application ont été comparés aux données de questionnaires rétrospectifs subséquents en ligne. Ethica s'est avéré un outil efficace pour recueillir des données sur les symptômes entériques et les comportements alimentaires dans une cohorte sentinelle de participants volontaires. Même si l'application avait déjà été utilisée pour recueillir des données historiques sur les aliments (62,63), c'était une première que les utilisateurs déclarent eux-mêmes des symptômes de maladies entériques.

vi. Surveillance dans les pharmacies

Les ventes au détail de produits pharmaceutiques sous ordonnance ou en vente libre peuvent être une source importante de données de surveillance syndromique. Les études réalisées jusqu'à maintenant sur les maladies infectieuses se sont généralement concentrées sur l'influenza, les maladies des voies respiratoires et les maladies entériques.

Les données sur les ventes au détail de médicaments en vente libre en Grande-Bretagne ont, par exemple, été utilisées en surveillance syndromique pour détecter les modèles spatiotemporels de l'activité de l'influenza et déterminer si des changements dans le comportement d'achat étaient liés à des messages de santé publique ou à l'intensité de l'attention médiatique et des préoccupations du public (65). Les produits soumis à la surveillance ont été les remèdes contre la grippe et le rhume chez les adultes, les remèdes pour la grippe et le rhume chez les enfants, les sirops contre la toux, les thermomètres, les produits antiviraux (y compris les gels pour les mains et les lingettes) et les papiers mouchoirs. On s'est servi des ventes hebdomadaires totales et des ventes de lait et de bananes pour prendre en compte d'autres raisons susceptibles d'expliquer les changements dans le volume total, par exemple les parts de marché ou les heures d'ouverture des magasins. Les ventes de produits antiviraux, suivies des ventes de remèdes pour les enfants, étaient les plus étroitement en corrélation avec le nombre de cas. Les ventes au détail n'ont pas été associées aux reportages dans les médias ni à la fréquence des recherches dans Internet.

L'Agence de la santé publique du Canada (ASPC) a fait état d'une étude sur les ordonnances d'antiviraux en Ontario et des produits en vente libre en Nouvelle-Écosse (66). Pour surveiller les maladies des voies respiratoires graves, les données sur les maladies qui ressemblaient à l'influenza étaient transmises à l'ASPC 10 jours environ après le début des symptômes et les données de laboratoire étaient fournies 17 jours plus tard. Les données sur les ventes de produits en vente libre, par ailleurs, étaient transmises à l'ASPC 48 heures après la fin de la transaction et les données sur les ordonnances d'antiviraux ont été étroitement mises en corrélation avec les dates du début de cas confirmés d'influenza et le nombre total de cas confirmés (66). Il y a également eu une association considérable entre les ventes de produits en vente libre et le nombre de cas de virus respiratoire syncytial et le nombre de détections d'autres virus des voies respiratoires (66). Les auteurs de l'étude ont conclu que les données sur les ventes de produits amélioraient à la fois la pertinence et la résolution géographique de l'information de surveillance lorsqu'on la compare aux autres sources de données.

Les auteurs de la même étude ont également examiné les ventes de médicaments en vente libre pour les problèmes gastro-intestinaux et constaté qu'il n'y avait aucune association avec les éclosions de maladies gastro-intestinales en Nouvelle-Écosse pendant la même période (66). Toutefois, la plupart des éclosions au cours de cette période avaient été rapportées dans des résidences telles que des foyers de soins de longue durée, la constatation à laquelle on se serait le moins attendu comme source d'influence sur les ventes de médicaments en vente libre.

Auparavant, les ventes sans ordonnance de médicaments antidiarrhéiques et antinauséeux étaient en corrélation avec l'activité du norovirus dans une collectivité, mais pas avec les autres causes virales (rotavirus), bactériennes et parasites de la diarrhée dans les conditions courantes où aucune éclosion n'avait été observée (67). Le rotavirus atteint surtout les enfants et dans ces cas, on a pensé que les parents étaient plus susceptibles de demander des conseils auprès d'un professionnel de la santé et que pour cette raison, ils étaient moins susceptibles d'utiliser des remèdes en vente libre. L'utilité de l'information sur les ventes semble la mieux dans les situations d'éclosion. Des études antérieures sur des éclosions historiques d'origine hydrique de *Cryptosporidium*, d'*E. coli* et de *Campylobacter* ont montré que les éclosions étaient associées à des augmentations des ventes de médicaments antidiarrhéiques et antinauséeux en vente libre (68).

vii. Surveillance des ventes au détail de produits alimentaires

Les données sur les ventes au détail des produits alimentaires ont également servi dans les enquêtes sur les éclosions de maladies d'origine alimentaire. Par exemple, en 2012, on s'est servi des cartes de fidélité de supermarchés pour déterminer qu'un mélange de fruits surgelés était la source d'une éclosion d'hépatite A en Colombie-Britannique (69). Les auteurs de l'étude ont obtenu l'autorisation de personnes qui avaient fait des achats dans de grandes chaînes alimentaires de publier leurs historiques détaillés d'achats sur trois mois, indiqués dans les dossiers des cartes de fidélité du magasin. Des données additionnelles sur les ventes ont servi à évaluer la proportion de tous les ménages qui avaient pu se procurer ce produit. Dans une éclosion précédente en 2007, un programme de cartes d'économies pour la clientèle avait servi à identifier qu'un basilic biologique était la source d'une éclosion de cyclosporiasis en Colombie-Britannique (70).

Récemment, IBM a créé un système d'analyse spatiotemporelle pour comparer les données obtenues par le balayage des produits vendus au détail dans les épiceries et les endroits où des cas de maladie d'origine alimentaire sont survenus. Selon ce qu'on en a dit, cette méthode a produit une liste des 12 aliments suspects

les plus probables (http://barfblog.com/tags/ibm/). Selon la description, l'algorithme a besoin d'au moins 10 cas déclarés. La méthode utilise des renseignements tels que la durée de conservation des produits, la date probable de consommation et la probabilité qu'un produit donné contienne un certain agent pathogène. La méthode a été utilisée en Norvège lors d'une éclosion comptant 17 cas confirmés *d'E. coli* pour cibler une liste de 10 produits alimentaires soupçonnés parmi lesquels une saucisse contaminée a subi des tests en laboratoire (71). Dans une étude précédente, des chercheurs ont utilisé les données de balayage des produits d'épicerie vendus au détail en Allemagne et des données spatiales pour montrer que la méthode pouvait être utilisée pour réduire la liste des aliments suspects et accélérer les premiers stades d'une enquête (72,73)

viii. Exploration des dossiers de santé électroniques à fort volume

On peut se demander, devant la disponibilité de plus en plus grande des dossiers médicaux électroniques, si ces renseignements devraient être accessibles en surveillance des maladies et dans quelle mesure. Les dossiers de santé électroniques peuvent être les dossiers des cabinets des médecins, les réclamations aux assurances, les dossiers de congé de l'hôpital et les certificats de décès (1), en plus des résultats des tests de laboratoire et d'imagerie. On s'est servi aux États-Unis des formulaires détaillés de réclamation d'assurance des médecins pour élaborer un modèle détaillé de la transmission locale de l'influenza (3). Les renseignements de patients à l'échelle individuelle permettraient de mieux définir les prédictions selon les différences de risque fondées sur l'âge et la comorbidité (3).

Aux États-Unis, ESPnet est né en 2007 et surveille en temps réel la santé publique au moyen des dossiers de santé électroniques pour appuyer la surveillance locale de la santé publique et les interventions connexes au Massachusetts (74). Ce réseau comprend la détection automatisée le jour suivant des maladies infectieuses à déclaration obligatoire et les rapports cumulatifs des états d'intérêt tels que le syndrome grippal. Les données réelles sont conservées dans les pratiques sentinelles. Les serveurs participants sont configurés pour accepter les demandes de renseignements précises du système ESPnet. Le système est configuré de telle sorte que le personnel de la santé publique peut présenter des demandes sans nécessairement posséder une connaissance approfondie de la programmation. Toutefois, pour que les résultats soient utiles, il est important de reconnaître les limites des données, les facteurs qui peuvent changer rapidement et qui sont susceptibles d'influencer la qualité des données.

Bien que les médecins et les hôpitaux canadiens se convertissent au système de surveillance des dossiers médicaux électroniques, le système canadien de pratique sentinelle, soit le programme Surveillance de l'influenza actuel, n'est pas fondé sur l'accès direct à ces dossiers de santé électroniques. Il exige plutôt la saisie manuelle du nombre total de visites des patients et du nombre total de cas de syndrome grippal vu par groupe d'âge une journée donnée, chaque semaine. Il existe toutefois un Réseau canadien de surveillance sentinelle en soins primaires ou RCSSP qui est axé sur les maladies chroniques et la santé mentale (http://cpcssn.ca/sentinel/potential-sentinels/). Les fournisseurs de soins primaires participants (p. ex. les médecins de famille) donnent accès à leurs systèmes de dossiers de santé électroniques. Le programme est axé sur cinq maladies chroniques et maladies mentales, soit l'hypertension, l'ostéoarthrite, le diabète, la maladie pulmonaire obstructive chronique (MPOC), la dépression et trois troubles neurologiques (l'Alzheimer et les démences connexes, l'épilepsie et la maladie de Parkinson).

Même si l'arrivée des dossiers électroniques semble être une excellente source de données de surveillance, certains facteurs pourraient limiter si ces données peuvent être utilisées et comment elles peuvent l'être pour améliorer la compréhension des maladies infectieuses et résoudre les éclosions. Les craintes pour le respect de la vie privée sont l'obstacle le plus important à l'accès (1). La confidentialité peut être un problème en raison des ensembles de données à forte granularité, même lorsque les données sont anonymisées et regroupées (3). Les demandes de prestations médicales et les opérations dans les pharmacies indiquent où se trouvent les établissements en soins de santé ou l'établissement de vente au détail et peuvent ne pas indiquer où se trouve le patient (12), ce qui peut être très différent en particulier lorsqu'un patient cherche à obtenir les soins d'un spécialiste ou qu'il vit dans une région rurale. Les données sur les congés de l'hôpital et les certificats de décès ne sont généralement pas disponibles en temps assez opportun pour aider les services sanitaires à détecter les éclosions de maladies (75). De plus, un seul patient pourrait avoir plusieurs fournisseurs de soins de santé, être vu dans différents contextes et dans certains cas, dans différentes compétences et avoir des assurances multiples différentes pour les médicaments d'ordonnance et les autres services, ce qui oblige à faire des liens entre un vaste nombre de systèmes (74). Finalement, en admettant que les questions de respect de la vie privée et autres questions techniques puissent être résolues, le coût d'achat des ensembles de données des compagnies privées d'assurance aux fins de surveillance peut être prohibitif (3).

La plupart des systèmes d'information électronique en santé ont été créés pour répondre aux besoins des fournisseurs de soins de santé aigus, pas la santé publique. Des questions de politique et de respect de la vie privée demeurent irrésolues et ne donnent pas de réponse sur leur utilisation possible en santé publique. Des

facteurs de main-d'œuvre et de budget dans de nombreux services de santé publique font en sorte qu'il est difficile d'accéder aux données des dossiers de santé électroniques et à d'autres sources de mégadonnées et d'en tirer profit. Les services de santé publique auront besoin de ressources pour investir dans la technologie et la formation et profiter ainsi des possibilités qu'offrent les dossiers de santé électroniques et d'autres nouvelles sources de données (76).

En dernier lieu, les dossiers de santé électroniques peuvent être d'une utilité limitée pour certaines activités de surveillance en santé publique étant donné que des variables intéressantes importantes ne sont généralement pas consignées, entre autres les facteurs de risque environnementaux ou comportementaux (76).

ix. Surveillance participative ou externalisée

Les systèmes de surveillance participative pour les maladies infectieuses peuvent être très sensibles, sont opportuns et indépendants des comportements propices à la santé (77). Les systèmes participatifs ou volontaires peuvent toutefois comporter un biais de sélection en raison de qui choisit de participer, de la difficulté d'adaptation pour les éléments perturbateurs, de la spécificité limitée des définitions des syndromes et des problèmes liés à une participation qui manque d'uniformité (77). La participation pourrait être limitée en particulier dans les régions rurales et éloignées en raison d'un accès inexistant ou peu fiable à Internet. Voici des exemples de systèmes participatifs : Influenzanet, FluTracking, Reporta et Flu Near You. Influenzanet a été évalué en Europe (78) et a détecté avec succès des modifications de l'activité du syndrome grippal avant la surveillance sentinelle des médecins. Le système peut être échelonnable et l'ajout de nouveaux participants n'augmente pas considérablement le coût de la surveillance. Influenzanet est également souple et peut être adapté sur-le-champ à des définitions différentes du syndrome grippal. D'autres facteurs de risque peuvent être ajoutés pour faciliter l'analyse individuelle détaillée ou étendre la plateforme de façon à se renseigner sur d'autres maladies.

Jusqu'à maintenant, cette méthode a été très largement utilisée pour surveiller l'influenza, mais elle peut servir pour examiner d'autres maladies (78). En France, les enquêtes sur la consommation alimentaire par le truchement d'Influenzanet ont servi à identifier la source d'une éclosion de salmonelle au début de 2016. D'autres sites Web sur mesure ont été utilisés pour obtenir des données externalisées de maladies entériques.

Par exemple, le service sanitaire de l'Utah a mis au point un site Web « I Got Sick » : https://health.utah.gov/phaccess/public/illness_report/.

Les limites des systèmes de surveillance participative comprennent l'autosélection de l'échantillon, le potentiel de fausses données intentionnelles et la déclaration volontaire de signes et de symptômes qui n'ont pas été validés par un médecin et des tests en laboratoire (78).

x. Dossiers de santé externalisés

Les personnes peuvent également partager une combinaison de données de santé déclarées volontairement, de données médicales et de données de laboratoire dans le cadre d'initiatives de surveillance en ligne, du soutien des patients et de recherche (79). Il existe plusieurs exemples de sociétés qui recueillent et partagent des données médicales. Le forum sur le Web patientslikeme.com compte plus de 500 000 membres. Le site offre aux patients la possibilité de faire part de leurs expériences et d'établir des liens avec des patients dans des projets de recherche clinique (https://www.patientslikeme.com/research/dataforgood).

23andMe.com fournit aux clients des États-Unis la possibilité de participer à des études de recherche. Les personnes qui vivent au Canada n'étaient pas autorisées à participer à ces études au moment de la rédaction du présent rapport. Il n'existe actuellement aucune protection au Canada pour éviter que des compagnies d'assurance ou des employeurs ne fassent preuve de discrimination à l'égard de personnes en raison de leurs renseignements génétiques. En 2014, 23andMe a indiqué que plus d'un million de clients avaient obtenu leur génotype et que plus de 80 % d'entre eux avaient consenti à l'utilisation anonyme de leurs données pour la recherche (https://mediacenter.23andme.com/en-ca/fact-sheet/).

La technologie portable contribue également à l'externalisation de mégadonnées pour la recherche. Fitbit, par exemple, a créé la fitabase, une bibliothèque en ligne de recherches à jour qui utilise l'information recueillie par l'appareil (http://www.fitabase.com/research-library/). La technologie portable peut fournir de l'information sur les modèles d'activité et de sommeil, les émotions, la glycémie, le rythme cardiaque et la pression artérielle (80). Le très grand nombre d'appareils offerts sur le marché et de types de données produites limitent actuellement l'utilité de ces renseignements pour la prise de décisions.

xi. Lignes téléphoniques d'infosanté

Les données des lignes téléphoniques d'infosanté sont également un élément important de tout système de surveillance syndromique. Les appels à Health Link (811) sont résumés par voie électronique et font partie de l'Alberta Real Time Surveillance System (ARTSS). Les données sont extraites et téléchargées toutes les 15 minutes (81). Lorsque les patients téléphonent à Health Link, ils parlent à une infirmière ou à un infirmier autorisé qui utilise une série d'algorithmes selon les symptômes énumérés pour suggérer des moyens à prendre pour gérer ces derniers et orienter les patients, au besoin, vers d'autres fournisseurs de soins de santé. En 2010, le système générait entre 100 000 et 125 000 enregistrements par année. Même si la quantité de données est assez petite si on la compare à d'autres types de mégadonnées, il faut des plateformes analogues pour stocker et gérer l'information et en rendre compte en temps réel. Même si plusieurs autres provinces ont également des lignes téléphoniques d'infosanté, toutes n'ont pas un système pour saisir rapidement les données, y compris la raison de l'appel.

Les données sur les appels peuvent être un élément d'information important pour construire des modèles prédictifs. Des chercheurs ont examiné les appels au service téléphonique National Health Services, NHS Direct line, au Royaume-Uni, concernant des vomissements ont été examinés pour en évaluer le lien avec l'activité de norovirus mesurée par les résultats des laboratoires. Ils ont conclu que lorsque 4 % ou plus des appels à la NHS Direct Line faisaient état de vomissements pendant plus de deux semaines consécutives dans tous les groupes d'âge, le système d'appel téléphonique servait d'avertissement d'un accroissement du nombre de cas dans les laboratoires (82).

xii. Dossiers d'absentéisme

On a également résumé, par voie électronique, les dossiers scolaires d'absentéisme de 250 écoles primaires de la ville d'Edmonton et du comté environnant dans le système ARTSS de l'Alberta (81). On a extrait et téléchargé les données tous les jours et mis à jour les données dénominatrices deux fois par année. Les raisons des absences ont été codées automatiquement au moyen d'une liste normalisée de rapports. Les écoles ont fourni les données et pour chaque élève, elles ont précisé l'année, l'âge et le code postal de la résidence, ainsi que l'emplacement de l'école. En 2010, il y avait de 500 000 à 650 000 dossiers par année. Ces chiffres devraient au moins doubler si le programme était étendu au reste de la province.

xiii. Données d'observation de la Terre

Les covariables environnementales ou les données spatiales qui pourraient présenter un intérêt pour la prévision des maladies infectieuses comprennent les précipitations, la température, le type de sol, la végétation, la densité démographique et les données du recensement selon des variables démographiques (14). Les données environnementales ont été le plus souvent utilisées jusqu'à maintenant pour prédire les maladies à transmission vectorielle telles que la fièvre de la vallée du Rift, le virus du Nil occidental, la dengue, l'encéphalite de Murray Valley et le virus Zika (14).

Les variables d'intérêt dans la prédiction des maladies ayant des facteurs de risques environnementaux comprennent : la couverture terrestre et l'utilisation des sols, le couvert végétal, les plans d'eau permanents et provisoires, les inondations, l'humidité du sol et les milieux humides, les précipitations, la température, l'altitude et le type de sol (83). Ces données sont ce qu'on appelle des produits de l'observation de la Terre et comprennent les observations de la télédétection de l'imagerie des satellites, des observations directes sur le terrain (p. ex. les données des stations météorologiques) et les extrants des chaînes de processus tels que les modèles de prédictions météorologiques (83).

Les images temporelles mondiales traitées de Fourier provenant de MODIS sont un exemple des données satellitaires traitées. Ces données ont été utilisées pour produire la réflectance infrarouge moyenne, la température à la surface du sol le jour et la nuit, l'indice de différence normalisée de végétation (NDVI) et l'indice de végétation augmentée (EVI) (84). Un certain nombre d'autres satellites procurent également des données à des résolutions diverses, gratuitement ou à un certain coût (83).

Les données de télédétection peuvent être classifiées selon la résolution spatiale ou la taille des pixels, ou encore l'intervalle de survol (83). Généralement, les données de télédétection donnent un instantané dans le temps, alors que les données obtenues sur le terrain sont peut-être plus susceptibles de renseigner sur les changements au fil du temps, par exemple les précipitations quotidiennes (83). La résolution temporelle des données utilisées dans la modélisation des maladies infectieuses devrait être examinée et justifiée pour éviter les conclusions trompeuses (83). L'incertitude des données spatiales peut et devrait, lorsque cela est possible, être directement prévue dans les modèles épidémiologiques au moyen des méthodes bayésiennes (83). Les perspectives d'une disponibilité accrue des données satellitaires chronologiques sont bonnes. Les satellites dont le lancement se faire après 2015 devraient avoir des résolutions pouvant atteindre de 10 à 60 mètres avec des délais de réobservation de cinq jours. D'autres satellites dont le déploiement est prévu devaient avoir des délais de réobservation de un à deux jours (83).

L'information des images satellitaires exige un traitement préalable considérable avant qu'elle ne puisse être utilisée. Il a fallu plus d'un an pour créer les versions initiales des cartes fondées sur les mégadonnées pour examiner le risque de virus de la dengue. L'information recueillie et les algorithmes ont par la suite été utilisés pour orienter les cartes sur le virus Zika. Comme ces dernières ont été créées à partir du travail précédent, il n'a fallu qu'un peu plus d'un mois pour les produire (14,85). Pour modéliser au mieux les répercussions des changements dans les données d'observation de la Terre sur les maladies infectieuses, l'épidémiologiste devra se servir de méthodes d'analyse chronologique et les faits nouveaux en analyse des images (83). Les véhicules aériens sans pilote ou les drones peuvent également constituer des sources nouvelles et de résolution très fine des données environnementales (83).

LES MÉGADONNÉES POUR LA MODÉLISATION DES MALADIES INFECTIEUSES

i. Les modèles comme outils de compréhension des mégadonnées

L'évaluation et l'application des mégadonnées en surveillance et gestion des maladies peuvent comprendre une description de ce qui s'est produit, des diagnostics des raisons pour lesquelles nous voyons ce que nous voyons, des prédictions de ce qui surviendra et des comparaisons d'autres scénarios de gestion (86). Les modèles prédictifs sont d'importants outils pour comprendre la transmission des maladies, expliquer ce qui se produit et ce qui pourrait se produire à l'avenir, puis pour communiquer cette information. Un certain nombre d'outils ont été mis au point au cours des dernières années pour modéliser les maladies infectieuses, la plus courante d'entre elles étant l'influenza.

EpiDMS, décrit par Liu et coll., est un exemple d'outil de modélisation prédictif (87). EpiDMS comprend des volets pour la gestion des données, leur analyse, leur visualisation et la simulation épidémique. Divers autres systèmes de modélisation ont aussi été mis au point pour mettre à profit les nouvelles sources de mégadonnées. Des chercheurs du Los Alamos National Laboratory ont observé que des modèles fondés sur les données de Wikipédia pouvaient prévoir le nombre de malades parfois jusqu'à quatre semaines d'avance, le meilleur rendement obtenu ayant été pour les modèles de la dengue et de l'influenza. Ces modèles ont été fondés sur les demandes de renseignements dans les pages Wikipédia dans sept langues. Des modèles comparables pour le choléra n'ont pas donné d'aussi bons résultats que ceux de l'influenza ou de la dengue, peut-être en raison des limites à l'accès à Internet dans la plupart des pays les plus touchés (http://www.livescience.com/49019-web-data-helps-forecast-infectious-diseases.html).

La modélisation permet de mobiliser les mégadonnées comme faits probants pour orienter les interventions et l'élaboration des politiques de santé publique. Lorsqu'un type de données ne fournit pas à lui assez de renseignements, on peut combiner plusieurs types de données. Par exemple, des données démographiques, socioéconomiques et environnementales de grande envergure ont été intégrées à l'aide des méthodes spatiales pour prédire les éclosions de *Cryptosporidia* (88). DEFENDER fait partie des autres exemples de modèles utilisant des mégadonnées : il intègre les données de Twitter et des médias d'information, de même que des algorithmes pour la détection des éclosions de maladies, la connaissance de la situation et les prévisions (51). Ce modèle comprend une fonction de prévision immédiate qui a prédit le nombre réel, mais encore inconnu, de cas mieux que tout autre modèle lorsque seules les données antérieures de dénombrement des cas ont été prises en compte (51).

Il faut des flux de données opportunes et exactes pour modéliser plus efficacement les éclosions de maladies infectieuses et mettre à jour les modèles en temps quasi réel. Le processus de modélisation doit être itératif et le modèle doit être régulièrement mis à jour pour intégrer les nouvelles données, avec observation, analyse des observations, modélisation, prévision, mise à jour suivie d'une autre répétition du cycle. Il faut perfectionner les modèles pour tenir compte des sources de données changeantes et des algorithmes d'extraction; certaines des difficultés concernant Google Flu peuvent partiellement s'expliquer par le fait que ce système n'a pas tenu compte de ces facteurs (33). Toutefois, même avec l'accès à d'excellentes données, les modèles prédictifs perdent de leur exactitude lorsqu'on tente de prévoir plus loin dans le temps. Par exemple, la limite de l'exactitude des modèles météorologiques est d'environ deux semaines (89). Les données météorologiques horaires de forte résolution sont recueillies de milliers de sites et un grand nombre des données qui en résultent sont rendues publiques. Diverses sources de données comparativement aussi grosses bien que moins spécifiques, y compris les médias sociaux et d'autres flux de données, deviennent de plus en plus à la disposition des chercheurs en santé. La difficulté réside dans l'apprentissage du meilleur usage à faire de ces données pour orienter l'élaboration des modèles. Des outils tels qu'ORBiT (Oak Ridge Bio-surveillance Toolkit) peuvent, par exemple, aider à créer les paramètres des modèles épidémiologiques découlant de mégadonnées (90).

Les difficultés liées aux données traditionnelles de surveillance des maladies limitent la capacité des concepteurs de modèles d'appuyer efficacement les décisions en raison du délai entre l'observation et le rapport, de même que l'agrégation géographique et parfois temporelle nécessaire pour protéger les renseignements personnels (89). En plus de l'augmentation possible de l'opportunité, de la granularité et de la variété des mégadonnées dont on pourrait se servir pour établir les paramètres de modèles prédictifs, on

peut aussi obtenir des données sur la perception des risques de sources de mégadonnées non traditionnelles. La volonté d'inclure les données comportementales répond à la reconnaissance grandissante que la perception des risques influence l'adoption des mesures de prévention, la probabilité de voies particulières de transmission et la réussite subséquente des stratégies d'intervention (89).

Les changements dans le comportement humain influencent le rythme de transmission des maladies dans les populations. Les changements de comportement par suite des éclosions de maladie, dits « prévalence-élastiques » peuvent comprendre la distanciation sociale, le port d'un masque et des modifications des comportements de voyage (91). Les données provenant des médias sociaux ont servi à orienter des modèles qui intègrent l'influence des médias de masse sur les attitudes et la transmission subséquente des maladies (91). La politique concernant le meilleur usage des médias pour informer le public et limiter la propagation de la maladie peut être améliorée par une meilleure compréhension des liens entre la couverture médiatique des éclosions et les changements de comportement.

Le potentiel de prédiction de la progression des maladies n'est pas le seul objectif de la modélisation. Les avantages immédiats de la construction de modèles dynamiques sont une meilleure compréhension des facteurs liés à la transmission de la maladie et la précision des lacunes d'information. Toutefois, un objectif ultime, bien qu'encore lointain, pourrait être d'avoir un système pour les modèles des maladies infectieuses qui se compare à celui des prévisions météorologiques dans lequel de gros volumes de données sont intégrés aux modèles en temps réel et les prédictions pertinentes locales immédiatement disponibles dans les téléphones intelligents (89). La combinaison des mégadonnées spatiales et des modèles dynamiques pourrait faciliter la gestion adaptative en temps réel des maladies infectieuses (12). Il faut toutefois tenir soigneusement compte des limites des données et des données stochastiques, puis tenir compte de l'incertitude dans les prédictions qui en résultent (12). Les méthodes bayésiennes ont été utilisées conjointement avec diverses sources de données pour créer des modèles prédictifs qui s'adaptent à la sous-déclaration et aux biais d'observation dans les flux de données (92) et fournir des estimations de l'incertitude qui en découle dans les résultats.

ii. Gestion des intrants et des extrants de données issus des modèles de simulation – un autre type de mégadonnées

Il existe également un besoin très réel de gérer les grands volumes de données nécessaires aux modèles de simulation et produits par eux pour différents scénarios avec différents paramètres, ciblant des échelles spatiales différentes et des interventions différentes (87). Les intrants et les extrants des modèles sont souvent eux-mêmes des mégadonnées. Les concepteurs des modèles doivent pouvoir produire, interroger, visualiser et analyser les intrants et les extrants des modèles en temps opportun et de manière efficiente. EpiDMS est un exemple de logiciel dont nous avons parlé et qui est conçu pour répondre à certains de ces besoins. Les données d'intrant sont, par exemple, les caractéristiques démographiques, les réseaux de contacts, les taux de contacts précis par âge ou sexe, les modèles de mobilité, ainsi que les détails des interventions à différents niveaux de résolution spatiale (87). Le système de gestion des données doit également tenir compte du potentiel de variations dynamiques dans le temps des paramètres et des données sous-jacentes. De même, il faut, dans l'analyse de sensibilité, prévoir la variation dans ces paramètres, pour connaître l'influence d'un éventail de valeurs paramétriques plausibles sur les résultats du modèle (87).

En plus des paramètres d'intrant, le système de gestion des données doit pouvoir tenir compte des très grands volumes de données complexes produits par un ensemble de réalisations stochastiques qui pourrait comprendre des centaines, voire des milliers de simulations (87). Des conditions changeantes exigeront également d'établir une série de nouvelles simulations au fil du temps. Pour que les mégadonnées puissent être utilisées efficacement par une modélisation de simulation dynamique pour orienter la gestion des maladies, il faut des outils qui aideront à exécuter les ensembles de simulation de grande envergure comportant divers scénarios et faciliter l'analyse, l'exploration, l'interprétation et la visualisation des résultats des modèles (87). EpiDMS peut interroger des données stockées à la suite de simulations pour trouver des correspondances de modèles de maladies observés ou des cibles d'intervention.

VISUALISATION DES MÉGADONNÉES POUR LES MALADIES INFECTIEUSES

Pour que les mégadonnées soient utiles à la prise de décisions en santé publique, il faut des outils pour résumer et appuyer la visualisation et reconnaître les modèles, les tendances, les corrélations et les valeurs aberrantes (93). Un certain nombre de plateformes existantes assurent un soutien de base aux utilisateurs (90). Le RMISP, par exemple, fournit des alertes sous forme textuelle aux utilisateurs. ProMED-mail fournit des cartes SIG et des alertes dont les priorités sont établies en fonction de l'occurrence et l'ampleur de la maladie, de même qu'un résumé sous forme textuelle des données. HealthMap approfondit les caractéristiques de rapport de la carte SIG par des calendriers et des tableaux/graphiques.

Dans de nombreux ministères de la Santé, l'influenza est un objet de surveillance et de rapports publics. Le CDC aux États-Unis a investi pour rendre les données de surveillance de la grippe disponibles au public sous une forme interactive et conviviale (www.cdc.gov/flu/weekly/fluviewinteractive.htm). Les données ne sont toutefois offertes qu'à l'échelle des États et il y a un délai de deux semaines dans le rapport des données de laboratoire et de mortalité (3). L'Agence de la santé publique du Canada tient également le rapport hebdomadaire mis à jour du programme Surveillance de l'influenza (https://www.canada.ca/fr/sante-publique/services/publications/maladies-et-affections/surveillance-influenza/2016-2017/semaine11-12-18-mars-2017.html) qui contient une série de cartes, de tableaux et de graphiques qui résument les indicateurs actuels clés de la surveillance. Des cartes animées sont fournies pour montrer les changements dans l'activité de l'influenza au fil du temps.

Par contraste, la plupart des renseignements disponibles sur les maladies d'origine alimentaire sont contenus dans des rapports périodiques ou présentés sous forme de textes et de tableaux (https://www.canada.ca/fr/sante-publique/services/maladie-origine-alimentaire-canada/surveillance-maladies-origine-alimentaire-canada.html). L'ASPC a également un infographique qui résume l'information récente sur les maladies d'origine alimentaire au Canada (www.healthycanadians.gc.ca/publications/eating-nutrition/foodborne-illness-infographic-maladies-origine-alimentaire-infographie/alt/pub-fra.pdf). Les diagrammes de séries chronologiques et les graphiques circulaires personnalisés existent pour toutes les maladies à déclaration obligatoire, y compris les maladies d'origine alimentaire. (http://maladies.canada.ca/declaration-obligatoire/liste-graphiques).

Les options de visualisation les plus souvent indiquées comprennent les cartes à points ou choroplèthes, les graphiques des mesures d'intérêt au fil du temps, les graphiques circulaires par catégorie démographique et les graphiques à barres (90). Ces outils sont utilisés pour communiquer la chronologie de la maladie, sa répartition géographique ou faire des comparaisons entre les groupes d'âge ou d'autres facteurs de risque.

i. Amélioration de l'efficacité de la visualisation et reconnaissance des obstacles

En 2014, Carroll et coll. ont publié un examen systématique des pratiques et des problèmes actuels en visualisation et en analyse de l'épidémiologie des maladies infectieuses. Plusieurs observations avaient trait aux facteurs qui contribuent à l'efficacité des évaluations. Il est incontestable que de bonnes visualisations améliorent la compréhension des données, rehaussent la détection des modèles et les inférences qui découlent des données. Les visualisations doivent être compréhensibles pour les utilisateurs de diverses disciplines et être créées de manière à réduire au minimum les risques de surcharge d'information ou de mésinterprétation par le lecteur.

Voici des facteurs communs qui peuvent être utilisés pour évaluer différents types d'outils de visualisation et d'analyse : utilité, simplicité d'apprentissage, capacité de mémorisation, prévention et correction des erreurs, efficience et satisfaction des utilisateurs (94). Les outils doivent prévoir des options pour cibler les données manquantes et l'incertitude. Pour optimiser l'adoption des outils et leur utilité, il est important de comprendre les besoins et les préférences des utilisateurs, leur formation, l'intégration de l'outil dans les milieux de travail, le degré de compréhension et l'utilisation précédente des visualisations, la confiance des utilisateurs et le soutien de l'organisation (94).

Carroll et coll. ont également abordé les obstacles à la création et à l'interprétation des visualisations par les professionnels de la santé publique (94). Les obstacles communément cités concernant l'utilisation des outils comprenaient des degrés variables de connaissances informatiques, un soutien TI insuffisant de la part de l'organisation, un accès limité au logiciel et une incompréhension de la raison d'être de l'outil ou du degré de difficulté de l'utilisation de ce dernier. Les auteurs ont laissé entendre que l'utilisation courante des outils d'analyse et de visualisation pourrait être améliorée par l'intégration des outils dans les flux de travail courants. L'agrégation et l'anonymisation des données ont également été citées comme des obstacles à la capacité de produire des recommandations utiles à partir des données.

Des systèmes en ligne ont été recommandés pour diminuer les coûts de la mise en œuvre des logiciels, améliorer l'accessibilité de tous les membres des équipes et la capacité de diffuser les résultats dans différents milieux (94). Le coût de nombreux progiciels a également été signalé comme un problème grave, car de nombreux organismes n'ont pas d'autres options que d'utiliser les ressources logicielles gratuites. Les courbes d'apprentissage de nombreux progiciels gratuits sont toutefois plus abruptes, les options peuvent être plus restreintes et les ressources de soutien limitées. Il faut de la documentation en ligne de haute qualité et un code source d'accès facile. OutbreakTools est un exemple de plateforme R de source libre pour la gestion

et l'analyse des données sur les éclosions, même si elle n'est pas spécifiquement réservée aux mégadonnées (22).

Les questions de confidentialité et de respect de la vie privée limitent également les stratégies de visualisation. Une personne a proposé que la solution au travail à une résolution spatiale supérieure dans certains cas consistait à résumer un ensemble de données semblable aux caractéristiques sous-jacentes (12). D'autres solutions régulièrement proposées comprenaient l'agrégation des données à une résolution spatiale moins grande, le résumé des données aux paramètres couramment utilisés dans les modèles épidémiologiques et l'affichage des résultats du modèle au lieu des données brutes. D'autres options ont également été proposées pour une cartographie et des analyses épidémiologiques spatiales utiles des données locales sur les maladies, sans pour autant compromettre la confidentialité (95,96).

ii. Domaines d'intérêt pour la visualisation des mégadonnées

Carroll et coll. ont proposé trois domaines d'intérêt pour une meilleure adoption et application des outils de visualisation : les systèmes d'information géographique (SIG), l'épidémiologie moléculaire et l'analyse des réseaux sociaux (ARS) (94). La visualisation SIG comprend la simplification, l'intégration et l'analyse des données spatiales. Il existe actuellement des systèmes qui mettent en permanence à jour des cartes à points de l'occurrence de maladies infectieuses (HealthMap) ou des cartes choroplèthes mises à jour chaque semaine (Cartes de surveillance de l'influenza et Surveillance du virus du Nil occidental). Il existe également des cartes spatiales statiques des risques de maladie, mais il n'y a pas encore d'option reconnue en ligne de cartes continuellement mises à jour pour montrer un risque spatial en continu (84). Les méthodes d'analyse spatiale en continu comprennent l'application de techniques telles que la régression de type kernel par lissage et les estimations de risque fondées sur des modèles. Par exemple, les taux d'infection au virus occidental du Nil selon la modélisation chez les moustiques femelles *Cx. tarsalis* ont été cartographiés en juillet et en août, de 2005 à 2008, dans les prairies canadiennes (97).

De plus, les résultats des analyses de grappes spatiales ou spatiotemporelles peuvent être cartographiés et les secteurs à haut risque surlignés au moyen des cartes de risque relatif, comme dans l'exemple d'une étude de l'infection *Salmonella Enteritidis* dans la région du Grand Toronto (98). Greene et coll., ont également fait mention de l'utilisation de SatScan par le service sanitaire de New York pour voir de manière prospective les grappes spatiotemporelles dans les données sur les maladies à déclaration obligatoire (99). Ce processus a

fait ressortir des éclosions de shigellose. L'application de la détection des grappes et d'autres méthodes spatiales dans les enquêtes sur les éclosions ont récemment été résumées dans une évaluation systématique (100). D'autres méthodes mentionnées pour l'exploration des données spatiales de surveillance comprenaient les appareils d'analyse géographique, DYCAST (analyse spatiotemporelle continue dynamique), la modélisation basée sur les agents, les statistiques spatiales et les méthodes auto-organisées (101). Les programmes commerciaux de modélisation de simulation tels que AnyLogic ont de plus en plus perfectionné le soutien pour intégrer des caractéristiques de SIG dans des modèles de simulation hybrides et basés sur des agents (http://www.anylogic.com/).

Les données moléculaires peuvent constituer l'un des domaines où il sera le plus difficile de produire des résumés, des tableaux et des cartes clairs et compréhensibles, en particulier pour les personnes qui ne connaissent pas l'information résultante. Les analyses phylogénétiques sont souvent représentées comme des arbres ou des dendrogrammes (102); toutefois, il existe un certain nombre d'options graphiques pour dépeindre des similarités et des différences entre les organismes (103). Les résultats des analyses moléculaires comportant de l'information sur l'emplacement peuvent également être cartographiés, comme on l'a fait pour le risque de génotype différent pour les infections par *Campylobacter* chez les humains en Nouvelle-Zélande (104).

Les résultats de l'analyse des réseaux sociaux (ARS) sont le plus souvent représentés par une série de graphiques faisant ressortir les nœuds et les sous-groupes importants dans le système, en plus des liens indispensables le long des divers nœuds et sous-groupes (94). Il existe plusieurs exemples intéressants de visualisation dans lesquels les résultats de l'ARS ont été combinés à des données moléculaires et au séquençage pangénomique. Dans un exemple, l'ARS a été utilisée conjointement avec le suivi des contacts et l'ARS pour découvrir la nature de deux souches moléculaires distinctes et une éclosion de tuberculose dans un réseau social à haut risque en Colombie-Britannique (102).

iii. Difficultés particulières de la visualisation des mégadonnées

Ola et Sedig ont examiné plusieurs problèmes particuliers de la conception de visualisations pour les mégadonnées complexes (93). Ces auteurs donnent à penser qu'il faut un cadre conceptuel pour rendre le processus de conception gérable et obtenir des visualisations efficaces (93). La conception des visualisations doit tenir compte des tâches engendrées par les mégadonnées et de la nécessité de mises à jour dynamiques

souples et permanentes pas seulement des données, mais aussi du contenu, du thème et des questions abordées dans les visualisations.

Il faut des visualisations interactives pour que les initiatives comportant des mégadonnées permettent à leurs utilisateurs d'explorer simultanément de nombreux éléments de données pour reconnaître les modèles et vérifier des hypothèses (93). ORBiT Toolkit (90) est un exemple de ce type de plateforme interactive. En plus des outils permettant d'intégrer des données de surveillance plus traditionnelles (salles d'urgence, ventes dans les pharmacies) à celles des médias sociaux, ORBiT présente les résultats de l'analyse dans une interface visuelle dynamique où l'utilisateur final peut interagir et faire des commentaires.

De simples diagrammes de dispersion, graphiques à barres, cartes choroplèthes ou cartes thermiques peuvent ne pas suffire pour l'examen de liens complexes entre des variables multiples provenant de sources multiples. De même, des animations peuvent ne pas suffire à montrer les changements au fil du temps parce qu'elles se fient à la mémoire de l'utilisateur et conviennent mieux aux présentations qu'aux analyses (93). De même, des visualisations statiques multiples indépendantes réparties dans les pages des tableaux de bord comptent aussi sur l'utilisateur pour intégrer l'information dans les diverses images (93). Idéalement, il devrait y avoir des visualisations qui permettent de présenter des aspects multiples des données par couche dans le même espace (93).

Même si les mégadonnées pouvaient considérablement améliorer l'efficience et entraîner des économies de coûts dans les soins de santé, les répercussions véritables des mégadonnées sur la gestion des maladies dépendent directement de la disponibilité des outils pour faire des grands ensembles de données de l'information digestible qui peut être utilisé pour prendre des décisions en temps opportun (93).

MISES EN GARDE CONCERNANT L'APPLICATION DE MÉGADONNÉES À LA SURVEILLANCE DES MALADIES INFECTIEUSES ET AUX ENQUÊTES SUR LES ÉCLOSIONS

i. Analyse des données et risque de faux positifs

L'exploration des mégadonnées est généralement orientée vers la recherche qui produit des hypothèses plutôt que sur la recherche basée sur des hypothèses (105). Les conclusions sont souvent fondées sur un raisonnement par induction plutôt que déduction, la pierre angulaire de toute recherche actuellement acceptée en santé. La validation des résultats est indispensable pour réduire au minimum le risque de conclusions de faux positifs (106). Les faux positifs viennent de corrélations fallacieuses qui peuvent causer beaucoup de

tort, par exemple, si l'on se demande si un traitement ou un vaccin est véritablement sûr et efficace. En plus du risque d'erreurs statistiques de type I, il risque aussi que se propagent rapidement des faux renseignements dans les médias sociaux et c'est pourquoi il faut faire preuve de prudence. Ce risque peut être géré par l'utilisation avec les nouvelles sources de données et approches d'analyse d'une bonne analyse statistique fondée sur des hypothèses, un raisonnement causal prudent et la consultation d'experts (1).

Souvent, les études de mégadonnées se caractérisent par l'absence de certaines valeurs, un manque de congruence, des sources d'information disparates, des mesures à différentes échelles, l'hétérogénéité et leur lourdeur (54). Une analyse doit comporter une exploration attentive des données, de leur classification et le suivi des modèles. La modélisation mixte de diverses sources de données est compliquée par l'hétérogénéité, le bruit et les corrélations fallacieuses, la dépendance non reconnue entre des données et des termes erronés, et la latence d'importantes variables d'intérêt (54). Les résultats des analyses de mégadonnées doivent être intégrés aux analyses de systèmes existants et être constamment validés (1).

ii. Nécessité de nouvelles compétences en gestion des données et d'un nouveau vocabulaire d'analyse

Les chercheurs et les praticiens qui utilisent les mégadonnées trouveront utile de connaître l'apprentissage automatique, l'exploration des données et les algorithmes créés par la machine pour chercher des modèles et des liens dans les données. Même si tous les chercheurs et praticiens ne voudront pas nécessairement apprendre toutes ces techniques, il est utile de pouvoir communiquer efficacement avec des collègues en sciences informatiques. Voici des exemples de termes couramment utilisés par de nombreuses personnes qui travaillent en surveillance des maladies : Hadoop (cadre de programmation qui permet le traitement de mégadonnées dans de nombreux ordinateurs); apprentissage non supervisé (analyses qui cherchent des modèles cachés dans les données); analyses graphiques (analyses qui utilisent les graphiques pour comprendre les liens et les modèles); et traitement du langage naturel (analyses permettant aux ordinateurs de trouver un sens au langage humain et, par conséquent, d'extraire des connaissances des documents) (106). L'analyse perfectionnée que peu de chercheurs en sciences de la santé connaissent est nécessaire pour élucider les interactions complexes et les liens causaux dans l'analyse des mégadonnées (106). Toutefois,

certains outils comme les réseaux bayésiens et les modèles graphiques, de même que les analyses spatiotemporelles, sont déjà utilisés dans certains domaines de l'épidémiologie.

Une grande partie des données brutes provenant des sources de mégadonnées d'intérêt sont sous forme de textes non structurés, d'images, de données de séquençage pangénomique et de rapports de laboratoire. Le prétraitement est nécessaire pour structurer ces données avant de les analyser. Des exemples de stratégies de classification des mégadonnées complexes comprennent l'apprentissage automatique non supervisé, les réseaux de croyance bayésiens, l'apprentissage approfondi et des méthodes d'ensemble (54). D'autres options de regroupement et de classification des données comprennent la modélisation à mélange gaussien, les forêts aléatoires et la règle des K plus proches voisins. Les machines à vecteurs de support (SVM) ont été utilisées pour créer des classifications linéaires binaires de données complexes fondées sur des caractéristiques « a priori » identifiées dans les données de formation (54). Des modèles linéaires généralisés plus simples peuvent également convenir dans certains cas (54). Les analyses à composantes principales ou indépendantes sont des exemples d'outils d'exploration non supervisée dans le cas des données quantitatives (54).

Les approches supervisées utilisent un ensemble de formation qui comprend des données déjà classifiées pour en faire des inférences ou classifier les nouvelles données, et les approches non supervisées déterminent la structure des données non étiquetées (54). Il existe aussi des algorithmes semi-supervisés ou partiellement supervisés (54). L'analyse non supervisée de mégadonnées permet de chercher des fils conducteurs communs qui pourraient lier des points de données en apparence sans lien et trouver des associations qui pourraient autrement passer inaperçues. Ce type d'analyse n'explique pas, cependant, le « comment » ni le « pourquoi » de ces associations (107). L'analyse des réseaux sociaux peut également servir à extraire des renseignements utiles de modèles et de liens dans les données issues des plateformes de médias sociaux et des appareils possédant des capteurs (54).

Pour gérer des mégadonnées très considérables, il faut des mécanismes qui facilitent le traitement parallèle des données entre des appareils distincts, mais reliés entre eux. L'informatique parallèle suppose l'exécution simultanée de tâches d'algorithmes dans un regroupement d'appareils ou des superordinateurs (18). Nous avons déjà mentionné Hadoop qui constitue une mise en œuvre à source ouverte de MapReduce de Google (54). C'est aussi un exemple d'une plateforme informatique nuagique qui permet le stockage centralisé de données et l'accès à Internet à distance à des ressources de calcul pour les infrastructures et les logiciels (18). Spark est un autre exemple qui, selon ce qu'on en dit, est plus rapide que certaines applications et facilite les interfaces avec d'autres programmes communément utilisés tels que Java, Scala, Python et R nécessaires aux

recherches, à la diffusion en continu, à l'apprentissage automatique et au traitement graphique (54). L'utilisation des mégadonnées exige des ressources et des compétences spécialisées en stockage et extraction des données, en identification des erreurs, en sécurité, en protocoles de partage et en analyse des données (18).

iii. Sécurité des données, gouvernance des données et respect de la vie privée

De nombreuses organisations ont souligné la nécessité d'assurer la sécurité des systèmes (14,108). Les questions préoccupantes comprennent, sans toutefois s'y limiter, la vulnérabilité des ensembles de données aux intrusions informatiques et l'utilisation potentielle des données du séquençage pangénomique pour concevoir des armes biologiques; l'inondation des ensembles de données par de faux renseignements; et le piratage des bases de données ou des systèmes informatiques. La sécurité devient de plus en plus complexe à mesure que les initiatives de mégadonnées englobent souvent plusieurs bases de données et du matériel à différents lieux et sous différentes gestions (109). L'intégrité des données pourrait aussi être menacée à leur origine par le potentiel quelque peu improbable que les processus de production des données externalisées soient manipulés pour répondre aux objectifs de certains groupes (38).

Les initiatives liées à des mégadonnées ont fait naître un nouveau besoin de compétences en gouvernance des données (108). Ces compétences comprennent la planification des coûts d'un projet, la gestion de la confidentialité et du respect de la vie privée, les difficultés d'ordre éthique liées au consentement éclairé. Les concepteurs de projet doivent reconnaître les coûts parfois substantiels liés au nettoyage, à la normalisation, au stockage, à la transmission et à la sécurité des ensembles de données très considérables et en croissance rapide (108). Même si les coûts de gestion sont un aspect important à considérer, la propriété des données et les questions de respect de la vie privée créent souvent des difficultés encore plus grandes en gouvernance des données.

Cheung et coll. ont résumé les préoccupations liées au respect de la vie privée parmi les premiers à adopter les nouvelles technologies de la santé et ces préoccupations pourraient être classées sous quatre grands thèmes : protection des renseignements personnels, contrôle des renseignements personnels, craintes liées à la discrimination et aspects de la contribution des renseignements personnels à la science (110). On ne sait pas vraiment si la plupart des gens connaissent l'ampleur des détails dans leurs renseignements personnels électroniques ou comment ces renseignements pourraient être combinés à d'autres sources de données

publiques (110). La plupart des gens ne savent pas non plus que l'HIPAA (loi américaine sur l'assurance maladie) ne protège pas les renseignements personnels qui peuvent être déduits par le lien avec des bases de données publiques (110).

Plusieurs problèmes de nature éthique se posent face à l'utilisation des mégadonnées et des nouvelles technologies (4). On se demande comment on pourrait définir un consentement éclairé suffisant. Certains proposent une « approche des moments critiques » selon laquelle le consentement est obtenu au début de l'étude, puis à chaque moment où de nouvelles données sont demandées ou qu'une intervention est amorcée. En plus du consentement, certains se demandent si l'on a tenu compte des besoins de la population à l'étude et si l'on a cherché un juste équilibre entre ces besoins et l'augmentation de l'accès aux données et les possibilités du développement rapide de la recherche et des outils d'innovation (4). Il y a également d'importantes questions à se poser au sujet de la nécessité de réduire au minimum la possibilité d'intrusion et de fatigue des participants, la nécessité d'exprimer des attentes claires et pertinentes quant aux préoccupations liées à la sécurité et la nécessité de protéger les nouveaux flux de données de nature délicate et potentiellement identifiables (4).

iv. Limites des données et biais possibles

L'une des premières limites importantes est l'absence de données démographiques de base dans de nombreuses sources de données (1). L'absence de données démographiques peut compliquer l'évaluation du potentiel de biais de sélection dans les résultats. La couverture, pour de nombreux types de tonnées, est limitée aux très jeunes enfants et aux populations de personnes âgées (1). De plus, les femmes sont plus susceptibles que les hommes de participer aux efforts d'externalisation comme Influenzanet (78).

L'instabilité dans les sources de données est également une menace réelle. Rien ne garantit que les plateformes des médias sociaux, les sites de discussion dans Internet, les sites Web aux données publiquement disponibles seront autant utilisés à l'avenir. La mention de MySpace et des forums de discussion Yahoo dans les études plus anciennes en est un exemple (48); ces deux plateformes ne sont plus aussi utilisées depuis les dernières années. Les taux de participation et de décrochage des plateformes de fournisseurs volontaires changent aussi en fonction de l'intérêt du public et des efforts faits pour maintenir l'engagement des participants (78). Flu Near You a, par exemple, observé des pointes isolées de participation et des taux élevés de décrochage aux États-Unis (78).

Les données sur les réclamations au titre des frais médicaux peuvent ne pas refléter avec exactitude les risques de maladies infectieuses, même dans les populations qui consultent pour obtenir des soins en raison des difficultés liées aux pratiques de facturation (12). Il peut aussi y avoir des erreurs spatiales et temporelles liées aux données des pharmacies et des hôpitaux, car le point de vente ou le lieu de réception des soins ne reflète pas nécessairement le lieu de résidence de la personne no le lieu où elle est devenue malade (1). De même, lorsqu'une personne a publié un gazouillis relatant qu'elle a été malade, son message peut ne pas indiquer où elle a été exposée à la maladie (12).

Différentes sources de données résumées à différents niveaux d'agrégation rendent les fusions et les analyses difficiles (12). De même, les comportements liés à la déclaration de la maladie et à la recherche de soins de santé varieront au fil du temps, selon les pays et les groupes d'âge, ce qui fait douter de la pertinence de fusionner des données de différentes sources (3). Les données spatiales sont probablement plus complètes et de meilleure qualité quand les participants vivent dans les milieux urbains plutôt que ruraux (12) parce que dans les deux cas, les CDR et les adresses postales auront des erreurs associées. Seules les coordonnées du GPS offriront des données de bonne qualité dans les régions rurales.

Les modèles fondés principalement sur les données individuelles sont sujets au surapprentissage et aux erreurs atomistiques (12); toutefois, le sophisme écologique est également un problème important dans les études fondées sur des données agrégées lorsque les résultats et les facteurs de risque ne peuvent pas être liés à l'échelle individuelle. Même dans les études où les chercheurs disposent de données pour une grande partie de la population, il leur faut insister sur la rigueur statistique par une validation externe, l'accès de sources ouvertes aux données et aux codes, et les évaluations de l'influence des valeurs aberrantes (3) et ils doivent songer à des seuils plus prudents de mesure de la signification statistique (111). La validation doit également comprendre la réplication des résultats de l'étude (111) en insistant sur l'importance d'expliquer leurs méthodes pour en arriver aux résultats pour s'assurer que la réplication est possible. L'une des critiques de Google Flu était l'incapacité à indiquer avec transparence les termes de recherche utilisés pour extraire les données (38).

Finalement, il faut se demander si le système est exact et utile en pratique pour éclairer les décisions de gestion. Il est important de se demander si le système est capable de déceler les changements dans la dynamique temporelle ou spatiale des infections ou les changements dans les groupes d'âge touchés (3). Les résultats associés à un manque de normalisation, une information géographique limitée et l'incapacité de vérifier ou de clarifier l'information originale provenant des sources de mégadonnées paraîtront incertains (89).

La surveillance de laboratoire traditionnelle contient beaucoup de renseignements, possède un rapport signal-bruit élevé et une haute spécificité, mais son volume de données est faible. Elle est généralement peu complexe et nécessite peu de traitement préalable. Par contraste, alors que les données de Twitter ont un volume élevé de données, ce volume est considérablement plus complexe à interpréter et nécessite un traitement préalable considérable. Par comparaison, la quantité de renseignements utiles est limitée, le rapport signal-bruit est faible et la spécificité limitée. La meilleure combinaison d'information, le meilleur rapport signal-bruit, la meilleure spécificité et le meilleur volume des données s'obtiennent de systèmes hybrides qui combinent la surveillance traditionnelle aux mégadonnées (3).

Résumé des méthodes d'exploration de ces sources de données pour prédire et atténuer les éclosions de maladies d'origine alimentaire au Canada

Selon les estimations de l'ASPC, on estime qu'environ quatre millions d'épisodes de maladies d'origine alimentaire contractées au pays surviennent chaque année au Canada, dont 1,6 million (40 %) sont liées à 30 agents pathogènes connus (112). Du nombre total de cas de maladies d'origine alimentaire, 11 600 environ nécessitent des hospitalisations et 238 occasionnent des décès; 4 000 (34 %) de ces hospitalisations et 105 (44 %) des décès liés aux maladies d'origine alimentaire contractées au pays sont attribuables à 30 agents pathogènes connus (113). Dans une autre étude canadienne récente, 63,5 % des enquêtes sur des éclosions de maladies d'origine alimentaire ont signalé un aliment précis comme source de l'éclosion (114). Par comparaison, entre 2003 et 2010, un véhicule transportant des aliments a été ciblé dans 692 des 1 441 éclosions de maladies d'origine alimentaire (48 %) aux États-Unis (115). Ces résultats soulignent la nécessité de recourir à des ressources et à des outils additionnels pour enquêter sur les éclosions de maladies d'origine alimentaire afin d'augmenter le pourcentage de cas où la source pourrait être identifiée et améliorer la prévention future et les efforts de lutte contre ces maladies. Comme 32,2 % des éclosions de maladies d'origine alimentaire ont été associées à un établissement de services alimentaires, il existe aussi des solutions immédiates pour améliorer les efforts de contrôle grâce à de meilleures données pour cibler les inspections de la santé publique (114).

D'autres ont reconnu que la surveillance des maladies d'origine alimentaire fondées sur les incidents peut être améliorée par l'information provenant d'Internet et des médias sociaux (75). La littérature décrite dans le présent aperçu a donné plusieurs exemples de la façon dont le séquençage pangénomique, les rapports

analytiques de nouvelles, les recherches dans Internet (Wikipédia), les données des forums d'Internet (évaluations de Yelp), les téléchargements de Twitter, les téléphones intelligents, la surveillance dans les pharmacies, la surveillance des ventes d'aliments au détail, les données de l'externalisation et de la participation et les lignes d'infosanté ont particulièrement contribué à la surveillance des maladies d'origine alimentaire et à la prédiction et l'atténuation des éclosions de maladies d'origine alimentaire.

Les prochains paragraphes du présent document relatent l'avis d'un chercheur en exercice sur les possibilités offertes par les mégadonnées, puis les avis de praticiens de la santé publique pour connaître leur compréhension des mégadonnées, comment elles sont actuellement utilisées et comment elles pourraient l'être à l'avenir.

Commentaire d'un chercheur – Nathaniel Osgood, Ph. D.

Pour compléter les résultats de présent examen de la littérature, nous avons demandé à Nathaniel Osgood, Ph. D., de l'Université de la Saskatchewan de faire un commentaire en tant que chercheur sur le potentiel de différents types de mégadonnées pour les enquêtes sur les éclosions de maladies. Comme l'application des données du séquençage pangénomique aux maladies d'origine alimentaire a fait l'objet de discussions détaillées récemment (19), il n'aborde pas vraiment ces types de données dans ses propos. M. Osgood est un expert reconnu à l'échelle internationale en modélisation en santé et l'acquisition de données de contact de haute résolution à l'aide de capteurs et leur application à la modélisation des maladies.

Comme il est dit ailleurs dans le présent document, les systèmes visant à déceler les cas de maladies d'origine alimentaire peuvent sonder de nombreuses sources de données. Pour améliorer les enquêtes sur les éclosions, en nous fondant en partie sur les travaux effectués au laboratoire d'épidémiologie génématique et d'informatique en santé publique de l'Université de la Saskatchewan, nous envisageons ici un système multiniveaux qui mobilise plusieurs types de surveillance et de détection à différents niveaux de portée et de profondeur de l'information fournie. La première étape consisterait en surveillance traditionnelle des cas de maladies gastro-intestinales hautement crédibles, mais compléterait les données ordinaires recueillies auprès de ces personnes par des données issues des transactions des cartes de crédit et de débit qui identifient les achats faits chez des distributeurs alimentaires. À ces renseignements s'ajouterait la surveillance des évaluations des restaurants locaux et des commentaires sur des sites tels que Yelp, mais aussi la surveillance des mises à jour des statuts

Facebook et Twitter qui, comme on l'a vu, offrent des volumes considérables de publications pertinentes. Ces renseignements pourraient, dans des circonstances précises, orienter les priorités d'enquête auprès des fournisseurs.

La source peut-être la plus importante de données prospectives pourrait consister en une surveillance syndromique d'un sous-ensemble démographique de personnes sentinelles (sous-ensembles de diverses tailles à partir de 4 % et plus de la population) qui portent une application leur permettant de déclarer facilement les symptômes subcliniques éprouvés, de même qu'une information (facultative) sur des modèles particuliers de mouvements géographiques. L'obtention fiable de renseignements déclarés volontairement de maladie subclinique et de consommation d'aliments serait probablement grandement améliorée par des déclarations volontaires proactives (p. ex. des boutons des applications) et l'utilisation d'évaluations écologiques momentanées par laquelle la personne fait état de sa consommation récente d'aliments et de l'expérience d'une maladie gastro-intestinale. Par une double vérification de l'intégralité des rapports proactifs, ces évaluations rendront probablement compte de nombreux cas de maladies et de consommation d'aliments passés inaperçus.

Les personnes qui déclarent des symptômes gastro-intestinaux de manière proactive ou en réponse à des questions peuvent aussi se faire demander si elles consentiraient à partager les enregistrements de leurs achats récents ou à répondre à des questions sur leur consommation d'aliments provenant de fournisseurs alimentaires particuliers. Pour résoudre les problèmes de protection des renseignements personnels pour ces cohortes sentinelles (et accroître ainsi peut-être de beaucoup la réserve de sentinelles), une politique d'actualisation du consentement pourrait être établie. Dans ce système, seuls les dénombrements agrégés de déclarations sur de vastes régions géographiques seraient fournis aux autorités dans une situation ordinaire. Les rapports géolocalisés à l'échelle individuelle resteraient privés jusqu'à la déclaration d'une éclosion possible, dans lequel cas les rapports pourraient être publiés pour faciliter l'enquête. Ces divulgations pourraient être faites automatiquement ou après l'autorisation explicite de la personne concernée, informée de l'urgence avérée de santé publique.

Les recherches menées précédemment par l'auteure et ses collaborateurs donnent à penser que de plus grands volumes de données subcliniques recueillies de cette manière plus détaillée pourraient offrir des avantages marqués qui permettraient d'identifier beaucoup plus rapidement la source de contamination dans des éclosions. En même temps, des travaux récents donnent à penser qu'un flux de dénombrements agrégés (et ainsi anonymisés) de maladies subcliniques peuvent être mis à profit par les techniques d'apprentissage automatique

pour obtenir une détection rapide d'un nouvel épisode d'éclosion, ce qui milite en faveur d'une mise en œuvre hâtive des efforts de lutte contre l'éclosion.

À l'avenir, ces évaluations écologiques momentanées de l'alimentation seront aussi vraisemblablement plus efficaces si elles sont déclenchées par des algorithmes de classification en ligne qui déduisent que le sujet va bientôt manger et qu'il se verra poser des questions sur sa consommation d'aliments à un moment qui se rapprochera des repas; ces évaluations seront donc moins sujettes au biais de rappels; de plus, ces évaluations seront probablement moins fastidieuses. Les explorations des auteurs donnent à penser que les algorithmes de classification pour identifier les comportements d'alimentation pourraient fructueusement tenir compte de l'orientation téléphonique, des données des accéléromètres et des gyroscopes, de même que l'emplacement selon les estimations du GPS et du sans-fil. Dans les cas de maladie, l'information qui pourrait susciter une question pourrait comprendre les modifications dans les modèles de mobilité entre une installation ou à l'intérieur de l'une d'elles et des rappels sonores possibles. À mesure que les montres intelligentes deviendront encore plus courantes et perfectionnées, la détection de la consommation probable d'aliments et la détection de maladies pourront aussi s'améliorer. Dans le cas de la consommation d'aliments, cette détection pourrait comprendre de l'information sur les modèles de mouvements de la montre qui pourraient suggérer un comportement alimentaire. La classification des cas de maladies d'origine alimentaire peut être facilitée par des données physiologiques recueillies par la montre, par exemple le rythme cardiaque, la variabilité du rythme cardiaque et l'information sur l'activité électrodermale.

Une fois des établissements donnés devenus suspects en raison de l'information réunie par un moyen ou un autre, d'autres données provenant de ces fournisseurs pourraient être colligées de façon à obtenir plus rapidement la confirmation ou la dénégation de l'existence d'un problème. Les données au point de vente des achats faits à cet endroit comprendront probablement le moment où l'achat a été fait, les coordonnées de l'acheteur et peut-être quelque indication quant au nombre de personnes liées à cet achat. Dans le cas de certaines enquêtes, ces renseignements comprendront probablement des données sur les aliments achetés, ce qui pourrait permettre d'identifier certains produits douteux. L'information sur la présence du consommateur dans ces données au point de vente peut, dans certains cas, faciliter l'identification d'autres personnes à risque par messagerie texte ou téléphone, de façon à obtenir des déclarations additionnelles d'éventuelles maladies d'origine alimentaire.

RÉSUMÉ DES COMMENTAIRES DES PRATICIENS EN SANTÉ PUBLIQUE – PATRICK SEITZINGER

Patrick Seitzinger, de l'Université de la Saskatchewan, a entrepris pour son mémoire de maîtrise en santé publique une étude pilote de l'application actuelle des mégadonnées dans les enquêtes sur les éclosions de maladies d'origine alimentaire au Canada. Un court sondage composé de quatre questions ouvertes a été remis, par le truchement de FluidSurvey, à des professionnels de la santé publique qui s'occupent actuellement d'enquêtes sur les éclosions de maladies d'origine alimentaire, entre autres des épidémiologistes, des microbiologistes et des chercheurs du Canada. Les questions ont évalué les perceptions et les utilisations réelles des mégadonnées, les lacunes dans la disponibilité des données et les idées d'applications nouvelles et créatives des mégadonnées dans les enquêtes sur les éclosions. Après avoir examiné le questionnaire et le protocole, le Comité d'éthique comportementale a accordé une exemption à l'Université de la Saskatchewan. Au total, 80 professionnels de la santé publique qui travaillent actuellement dans le domaine des enquêtes sur les éclosions de maladies d'origine alimentaire ont été contactés par courriel. Dix-huit participants de six provinces y ont répondu. La participation était anonyme, libre et volontaire, et chacun des participants a donné son consentement. Les thèmes, les idées et les concepts généraux qui se dégagent des réponses sont décrits ci-après.

i. Perceptions des mégadonnées dans le contexte des enquêtes sur les éclosions de maladies d'origine alimentaire

Les définitions des mégadonnées fournies par les participants tournaient autour de l'idée d'un ensemble de données volumineux et complexe composé de renseignements issus de nombreuses sources et utilisés à diverses fins pour définir des tendances, déceler des changements et prévoir des résultats. Les sources génériques de mégadonnées indiquées comprennent les moteurs de recherche (Google), les médias sociaux dans Internet (Twitter, Facebook) et les résumés analytiques de nouvelles. Les exemples de sources de données signalées comme des sources particulièrement pertinentes pour les enquêtes sur les éclosions de maladies d'origine alimentaire étaient les antécédents médicaux, les appels aux lignes infosanté, les données sur les ordonnances pharmaceutiques, les résultats de traçabilité des aliments et les données génomiques (obtenues par électrophorèse en champ pulsé et le séquençage pangénomique d'échantillons humains, animaux et alimentaires). Selon la compréhension des participants, le but de l'application des mégadonnées aux enquêtes sur les éclosions de maladies d'origine alimentaire était « d'améliorer la réaction aux éclosions de maladies [d'origine alimentaire] » (participant n^{o} 4) et « d'élargir ou de dépasser les capacités actuelles d'analyse des technologies de l'information » (participant n^{o} 7).

ii. Exemples de la façon dont les mégadonnées sont actuellement utilisées en pratique et en recherche en santé publique

Parmi les professionnels en santé publique qui ont répondu à cette question, 59 % (10/17) ont indiqué qu'ils avaient utilisé les mégadonnées directement en pratique ou en recherche en santé publique. Les exemples de mégadonnées utilisées pendant des enquêtes sur des éclosions précédentes ont été les suivants : renseignements des cartes de fidélité des consommateurs, données des pharmacies, données métagénomiques et information des rapports historiques d'éclosions et du Réseau canadien d'information sur la sécurité alimentaire (RCISA). Ces données ont été utilisées pour guider l'élaboration d'hypothèses et cibler les efforts de vérification et d'échantillonnage d'aliments. En ce qui concerne les systèmes de surveillance, les antécédents médicaux et les résultats de laboratoire ont été utilisés pour surveiller les tendances de la surveillance syndromique et prévoir différents résultats de santé. Des exemples de l'application des mégadonnées en recherche ont été donnés, par exemple « des projets de validation des principes pour évaluer les avantages et les inconvénients du séquençage pangénomique de bactéries par opposition aux techniques traditionnelles d'épidémiologie moléculaire » (participant n^{o} 10). Même si ces exemples illustrent un large éventail d'applications actuelles des mégadonnées dans la détection des éclosions, les interventions et la recherche, il est important de signaler que 15 % (3/17) des répondants ont indiqué qu'ils n'avaient utilisé les mégadonnées que de manière limitée et que 24 % (4/17) ont indiqué qu'ils n'avaient pas du tout utilisé les mégadonnées.

iii. Sources des mégadonnées actuellement disponibles pour faciliter les enquêtes sur les éclosions de maladies d'origine alimentaire au Canada

Lorsqu'on leur a demandé d'énumérer des sources précises de mégadonnées actuellement à la disposition des praticiens en santé publique au Canada, 35 % (6/17) des participants ont explicitement dit qu'ils étaient incertains des ressources actuellement à leur disposition. Les sources de mégadonnées que les praticiens connaissaient appartenaient à deux grandes catégories : les données gouvernementales (dossiers de la santé publique, Recensement, PulseNet Canada, le Laboratoire national de microbiologie, les systèmes intégrés d'information en santé publique provinciaux, le Réseau canadien de renseignements sur la santé publique (RCRSP), la surveillance de la résistance antimicrobienne, l'échantillonnage de produits alimentaires par l'organisme de réglementation et/ou les résultats de la traçabilité des produits alimentaires), et les données

recueillies par l'industrie (cartes de fidélité des consommateurs, données sur les parts de marché, GoogleFlu, GoogleTrends et les données historiques des pharmacies de Rx Canada). Un grand nombre de ces exemples s'accompagnaient de mises en garde, par exemple le manque de rapidité de diffusion de l'information et les difficultés à trouver des méthodes pertinentes d'exploration des données. Malgré la variété des exemples fournis, il s'en dégageait un sentiment général d'insuffisance des ressources et de l'information actuelles dans toutes les réponses au sondage.

iv. Sources de mégadonnées qui pourraient améliorer les enquêtes futures sur les éclosions de maladies d'origine alimentaire

Lorsqu'on leur a demandé quelles sortes de sources de données, si elles étaient mises à la disposition des praticiens de la santé publique et des chercheurs au Canada, amélioreraient le plus les enquêtes sur les éclosions de maladies d'origine alimentaire, les participants ont proposé un large éventail d'idées et de stratégies pour améliorer les systèmes actuels. Les participants ont souhaité pouvoir accéder à plus de données sur la traçabilité des produits, les messages des médias sociaux portant sur des problèmes de repas et de déclaration, des données quantitatives sur les recherches dans Google, les données des essais pour la conformité réglementaire, les données de surveillance et les statistiques sur la production alimentaire, des demandes de tests médicaux en temps réel et les motifs des consultations médicales. Un participant a parlé de l'avantage possible de données sur les ordonnances en pharmacie entièrement financées. En particulier, fournir « les emplacements géographiques des ordonnances de médicaments antidiarrhéiques et antinauséeux, jumelés aux tendances des isolats en laboratoire chez les humains » (participant n^{o} 11) semblait une stratégie prometteuse pour reconnaître et situer des éclosions en temps plus opportun. Les historiques de consommation des aliments, colligés par les applications et les registres alimentaires, constituaient une importante source d'information qui aiderait les enquêtes sur les éclosions de maladies d'origine alimentaire. Les données sur les épiceries constituaient également un moyen important d'améliorer les connaissances sur les parts de marché, car elles pouvaient fournir des dénominateurs à l'égard du nombre de personnes exposées à un certain produit.

Les participants ont réitéré les raisons pour lesquelles un grand nombre de ces options pourraient ne pas convenir, ne pas s'appliquer et/ou être opportunes. Les raisons les plus souvent citées étaient notamment les problèmes de protection de la vie privée, les limites juridiques, la réticence des entreprises à publier des données et le manque de financement pour obtenir et analyser ces données. Les participants ont à maintes

reprises parlé des obstacles à la fusion, à l'analyse, au stockage et à la visualisation. Ils ont aussi mentionné les difficultés suivantes : obtenir l'accès à l'information, résoudre les problèmes de formatage, le manque de formation et de compétences nécessaires pour interpréter les mégadonnées et, comme l'a dit le participant n° 10, [traduction] « actuellement, le personnel en santé publique et dans les laboratoires n'a pas les compétences pour analyser ou interpréter les "mégadonnées" et les groupes de technologies de l'information des gouvernements luttent pour obtenir le soutien des infrastructures des "mégadonnées" ». Les participants ont en outre souligné les problèmes de compétence qui existent déjà non seulement dans les organisations et les régions géographiques, mais parfois à l'intérieur même des ministères et des groupes. En dernier lieu, les participants ont insisté sur la nécessité de faire dûment compte de prudence dans l'interprétation des résultats des mégadonnées pour éviter les problèmes de sophisme écologique et de mal répartir les ressources.

En résumé, les praticiens de la santé publique dans ce domaine ont indiqué qu'ils appuyaient l'utilisation des mégadonnées dans les enquêtes sur les éclosions de maladies d'origine alimentaire. Ils ont souligné les risques prévisibles de même que les répercussions importantes. Les mégadonnées étaient généralement perçues comme un élément qui en valait la peine et qui avait le potentiel de [traduction] « contribuer à un "tableau" plus complet des problèmes en santé publique » (participant n° 7).

CONCLUSIONS

Les praticiens de la santé publique et les chercheurs devraient prendre soin d'éviter « la surestimation des mégadonnées » (33,38). Elles sont un complément, pas un substitut à la surveillance fondée sur la collecte de données traditionnelles. La santé publique doit s'orienter vers des systèmes hybrides qui améliorent l'opportunité et la profondeur de l'information plutôt que le remplacement de la surveillance traditionnelle (1). Un atelier sur les mégadonnées et l'analyse des maladies infectieuses, organisé par les National Academies of Science, Engineering, and Medicine, a conclu que les mégadonnées ne remplaceront pas la prise de décisions humaine, mais qu'elles peuvent être utilisées pour donner des connaissances qui peuvent améliorer l'efficacité et orienter les travaux de recherche futurs (14).

[Traduction] « Distinguer le signal véritable de la quantité gigantesque de bruit n'est jamais facile ni simple, mais c'est ce défi à relever si l'on veut que l'information soit un jour transformée en bien-être pour la société. » (111).

BIBLIOGRAPHIE

1. Bansal S, Chowell G, Simonsen L, Vespignani A, Viboud C. Big data for infectious disease surveillance and modeling. J Infect Dis. 2016 Nov 14;214(4):S375–9.

2. Fong D, Otterstatter M, Taylor M, Galanis E. Analysis of enteric disease outbreak metrics, British Columbia Centre for Disease Control. Can Commun Dis Rep. 2017 Jan 5;43(1):1–6.

3. Simonsen L, Gog JR, Olson D, Viboud C. Infectious disease surveillance in the big data era: Towards faster and locally relevant systems. J Infect Dis. 2016;214(4):S380–5.

4. Pisani A, Wyman P, Mohr D, Perrino T, Gallo C, Villamar J, et al. Human subjects protection and technology in prevention science: Selected opportunities and challenges. Prev Sci. 17(6):765–78.

5. Choi BCK. The past, present, and future of public health surveillance. Scientifica. 2012;2012:1–26.

6. Link MW, Battaglia MP, Frankel MR, Osborn L, Mokdad AH. Reaching the U.S. cell phone generation: Comparison of cell phone survey results with an ongoing landline telephone survey. Public Opin Q. 2007;71(5):814–39.

7. Hope K, Durrheim DN, d'Espaignet ET, Dalton C. Syndromic surveillance: Is it a useful tool for local outbreak detection? J Epidemiol Community Health. 2006 May;60(5):374–5.

8. Velsko S, Bates T. A Conceptual architecture for national biosurveillance: Moving beyond situational awareness to enable digital detection of emerging threats. Health Secur. 2016;14(3):189–201.

9. Flahault A, Bar-Hen A, Paragios N. Public health and epidemiology informatics. Yearb Med Inform. 2016;10(1):240–6.

10. Wesolowski A, Buckee CO, Engø-Monsen K, Metcalf CJE. Connecting mobility to infectious diseases: The promise and limits of mobile phone data. J Infect Dis. 2016 Dec 1;214(suppl_4):S414–20.

11. Links MG. Big Data is changing the battle against infectious diseases. Can Commun Dis Rep. 2015 Sep 3;41(9):215–7.

12. Lee EC, Asher JM, Goldlust S, Kraemer JD, Lawson AB, Bansal S. Mind the scales: Harnessing spatial big data for infectious disease surveillance and inference. J Infect Dis. 2016;214(4):S409–13.

13. G. V Asokan, Vanitha Asokan. Leveraging "big data" to enhance the effectiveness of "one health" in an era of health informatics. J Epidemiol Glob Health. 2015;5(4):311–4.

14. National Academies of Sciences, Engineering, and Medicine. Big data and analytics for infectious disease research, operations, and policy: Proceedings of a workshop. In: The National Academies of Sciences, Engineering, and Medicine (NASEM). Washington, DC: The National Academies Press; 2016.

15. Jonathan Cinnamon, Sarah K Jones, W Neil Adger. Evidence and future potential of mobile phone data for disease disaster management. Geoforum. 2016;75:253–64.

16. Sintchenko V, Holmes EC. The role of pathogen genomics in assessing disease transmission. BMJ [Internet]. 2015 May 11;350. Available from: http://www.bmj.com/content/350/bmj.h1314.abstract

17. Mather A, Reid S, Maskell D, Parkhill J, Fookes M, Harris S, et al. Distinguishable epidemics of multidrug-resistant Salmonella typhimurium DT104 in different hosts. Science. 2013;341(6153):1514–7.

18. Luo J, Wu M, Gopukumar D, Zhao Y. Big data application in biomedical research and health care: A literature review. Biomed Inform Insights. 2016 Jan 19;2016(8):1–10.

19. Oliver H, Abdo Z, Ricke S. Using WGS to protect public health and enhance food safety: Meeting Summary [Internet]. Available from: www.uspoultry.org/foodsafety/docs/WGS_Meeting_Summary_072916-02.pdf

20. Stasiewicz M, den Bakker H. Introduction to the interpretation of whole genome sequence data in food safety [Internet]. 2016. Available from: https://www.uspoultry.org/foodsafety/docs/WGS_pathogen_characterization_072916-03.pdf

21. Aarestrup F, Koompmans M. Sharing data for global infectious disease surveillance and outbreak detection. Trends Microbiol. 2016 Apr;24(4):241–5.

22. Jombart T, Cori A, Didelot X, Cauchemez S, Fraser C, Ferguson N. Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. PLOS Comput Biol. 2014 Jan 23;10(1):e1003457.

23. Deng X, den Bakker H, Hendriksen R. Genomic epidemiology: whole-genome-sequencing-powered surveillance and outbreak investigation of foodborne bacterial pathogens. Annu Rev Food Sci Technol. 2016 Feb;7:353–74.

24. Pettengill JB, Pightling AW, Baugher JD, Rand H, Strain E. Real-time pathogen detection in the era of whole-genome sequencing and big data: comparison of k-mer and site-based methods for inferring the genetic distances among tens of thousands of Salmonella samples. PLOS One. 2016;11(11):e0166162.

25. Tae H, Karunasena E, Bavarva JH, Garner HR. Updating microbial genomic sequences: improving accuracy & innovation. BioData Min. 2014;7(1):25.

26. Mehta S, Vinterbo S, Little S. Ensuring privacy in the study of pathogen genetics. Lancet Infect Dis. 2014;14(8):773–7.

27. Raza S, Luheshi L. Big data or bust: realizing the microbial genomics revolution. Microb Genomics. 2016;2(2):e000046.

28. Dion M, AbdelMalik P, Mawudeku A. Big data and the global public health intelligence network (GPHIN). Can Commun Dis Rep. 2015 Sep 3;41(9):209–14.

29. VELASCO E, AGHENEZA T, DENECKE K, KIRCHNER G, ECKMANNS T. Social Media and Internet-Based Data in Global Systems for Public Health Surveillance: A Systematic Review. Milbank Q. 2014 Mar;92(1):7–33.

30. Bahk CY, Scales DA, Mekaru SR, Brownstein JS, Freifeld CC. Comparing timeliness, content, and disease severity of formal and informal source outbreak reporting. BMC Infect Dis. 2015;15:135.

31. Chowell G, Cleaton JM, Viboud C. Elucidating transmission patterns from internet reports: Ebola and Middle East Respiratory Syndrome as case studies. J Infect Dis. 2016;214(4):S421–6.

32. MacFadden DR, Fisman D, Andre J, Ara Y, Majmuder MS, Bogoch II, et al. A platform for monitoring regional antimicrobial resistance, using online data sources: ResistanceOpen. J Infect Dis. 2016 Nov 14;214(4):S393–8.

33. Salathé M. Digital pharmacovigilance and disease surveillance: combining traditional and big-data systems for better public health. J Infect Dis. 2016 Nov 14;214(4):S399–403.

34. Nicholas Generous, Geoffrey Fairchild, Alina Deshpande, Sara Y Del Valle, Reid Priedhorsky. Global disease monitoring and forecasting with Wikipedia. PLoS Comput Biol. 2014;10(11):e1003892.

35. Bernardo MT, Rajic A, Young I, Robiadek K, Pham TM, Funk AJ. Scoping Review on Search Queries and Social Media for Disease Surveillance: A Chronology of Innovation. J Med Internet Res. 2013 Jul 18;15(7):e147.

36. Michael Edelstein, Anders Wallensten, Inga Zetterqvist, Anette Hulth. Detecting the norovirus season in Sweden using search engine data – Meeting the needs of hospital infection control teams. PLoS ONE. 2014;9(6):e100309.

37. Davidson MW, Haim DA, Radin JM. Using networks to combine "'Big Data'"and traditional surveillance to improve influenza predictions. Sci Rep. 2015;5:8154.

38. Lazer D, Kennedy R, King G, Vespignani A. The Parable of Google Flu: Traps in Big Data Analysis. Science. 2014;343(6176):1203–5.

39. Domnich A, Panatto D, Signori A, Lai PL, Gasparini R, Amicizia D. Age-related differences in the accuracy of web query-based predictions of influenza-like illness. Plos ONE. 2015 May 26;10(5):e0127754.

40. Martin L, Lee B, Yasui Y. Google Flu Trends in Canada: a comparison of digital disease surveillance data with physician consultations and respiratory virus surveillance data, 2010–2014. Epidemiol Infect. 2016;144(2):325–32.

41. Pelat C, Turbelin C, Bar-Hen A, Flahault A, Valleron A. More diseases tracked by using Google Trends. Emerg Infect Dis. 2009;15(8):1327–8.

42. Rishi Desai, Benjamin A Lopman, Yair Shimshoni, John P Harris, Manish M Patel, Umesh D Parashar. Use of internet search data to monitor impact of rotavirus vaccination in the United States. Clin Infect Dis. 2012;54(9):e115–8.

43. Kang JS, Kuznetsova P, Luca M, Choi Y. Where Not to Eat? Improving Public Policy by Predicting Hygiene Inspections Using Online Reviews. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing [Internet]. 2013. Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.431.7286&rep=rep1&type=pdf

44. Harrison C, Jorder M, Henri Stern, Stavinsky F, Reddy V, Hanson H, et al. Using online reviews by restaurant patrons to identify unreported cases of foodborne illness—New York City, 2012–2013. Morb Mortal Wkly Rep. 2014;63(20):441–5.

45. Nsoesie EO, Kluberg SA, Brownstein JS. Online reports of foodborne illness capture foods implicated in official foodborne outbreak reports. Prev Med. 2014 Oct;67:264–9.

46. Schomberg JP, Haimson OL, Hayes GR, Anton-Culver H. Supplementing public health Inspection via social media. PLoS ONE. 2016;11(3):e0152117.

47. Park H, Kim J, Almanza B. Yelp versus inspection reports: Is quality correlated with sanitation in retail facilities? J Environ Health. 2016;78(10):8–12.

48. Charles-Smith LE, Reynolds TL, Cameron MA, Conway M, Lau EH, Olsen JM, et al. Using social media for actionable disease surveillance and outbreak management: a systematic literature review. PLoS ONE. 2015;10(10):e0139701.

49. Gittelman S, Lange V, Gotway Crawford C, Okoro C, Lieb E, Dhingra S, et al. A new source of data for public health surveillance: Facebook likes. J Med Internet Res. 2015 Apr 20;17(4):e98.

50. Jurdak R, Zhao K, Liu J, AbouJaoude M, Cameron M, Newth D. Understanding human mobility from Twitter. PLOS ONE. 2015;10(7):e0131469.

51. Thapen N, Simmie D, Hankin C, Gillard J. DEFENDER: Detecting and forecasting epidemics using novel data-analytics for enhanced response. PLoS ONE. 2016;11(5):e0155417.

52. Hawkins J, Tuli G, Kluberg S, Harris J, Brownstein J, Nsoesie E. A Digital Platform for Local Foodborne Illness and Outbreak Surveillance. Online J Public Health Inform. 2016;8(1):e60.

53. Allen C, Tsou M-H, Aslam A, Nagel A, Gawron J-M. Applying GIS and machine learning methods to Twitter data for multiscale surveillance of influenza. Plos ONE. 2016 Jul 25;11(7).

54. Ivo D Dinov. Methodological challenges and analytic opportunities for modeling and interpreting Big Healthcare Data. GigaScience. 2016;5(1):1–15.

55. Krieck M, Dreesman J, Otrusina L, Denecke K. A new age of public health: Identifying disease outbreaks by analyzing tweets. In: Proceedings of Health WebScience Workshop. Koblenz, Germany; 2011.

56. Harris JK, Mansour R, Choucair B, Olson J, Nissen C, Bhatt J. Health department use of social media to identify foodborne illness-Chicago, Illinois, 2013-2014. Morb Mortal Wkly Rep. 2014;63(32):681–5.

57. Harris JK, Hawkins JB, Nguyen L, Nsoesie EO, Tuli G, Mansour R, et al. Using Twitter to Identify and Respond to Food Poisoning: The Food Safety STL Project. J Public Health Manag Pract [Internet]. 2017;Publish Ahead of Print. Available from: http://journals.lww.com/jphmp/Fulltext/publishahead/Using_Twitter_to_Identify_and_Respond_to_Food.99618.aspx

58. Sadilek A, Kautz H, DiPrete L, Labus B, Portman E, Teitel J, et al. Deploying nEmesis: Preventing Foodborne Illness by Data Mining Social Media. In 2016. Available from: https://www.aaai.org/ocs/index.php/IAAI/IAAI16/paper/view/11823

59. Talbot D. Big data from cheap phones [Internet]. MIT Technology Review. 2013. Available from: https://www.technologyreview.com/s/513721/big-data-from-cheap-phones/

60. Chen Y, Crespi N, Ortiz AM, Shu L. Reality mining: A prediction algorithm for disease dynamics based on mobile big data. Inf Sci. 2017 Feb 10;379:82–93.

61. Farrahi K, Emonet R, Cebrian M. Epidemic contact tracing via communication traces. Plos ONE. 2014;9(5):e95133.

62. Mohammad Hashemian, Dylan Knowles, Jonathan Calver, Weicheng Qian, Michael C Bullock, Scott Bell, et al. iEpi: an end to end solution for collecting, conditioning and utilizing epidemiologically relevant data. In: Proceedings of the 2nd ACM international workshop on Pervasive Wireless Healthcare, June 11-11, 2012. Hilton Head, South Carolina, USA; 2012. p. 3–8.

63. Mohammad Hashemian, Dylan Knowles, Kevin G Stanley, Jonathan Calver, Nathaniel Osgood. Human network data collection in the wild: the epidemiological utility of micro-contact and location data. In: Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium, January 28-30, 2012. Miami, Florida, USA; 2012.

64. Waldner C, Martin W, Osgood N, Seitzinger P, Teyhouee A, Relf-Eckstein J-A. Exploring New Technologies to Support Investigation of Foodborne Disease. 2016 Jun 14; Canadian Public Health Association.

65. Todd S, Diggle PJ, White PJ, Fearne A, Read JM. The spatiotemporal association of non-prescription retail sales with cases during the 2009 influenza pandemic in Great Britain. BMJ Open. 2014 Apr 29;4(4).

66. Muchaal P, Parker S, Meganath K, Landry L, Aramini J. Evaluation of a national pharmacy-based syndromic surveillance system. Can Commun Dis Rep. 2015 Sep 3;41(9):204–10.

67. Edge VL, Pollari F, Ng LK, Michel P, McEwen SA, Wilson J, et al. Syndromic Surveillance of Norovirus using Over-the-counter Sales of Medications Related to Gastrointestinal Illness. Can J Infect Dis Med Microbiol. 2006;17(4):235–41.

68. Edge VL, Pollari F, Lim G, Aramini J, Sockett P, Martin SW, et al. Syndromic Surveillance of Gastrointestinal Illness Using Pharmacy Over-the-Counter Sales: A Retrospective Study of Waterborne Outbreaks in Saskatchewan and Ontario. Can J Public Health Rev Can Santee Publique. 2004;95(6):446–50.

69. Swinkels H, Kuo M, Embree G, Fraser Health Environmental Health Investigation Team, Andonov A, Henry B, et al. Hepatitis A outbreak in British Columbia, Canada: the roles of established surveillance, consumer loyalty cards and collaboration, February to May 2012. Euro Surveill [Internet]. 19(18). Available from: https://www.ncbi.nlm.nih.gov/pubmed/24832119

70. Shah L, MacDougall L, Ellis A, Ong C, Shyng S, LeBlanc L, et al. Challenges of investigating community outbreaks of cyclosporiasis, British Columbia, Canada. Emerg Infect Dis. 2009;15(8):1286–8.

71. Norström M, Kristoffersen AB, Görlach FS, Nygård K, Hopp P. An adjusted likelihood ratio approach analysing distribution of food products to assist the investigation of foodborne outbreaks. PLoS ONE. 2015 Aug 3;10(8):e0134344.

72. Kaufman J, Lessler J, Harry A, Edlund S, Hu K, Douglas J, et al. A likelihood-based approach to identifying contaminated food products using sales data: performance and challenges. PLOS Comput Biol. 2014 Jul 3;10(7):e1003692.

73. Hu K, Edlund S, Davis M, Kaufman J. From Farm to Fork: How Spatial-Temporal Data can Accelerate Foodborne Illness Investigation in a Global Food Supply Chain. SIGSPATIAL. 2016;8(1):3–11.

74. Curtis L, Brown J, Platt R. Four health data networks illustrate the potential for a shared national multipurpose big-data network. Health Aff (Millwood). 2014;33(7):1178–86.

75. Ford L, Miller M, Cawthorne A, Fearnley E, Kirk M. Approaches to the surveillance of foodborne disease: A review of the evidence. Foodborne Pathog Dis. 2015 Dec 11;12(12):927–36.

76. Tomines A, Readhead H, Readhead A, Teutsch S. Applications of Electronic Health Information in Public Health: Uses, Opportunities and Barriers. Gener Evid Methods Improve Patient Outcomes [Internet]. 2013;1(2). Available from: http://repository.edm-forum.org/egems/vol1/iss2/5/

77. Wójcik O, Brownstein JS, Chunara R, Johansson MA. Public Health for the People: Participatory Infectious Disease Surveillance in the Digital Age. Emerg Themes Epidemiol. 2014;11:7.

78. Caroline Guerrisi, Clement Turbelin, Thierry Blanchon, Thomas Hanslik, Isabelle Bonmarin, Daniel Levy-Bruhl, et al. Participatory syndromic surveillance of influenza. Eur J Infect Dis. 2016;214(4):S386–92.

79. Swan M. Scaling crowd-sourced health studies: the emergence of a new form of contract research organization. Pers Med. 2012 Mar;9(2):223–34.

80. Mezghani E, Exposito E, Drira K, Da Silveira M, Pruski C. A semantic big data platform for integrating heterogeneous wearable data in healthcare. J Med Syst. 2015 Oct;39(1):185.

81. Fan S, Blair C, Brown A, Gabos S, Honish L, Hughes T, et al. A multi-function public health surveillance system and the lessons learned in its development: The Alberta Real Time Syndromic Surveillance Net. Can J Public Health. 2010;101(6):454–8.

82. Loveridge P, Cooper D, Elliot A, Harris J, Gray J, Large S, et al. Vomiting calls to NHS Direct provide an early warning of norovirus outbreaks in hospitals. J Hosp Infect. 2010 Apr;74(4):385–93.

83. Nicholas A S Hamm, Ricardo J Soares Magalhães, Archie C A Clements. Earth observation, spatial data quality, and neglected tropical diseases. PLoS Negl Trop Dis. 2015;9(12):e0004164.

84. Hay S, George D, Moyes C, Brownstein J. Big data opportunities for global infectious disease surveillance. PLoS Med. 2013;10(4):e1001413.

85. Messina JP, Kraemer MU, Brady OJ, Pigott DM, Shearer FM, Weiss DJ, et al. Mapping global environmental suitability for Zika virus. eLIFE. 2016 Apr 19;5:e15272.

86. Hilton BN. Overview of Spatial Big Data and Analytics [Internet]. 2015 Dec 13; Fort Worth, Texas. Available from: https://www.redlands.edu/globalassets/depts/school-of-business/gisab/workshops-conferences/brian-hilton-icis_2015_bnh.pdf

87. Liu S, Poccia S, Candan KS, Chowell G, Sapino ML. epiDMS: Data management and analytics for decision-making from epidemic spread simulation ensembles. J Infect Dis. 2016 Nov 14;214(4):S427–32.

88. Lal A. Spatial modelling tools to integrate public health and environmental science, illustrated with infectious Cryptosporidiosis. Int J Environ Res Public Health. 2016 Feb;13(2):1–8.

89. Moran KR, Fairchild G, Generous N, Hickmann K, Osthus D, Priedhorsky R, et al. Epidemic forecasting is messier than weather forecasting: The role of human behavior and internet data streams in epidemic forecast. J Infect Dis. 2016;214(4):S404–8.

90. Ramanathan A, Pullum L, Steed C, Quinn S, Chennubhotla C, Parker T. Integrating heterogeneous healthcare datasets and visual analytics for disease. In: 3rd IEEE Workshop on Visual Text Analytics. 2013.

91. Mitchell L, Ross JV. A data-driven model for influenza transmission incorporating media effects. R Soc Open Sci. 2016 Oct 26;3(10):160481.

92. Yang W, Lipsitch M, Shaman J. Inference of seasonal and pandemic influenza transmission dynamics. Proc Natl Acad Sci. 2015 Mar 3;112(9):2723–8.

93. Ola O, Sedig K. Beyond simple charts: Design of visualizations for big health data. Online J Public Health Inform. 2016;8(3):e195.

94. Carroll LN, Au AP, Detwiler LT, Fu T, Painter IS, Abernethy NF. Visualization and analytics tools for infectious disease epidemiology: A systematic review. J Biomed Inform. 2014 Oct;51:287–98.

95. Mazumdar S, Rushton G, Smith BJ, Zimmerman DL, Donham KJ. Geocoding accuracy and the recovery of relationships between environmental exposures and health. Int J Health Geogr. 2008;7(1):13.

96. Zandbergen PA. Ensuring confidentiality of geocoded health data: assessing geographic masking strategies for individual-level data. Adv Med. 2014 Apr 29;2014(1):567049.

97. Chen C, Epp T, Jenkins E, Waldner C, Curry P, Soos C. Modeling monthly variation of Culex tarsalis (Diptera: Culicidae) abundance and West Nile Virus infection rate in the Canadian prairies. Int J Environ Res Public Health. 2013;10(7):3033–51.

98. Varga C, Pearl DL, McEwen SA, Sargeant JM, Pollari F, Guerin MT. Evaluating area-level spatial clustering of SalmonellaEnteritidis infections and their socioeconomic determinants in the greater Toronto area, Ontario, Canada (2007 – 2009): a retrospective population-based ecological study. BMC Public Health. 2013;13(1):1078.

99. Greene SK, Peterson ER, Kapell D, Fine AD, Kulldorff M. Daily Reportable Disease Spatiotemporal Cluster Detection, New York City, New York, USA, 2014–2015. Emerg Infect Dis. 2016 Oct;22(10):1808–12.

100. Smith C, Le Comber S, Fry H, Bull M, Leach S, Hayward A. Spatial methods for infectious disease outbreak investigations: systematic literature review. Euro Surveill. 2015 Oct;20(39).

101. Musa GJ, Chiang P-H, Sylk T, Hoven CW. Use of GIS mapping as a public health tool-from cholera to cancer. Health Serv Insights. 2013 Nov 19;2013(6):111–6.

102. Jennifer L Gardy, James C Johnston, Shannan J Ho Sui, Victoria J Cook, Lena Shah, Elizabeth Brodkin, et al. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. N Engl J Med. 2011;364:730–9.

103. Lauren A Cowley, Stephen J Beckett, Margo Chase-Topping, Neil Perry, Tim J Dallman, David L Gally, et al. Analysis of whole genome sequencing for the Escherichia coli O157:H7 typing phages. BMC Genomics. 2015 Apr 8;16(1):271.

104. Muellner P, Shadbolt T, Collins-Emerson J, French NP. Molecular and spatial epidemiology of human Campylobacteriosis: source association and genotype-related risk factors. Epidemiol Infect. 2010;138(10):1372–83.

105. Grant E. The promise of big data [Internet]. Harvard T.H. Chan School of Public Health. 2012. Available from: https://www.hsph.harvard.edu/news/magazine/spr12-big-data-tb-health-costs/

106. Krumholz HM. Big data and new knowledge in medicine: the thinking, training, And tools needed for a learning health system. Health Aff (Millwood). 2014;33(7):1163–70.

107. Austin C, Kusumoto F. The application of big data in medicine: current implications and future directions. J Interv Card Electrophysiol. 2016;47(1):51–9.

108. Kruse C, Goswamy R, Raval Y, Marawi S. Challenges and opportunities of big data in health care: a systematic review. JMIR Med Inform. 2016 Nov 21;4(4):e38.

109. American Association for the Advancement of Science in conjunction with the Federal Bureau of Investigation and the United Nations Interregional Crime and Justice Research Institute (AAAS-FBI-UNICRI). National and Transnational Security Implications of Big Data in the Life Sciences [Internet]. 2014 Nov. Available from: https://www.aaas.org/report/national-and-transnational-security-implications-big-data-life-sciences

110. Cheung C, Bietz MJ, Patrick K, Bloss CS. Privacy attitudes among early adopters of emerging health technologies. Plos ONE. 2016;11(11):e0166389.

111. Khoury M, Ioannidis J. Medicine. Big data meets public health. Science. 2014 Nov 28;346(6213):1054–5.

112. Thomas M, Murray R, Flockhart L, Pintar K, Pollari F, Fazil A, et al. Estimates of the burden of foodborne illness in Canada for 30 specified pathogens and unspecified agents, Circa 2006. Foodborne Pathog Dis. 2013;10(7):639–48.

113. Thomas MK, Murray R, Flockhart L, Pintar K, Fazil A, Nesbitt A, et al. Estimates of Foodborne Illness–Related Hospitalizations and Deaths in Canada for 30 Specified Pathogens and Unspecified Agents. Foodborne Pathog Dis. 2015 Oct 1;12(10):820–7.

114. Belanger P, Tanguay F, Hamel M, Phypers M. An overview of foodborne outbreaks in Canada reported through Outbreak Summaries: 2008-2014. Can Commun Dis Rep. 2015 Nov 5;44(11):254–62.

115. Mecher T, Stauber C, Gould LH. Contributing Factors in a Successful Foodborne Outbreak Investigation: an Analysis of Data Collected by the Foodborne Diseases Active Surveillance Network (FoodNet), 2003-2010. [Internet]. Georgia State University; 2015. Available from: http://scholarworks.gsu.edu/iph_theses/382