Ecologie de la santé humaine : contribution à l'étude et à la surveillance des épidémies de gastro-entérite aigüe d'origine hydrique

Damien Mouly

UNIVERSITE BLAISE PASCAL
N° D.U. : 2706

UNIVERSITE D'AUVERGNE
ANNEE : 2016

ÉCOLE DOCTORALE DES SCIENCES DE LA VIE, SANTÉ, AGRONOMIE, ENVIRONNEMENT

N° d'ordre : 690

Thèse

Présentée à l'Université Blaise Pascal pour l'obtention du grade de

DOCTEUR D'UNIVERSITE
(Spécialité : **Écologie Microbienne**)

Soutenue le 23 juin 2016

Par Damien Mouly

Ecologie de la santé humaine : contribution à l'étude et à la surveillance des épidémies de gastro-entérite aigüe d'origine hydrique

---

Mr. Télesphore Sime-Ngando (Directeur de recherche, LMGE, CNRS), Président

Mr. Christian Ducrot (Directeur de recherche, Inra, EpiA), directeur de thèse

Mme. Anne Gallay (Docteur, Santé publique France, direction des régions), co-directrice de thèse

Mr. Didier Calavas (Docteur, Anses), rapporteur

Mr. Philippe Quenel (Professeur, EHESP, LERES), rapporteur

Adresse laboratoire d'accueil

INRA Clermont-Ferrand-Theix
Unité d'Épidémiologie Animale (UR346, département Santé Animale)
Route de Theix
63122 Saint Genès Champanelle

Adresse laboratoire de rattachement à l'Université Blaise Pascal

Laboratoire Microorganismes : Génome et Environnement, UMR CNRS 6023
Université Blaise Pascal
Campus Universitaire des Cézeaux
1 Impasse Amélie Murat
TSA 60026
CS 60026
63178 AUBIERE Cedex

REMERCIEMENTS

Je tiens à adresser mes remerciements chaleureux aux nombreuses personnes qui m'ont accompagné et soutenu durant ma thèse.

A Christian Ducrot et Anne Gallay d'avoir accepté de diriger cette thèse. La confiance qu'ils m'ont accordé, leur enthousiasme, leur bienveillance, leur disponibilité, leurs conseils toujours justes et avisés ont été d'un précieux soutien. Merci pour l'encadrement de qualité, l'expérience riche et partagée.

A Télesphore Sime-Ngando pour avoir accepté d'assurer le lien avec l'UBP et pour toute l'aide apportée lors de la préparation de mon mémoire et de ma soutenance.

A Philippe Quénel et Didier Calavas d'avoir accepté d'être rapporteurs de cette thèse.

Aux collègues de la Cire Occitanie (Amandine, Leslie, Franck, Cécile, Olivier, Cyril, Tiphanie, Anne, Jérôme, Mattéo) et aux voisins du DST (Eloi et Stéphanie) pour leur soutien, leurs encouragements et surtout leur patience ! Une mention particulière à Jérôme pour sa contribution technique et créative à l'élaboration de l'outil EpiGEH et à Franck pour sa disponibilité et son soutien au moment de la fusion des équipes.

Aux collègues de Santé publique France, en particulier Pascal Beaudeau et Henriette De Valk pour m'avoir initié et formé à la surveillance des maladies infectieuses portées par l'eau du robinet. Leur expertise du sujet, les connaissances transmises depuis plus de 10 ans ont été déterminants dans la réalisation de cette thèse. A Magali Corso et Grégoire Falq pour les giga-octets de cas de gastro-entérites aigües transmis ! Catherine Galey, Loïc Rambaud, Sarah Goria, Dieter Van Cauteren pour leur soutien et leur collaboration aux travaux de recherche. A Yann Le Strat pour ses remarques bienveillantes et constructives dans le cadre du comité de thèse et des articles. A Christel Guillaume et Asli Kilinc, toujours aux petits soins dans les démarches administratives !

Aux anciens collègues auvergnats. En particulier Nicolas Vincent et Emmanuelle Vaissière de la Cire, pour les investigations partagées, les nombreux échanges à la naissance du projet et leur aide dans les étapes de développement ; à Gilles Bidet de l'ARS pour m'avoir initié à l'investigation de terrain lors de l'épidémie de Pérignat, pour son soutien et son enthousiasme sans faille ! A Alain Blineau et Armelle Mathieu-Hermet de l'ARS pour l'intérêt porté à ce projet et leur soutien pour sa mise en œuvre au niveau régional.

A la jeune et dynamique équipe de l'unité EpiA pour les bons moments partagés à l'occasion des journées des doctorants. En particulier à Sylvain Coly pour sa contribution active aux travaux de recherche lors de son passage à la Cire et pour avoir assuré le lien avec les autres co-doctorants et membres de l'unité, pour ses idées profuses, les échanges riches et constructifs. A Gwenael Vourch, directrice de l'unité EpiA, pour avoir facilité la réalisation des travaux à distance dans un climat de confiance, serein et adapté. A Myriam Charras-Garrido et David Abrial pour leur disponibilité et leur contribution aux réflexions méthodologiques.

A Michael Mounié pour sa contribution aux travaux de simulation lors de son passage à la Cire et le partage de compétences. A Frédéric Bounoure, qui a ouvert la voie de la surveillance syndromique des gastro-entérites aigües à partir des données de l'Assurance Maladie, pour les échanges depuis 10 ans.

A Henri Davezac, en charge de l'administration de la base nationale de données SISE-eaux, pour son expertise et la fourniture des données. A l'Assurance Maladie pour la mise à disposition de ses données à Santé publique France qui a permis la réalisation des travaux présentés dans cette thèse.

A Thierry Cardoso et la direction de Santé publique France pour m'avoir fait confiance et permis de réaliser cette thèse dans un cadre professionnel adapté.

A Aurélie, Aubin, Noé, Romane pour leur soutien quotidien, leurs encouragements et leur patience.

RESUME

Les épidémies de gastro-entérite aigüe liées à l'eau du robinet demeurent un enjeu de santé publique au 21ème siècle dans les pays développés. La majorité des dispositifs de surveillance mis en place dans les pays se caractérisent par une sous-déclaration. En France, l'amélioration de la surveillance de ces évènements repose sur l'exploitation des données de l'Assurance Maladie.

L'objectif de notre travail est de proposer une méthode pour améliorer la sensibilité et la spécificité de la détection des épidémies de gastro-entérite aigüe d'origine hydrique. Trois études ont été menées pour i) évaluer la capacité des données de l'Assurance Maladie à décrire des épidémies de gastro-entérite aigüe d'origine hydrique, ii) adapter une méthode de détection spatio-temporelle en intégrant l'exposition à l'eau du robinet, iii) évaluer les performances de cette méthode.

Notre travail a permis de développer une méthode de détection spatio-temporelle des épidémies hydriques en tenant compte des unités géographiques de distribution d'eau avec une bonne sensibilité et une bonne valeur prédictive positive. Les performances de détection sont principalement liées à la taille de l'épidémie. La capacité des données de l'Assurance Maladie à détecter des épidémies d'origine hydrique peut être influencée par les habitudes de recours aux soins, la sensibilisation de la population au risque infectieux d'origine hydrique et le niveau d'agrégation temporel des cas de gastro-entérite aigüe.

La finalité de nos travaux de recherche est l'application opérationnelle pour la détection automatisée des épidémies hydriques dans une finalité de prévention et de réduction de leur impact sanitaire.

Mots-clés : surveillance, détection, épidémie d'origine hydrique, gastro-entérite aigüe, eau du robinet, Assurance Maladie

TITLE

Ecology of human health: contribution to the study and to the surveillance of waterborne disease outbreaks of gastrointestinal illness

ABSTRACT

Waterborne disease outbreaks (WBDO) of acute gastrointestinal illness remain a public health concern in the 21st century in developed countries. Almost all surveillance systems implemented in countries are characterized by underreporting. In France, the improvement of monitoring these events is based on the use of data from the French Health Insurance.

The aim of our work was to propose a method to improve the sensitivity and specificity of the detection of WBDO. Three studies were conducted to i) evaluate the ability of French Health Insurance data to describe WBDO ii) apply a method of space-time detection by integrating the exposure to tap water, iii) evaluate the performance of this method.

Our work allowed developing a method of space-time detection of WBDO by taking into account geographical units for drinking water networks units with a good sensitivity and positive predictive value. The performances for detection are mainly related to the size of the epidemic. The ability of the data from the French Health Insurance to detect WBDO can be influenced by the health-seeking behavior of people, awareness of the population to the risk of WBDO and the temporal aggregation level of cases of acute gastrointestinal illness.

The purpose of our research is to carry out an operational design for the automated detection of WBDO in a purpose of prevention and reduction of health impact.

Keywords: surveillance, detection, waterborne disease outbreak, acute gastrointestinal illness, drinking water, French Health Insurance

AVANT-PROPOS

Cadre de travail de la thèse

Ce travail de thèse a été conduit dans le cadre de mon activité professionnelle au sein de la direction des régions de Santé publique France (ex-Institut de veille sanitaire, InVS). Il recoupe les objectifs scientifiques de surveillance de la santé des populations de Santé publique France et a été réalisé en partenariat avec la direction santé environnement, en charge de la coordination nationale de la surveillance des risques d'origine hydrique, et la direction des maladies infectieuses en charge de la surveillance des épidémies de gastro-entérites aigües.

Les épidémies de gastro-entérites aigües d'origine hydrique font l'objet d'un programme de surveillance au sein de Santé publique France depuis sa création en 1998. Nos travaux contribuent au projet de détection automatisée d'agrégats spatio-temporels de gastro-entérite aigüe par contamination hydrique en France.

PRODUCTIONS SCIENTIFIQUES REALISEES DANS LE CADRE DE LA THESE

Articles en anglais (revues scientifiques à comité de lecture)

1/ **Mouly, D**., Van Cauteren, D., Vincent, N., Vaissiere, E., Beaudeau, P., Ducrot, C. & Gallay, A. 2016 Description of two waterborne disease outbreaks in France: a comparative study with data from cohort studies and from health administrative databases. *Epidemiol Infect.* 144, 591-601.

2/ Coly, S. Vincent N. Vaissière E. Charras-Garrido M. Gallay A. Ducrot C. **Mouly D.** Detection of Waterborne Disease Outbreaks: an integrated approach using health administrative databases. Journal of Water and Health (accepté le 12 juin 2016, en cours d'édition)

3/ **Mouly D**, Goria S, Mounié M, Rambaud L, Beaudeau P, Gallay A, Ducrot C, Le Strat Y. Detection of waterborne disease outbreak using health administrative databases: a simulation-based study (soumis à PlosOne le 4 septembre 2016 – accepté le 25 octobre 2016, en cours de révision)

Articles en français

4/ Nicolas Vincent, Emmanuelle Vaissière, Gilles Bidet, Alain Blineau, Sébastien Magne, Armelle Mathieu-Hermet, Alain Bruneau, Christine Louis, Thierry Chesnot, Benoît Gassillou, Sylvain Coly, Catherine Galey, **Damien Mouly**. Les risques infectieux d'origine hydrique en Auvergne. Bulletin de Veille Sanitaire, Cire Auvergne-Rhône-Alpes, n°1, février 2016.

Communications orales

5/ **Mouly D**. Adaptation de la méthode de Kulldorff à la détection des épidémies de GEA hydriques : test sur des données réelles et simulées. Séminaire bio-statistiques, Institut de veille sanitaire, 7 mai 2015.

6/ Coly S, **Mouly D**, Vincent N. Détection automatisée d'épidémies de gastro-entérite aigüe d'origine hydrique à partir des données de l'Assurance Maladie : définition d'une unité spatiale et test de la méthode de Kulldorff. 46èmes Journées de Statistique de la Société Française de statistique, ENSAI, Rennes, juin 2014.

Poster

7/ **Mouly D**, Coly S, Mounié M, Vincent N, Gallay A, Ducrot C. Evaluation d'une méthode de détection des épidémies de gastro-entérite aigüe d'origine hydrique en France à partir des données de l'Assurance Maladie. Les Journées de l'Ecole Doctorale (JED), Clermont-Ferrand, mai 2015.

Autres publications scientifiques issues de travaux précédents en lien avec le sujet

Articles en anglais

8/ Bounoure F., Beaudeau P., **Mouly D**., Skiba Mo., Skiba Ma. (2010) Syndromic surveillance of acute gastro-enteritis based on drug consumption, France. Epidemiol Infect. Nov 26:1-8.

9/ Beaudeau P., Valdes D., **Mouly D**., Stempfelet M. and Seux R. (2010) Natural and technical factors in faecal contamination incidents of drinking water in small distribution networks, France, 2003–2004: a geographical study. J. Water and Health. Vol 08 No 1 pp 20–34

10/ Beaudeau P, De valk H, Vaillant V, Mannschott C, Tillier C, **Mouly D**, and Ledrans M. Lessons (2008) Learned from 10 investigations of waterborne gastroenteritis outbreaks, France, 1998-2006. J.Water Health 06.4 :491-503.

Articles en français

11/ Rambaud L., **Mouly D**., Schmitt M., Kerrien F., Beaudeau P. (2011) Utilisation des données de remboursement des médicaments de l'Assurance maladie pour identifier et caractériser une épidémie de gastroentérites d'origine hydrique, Bourg Saint-Maurice (Arcs 1800), 2006. Bull Epid Hebd n°31 – 6 septembre 2011.

Communication orale

12/ Catherine Galey, Grégoire Falq, Agnès Guillet, Christel Lamat, **Damien Mouly** et Pascal Beaudeau. Surveillance des épidémies de gastroentérites aigües d'origine hydrique, France : Connexion avec la gestion de terrain. JIE Apten 2012 (colloque 25-26 et 27 septembre 2012). Communication orale

Comité de thèse

Un comité a été réuni deux fois durant la thèse :

Le 27 novembre 2013 à l'InVS, Saint-Maurice

Le 7 avril 2015 à l'InVS, Saint-Maurice

Membres du comité de thèse

Beaudeau Pascal (Santé publique France – ex-InVS, direction santé-environnement)

Frédéric Bounoure (Faculté de pharmacie de Rouen)

Henriette De Valk (Santé publique France – ex-InVS, direction des maladies infectieuses)

Benoît Gassilloud (Anses, laboratoire d'hydrologie)

Yann Le Strat (Santé publique France – ex-InVS, direction des maladies infectieuses)

Télesphore Sime-Ngando (LMGE, Laboratoire Microorganismes : Génome et Environnement, UMR CNRS 6023, Université Blaise Pascal, Clermont-Ferrand)

TABLE DES MATIERES

Introduction

Le risque lié à l'eau de distribution (ou eau du robinet) est présent dans les pays développés comme la France. Il peut être la conséquence de contaminations chimiques ou microbiologiques et s'exprimer sous forme chronique ou endémique, aigüe ou épidémique, sub-chronique ou hyper-endémique. Leur étude ainsi que la mise en place de dispositifs de surveillance permettent de réduire l'impact sanitaire et de prévenir les maladies portées par l'eau du robinet.

Les travaux présentés dans ce mémoire se concentrent sur le risque épidémique de nature infectieuse lié à l'eau du robinet.

La contamination de l'eau de distribution par des agents pathogènes est susceptible d'entraîner des épidémies d'ampleurs variables en fonction du niveau de pollution, de la taille de la population desservie par les réseaux d'eau contaminés et de la proportion de personnes utilisant cette eau pour la boisson, le lavage des aliments ou l'hygiène (Tillett 1998). Bien que les épidémies d'origine hydrique ne représentent pas le seul risque sanitaire lié à l'eau, leur étude donne un aperçu des causes et des facteurs de risque pouvant contribuer à la survenue de maladies pour lesquelles le lien avec de l'eau contaminée a pu être établi. Ce lien est établi grâce à des données de laboratoire et des données épidémiologiques (Hrudey & Hrudey 2004).

Historiquement, l'épidémie de choléra survenue à Londres en 1854 reste à ce jour la plus meurtrière parmi les épidémies hydriques rapportées (plus de 600 décès estimés) (Smith 2002). L'épidémie de cryptosporidiose survenue à Milwaukee (Wisconsin, Etats-Unis) en 1993 demeure quant à elle la plus importante avec 403 000 personnes atteintes dont 4 400 ont été hospitalisées et 50 sont décédées (MacKenzie 1995; Hoxie 1997; Corso 2003). Parmi les autres épidémies historiques, on peut également citer l'épidémie à *Escherichia coli* O157:H7 et *Campylobacter* survenue en 2000 à Walkerton (Ontario, Canada) (Canada 2000) qui a entraîné plusieurs décès et plus de 2 000 malades. Les nombreuses publications à la suite de ces deux épidémies, jusqu'à plus de 10 ans après leur survenue (18 articles identifiés pour Milwaukee (Mac Kenzie 1994; MacKenzie 1995; Addiss 1996; Morris 1996; Vakil 1996; Cicirello 1997; Cordell 1997; Hoxie 1997; Manthey 1997; Eisenberg 1998; Morris 1998; McDonald 2001; Corso 2003; Naumova 2003; Zhou 2003; Gupta & Haas 2004; Eisenberg 2005) et 20 pour Walkerton (Canada 2000; Journal 2000; Glouberman 2001; Hrudey & Hrudey 2002; Krewski 2002; Mackay 2002; McQuigge 2002; Ritter 2002; Brown & Hussain 2003; Clark 2003; Holme 2003; Hrudey 2003; Ali 2004; Auld 2004; Marshall 2004; Clark 2005; Richards 2005; Matsell & White 2009; Salvadori 2009; Wang & Chang 2011)), témoignent de leur impact sanitaire, social et scientifique. En Europe, de nombreuses épidémies hydriques ont pu également être recensées et étudiées (Nazareth 1994; Poullis 2002;

Koutsotoli 2006; Beaudeau 2008; Daures 2011; Rambaud 2011; Mouly 2013; Murphy 2014). Une revue globale montre qu'en dehors de quelques épidémies qui ont marqué l'histoire par leur intensité, la majorité d'entre elles surviennent dans des réseaux d'eau de petite ou moyenne taille (entre moins de 500 habitants à quelques milliers) (Beaudeau 2008; Beaudeau 2012a; Pons 2015). De même, la plupart des épidémies rapportées sont de type gastro-entérites aigües (Craun 2010). Enfin, des épidémies d'infection liées à l'exposition à des eaux récréatives sont également rapportées dans la littérature mais ne seront pas abordées dans ce travail.

La quantification de l'impact sanitaire attribuable aux épidémies d'origine hydrique reste difficile à estimer tant au niveau local, régional que national en raison de l'hétérogénéité des méthodes d'estimation et de surveillance (Murphy 2014). Aux Etats-Unis, Messner (Messner 2006) et Colford (Colford 2006) ont proposé des méthodes pour estimer à un niveau national le poids des maladies portées par l'eau du robinet. Selon la méthode utilisée, le nombre annuel moyen se situe entre 3,1 et 18,5 millions cas de gastro-entérite aigüe (Murphy 2014).

Le coût économique des épidémies d'origine hydrique lié à la prise en charge des malades (traitement, hospitalisations, arrêts de travail) et à la diminution de la productivité des travailleurs malades est variable. Les quelques études qui se sont penchées sur la question ont estimé que le coût de l'épidémie de cryptosporidiose de Milwaukee s'élevait à près de 96 millions de dollars (Corso 2003), celui d'une épidémie liée à une contamination par des eaux usées en Finlande (Laine 2011) à près de 2 millions d'euros (Halonen 2012) ; enfin, celui d'une épidémie liée à un retour d'eaux usées au Danemark à environ 200 000 euros (1,6 millions Danish Kroner) (Laursen 1994). En tenant compte du nombre de personnes atteintes, le coût par personne malade serait d'environ 135 à 300 euros selon les épidémies rapportées. Au Canada, Payment (Payment 1997) a estimé que le poids des maladies portées par l'eau du robinet pouvait coûter chaque année plusieurs millions de dollars aux canadiens. La part des épidémies de gastro-entérite aigüe d'origine hydrique n'est pas connue précisément.

Par ailleurs, les épidémies investiguées de par le monde (Etats-unis (Craun 2010), Canada (Pons 2015), Europe (Furtado 1998; Beaudeau 2008), Chine (Yang 2011)) ont permis d'identifier les principaux facteurs de risque associés : phénomènes environnementaux (contamination de la ressource à la suite du ruissellement d'eaux de pluies et d'une protection des captages d'eau potable insuffisante), insuffisance des systèmes de traitement d'eau (sous dimensionnement des filières de traitement, système de désinfection absent ou inadapté), incidents dans le traitement ou la distribution de l'eau (panne de filtration ou de désinfection, rupture de canalisations, retours d'eaux usées, etc.). Dans la plupart des cas, la mise en place de moyens techniques et de dispositifs de surveillance adaptés au

niveau de la ressource (protection captage), du traitement (système de pré-traitement, filtration, désinfection), ou de la distribution de l'eau permettraient de réduire l'impact de ces épidémies.

Ainsi, le nombre d'épidémies signalées au cours des dernières années montre que la transmission d'agents pathogènes par l'eau potable demeure une préoccupation sanitaire importante et que, en dépit des progrès réalisés ces dernières années dans le domaine du traitement de l'eau et des dispositifs de contrôle/surveillance, l'accès à une eau potable sûre reste un enjeu de santé publique dans les pays développés comme la France. L'existence de moyens de prévention connus renforce l'intérêt en santé publique de surveiller les épidémies d'infection d'origine hydrique.

Dans ce contexte, de nombreux pays ont mis en place une stratégie de surveillance des épidémies d'origine hydrique afin de limiter leur impact et de permettre l'identification des facteurs de risque dans le but d'améliorer les pratiques et la prévention de ces évènements. Les dispositifs de surveillance et les moyens de détection actuels reposent essentiellement sur l'identification d'un signal révélant des cas groupés d'infections (le plus souvent de type gastro-entérite aigüe) par des professionnels de santé, des responsables d'établissements recevant du public (écoles) ou des représentants de la population ; et/ou d'un signal environnemental (analyse d'eau non conforme, incident d'exploitation) par l'exploitant ou l'autorité sanitaire en charge du contrôle de la qualité de l'eau (Risebro & Hunter 2007).

Face à une sous-déclaration manifeste de ces évènements par les systèmes de surveillance traditionnels évoqués ci-dessus, Santé publique France (ex-InVS) explore depuis plusieurs années l'utilisation des données de l'Assurance Maladie pour améliorer le dispositif de surveillance. Des travaux précédents, réalisés en collaboration avec l'Université de Rouen, ont notamment abouti à la création d'un algorithme qui permet d'identifier les cas de gastro-entérite aigüe médicalisés ayant fait l'objet d'une prescription médicale suivi d'un achat de médicaments remboursés à la pharmacie (Bounoure 2011). D'autres travaux ont montré l'intérêt de la surveillance syndromique des cas de gastro-entérites aigües médicalisés pour prévenir le risque infectieux lié à l'ingestion d'eau du robinet (Beaudeau 2012a).

Notre démarche se situe dans le cadre scientifique général de l'écologie de la santé humaine. Elle vise à améliorer la sensibilité et la spécificité de la détection des épidémies d'origine hydrique en partant des travaux réalisés pour l'identification des cas de gastro-entérites aigües médicalisées à partir des données de l'Assurance Maladie.

L'objectif principal de notre travail présenté dans ce mémoire est de proposer une démarche scientifique pour améliorer l'étude et la surveillance des épidémies d'origine hydrique à partir des cas de

gastro-entérite aigüe médicalisés. Les objectifs spécifiques de la thèse sont les suivants : i) étudier les capacités des données de l'Assurance Maladie à décrire des épidémies de gastro-entérite aigüe d'origine hydrique connues ; ii) développer une méthode intégrée pour permettre la détection des épidémies d'origine hydrique, de façon automatisée, à partir des cas de gastro-entérite aigüe médicalisés, en tenant compte de l'exposition à l'eau du robinet des cas ; iii) évaluer, sur la base d'une étude de simulation, les performances et les facteurs influençant la détection des épidémies d'origine hydrique.

La finalité de nos travaux est de contribuer à la mise en place d'un système de détection automatisé des agrégats de cas de gastro-entérites aigües médicalisés liés à la consommation d'eau du robinet. Ce dispositif devrait permettre d'améliorer la connaissance de l'impact sanitaire de ces épidémies et des circonstances de leur survenue. Le système de détection automatisé devrait également apporter un support décisionnel pour formuler des préconisations concernant la gestion des réseaux d'eau potable identifiés comme étant fragiles face au risque de contamination, dans une finalité de prévention et de réduction de l'impact sanitaire.

La première partie de ce mémoire présente une synthèse bibliographique sur les épidémies d'origine hydrique : définitions, systèmes de surveillance et tendances épidémiologiques. Elle présente également les données de l'Assurance Maladie, les données d'exposition à l'eau du robinet disponibles et les possibilités de leur intégration dans la détection, le choix d'une méthode statistique adaptée. La deuxième partie décrit la démarche scientifique en trois étapes, chacune ayant conduit à la production d'un article scientifique. Enfin, la troisième et dernière partie présente une discussion des résultats et des perspectives de ce travail dans un objectif d'amélioration de la surveillance épidémiologique des épidémies d'origine hydrique et d'amélioration de la gestion des systèmes de production d'eau potable en France.

Première partie : synthèse bibliographique, données et outils disponibles

1 Les épidémies liées à l'eau du robinet : synthèse bibliographique

1.1 Définitions des épidémies d'origine hydrique

1.1.1 Définitions

Une épidémie correspond à une augmentation, en général rapide, du nombre de cas d'une même maladie, le plus souvent de nature infectieuse, au-dessus de ce qui est normalement attendu dans une population et un lieu donnés. Les sources et modes de transmission d'une épidémie peuvent être multiples (alimentaire, hydrique, inter-humain, zoonotique, vectorielle, etc.). La notion d'épidémie peut donc se rapporter à un symptôme ou un groupe de symptômes (expression de la maladie), indépendamment de la cause et du mode de transmission.

Il n'existe pas dans la littérature une définition universelle des épidémies d'origine hydrique. L'Organisation mondiale de la santé (OMS) retient une définition commune pour les épidémies d'origine hydrique et alimentaire : *« au moins deux personnes atteintes d'une même maladie après l'ingestion d'un même type d'aliment ou d'eau provenant de la même source et pour lesquelles les investigations épidémiologiques impliquent l'aliment ou l'eau comme origine de la maladie » (Schmidt 1995)*. Dans certains pays qui s'intéressent à la surveillance des épidémies d'origine hydrique, des définitions plus ou moins précises sont utilisées. Par exemple, les Etats Unis retiennent comme définition la notion de *« deux personnes ou plus, atteintes d'une maladie semblable survenue après une exposition à l'eau, liées épidémiologiquement dans le temps et au même lieu d'exposition à l'eau »* (Craun 2010). Cette définition, proche de celle de l'OMS, inclut les épidémies d'origine hydrique quels que soient l'origine de la maladie (infectieuse ou chimique) et le type d'exposition (eau potable, eau récréative incluant les piscines). Ainsi, par exemple, une épidémie de gastro-entérite aigüe d'origine infectieuse (ex : *Campylobacter, Salmonella, Giardia*) ou chimique (ex : cuivre, fluorure) liée à l'ingestion d'eau d'une commune desservie par un même réseau d'eau entre dans cette définition ; de la même façon qu'une épidémie de légionellose (infection respiratoire) liée à une exposition aux douches d'un même camping ou une épidémie d'infections cutanées liées à la fréquentation d'une même piscine. Au Canada, plusieurs définitions ont été utilisées lors de la réalisation de bilans de surveillance. Ainsi, Schuster en 2005 retient la définition d'une épidémie comme *« l'apparition d'au moins deux cas de maladie qui surviennent après l'ingestion d'eau provenant d'une même source d'eau potable»* (Schuster 2005). Par la suite, Wilson en 2009 utilise une définition large pour les épidémies d'origine hydrique : *« toute maladie aigüe suspectée ou confirmée liée à l'exposition à un agent biologique, chimique ou radiologique présent dans l'eau potable et impliquant au moins deux individus »* (Wilson 2009). En Europe, seuls l'Angleterre et le Pays de Galles ont adopté une définition officielle pour les épidémies

d'origine hydrique : *« une épidémie de maladie infectieuse intestinale pour laquelle les données épidémiologiques ou microbiologiques montrent que l'eau est la source la plus probable »* (Furtado 1998). Cette définition, restreinte aux épidémies infectieuses intestinales, peut inclure plusieurs sources d'exposition à l'eau (réseau d'eau potable, piscines). En France, les toxi-infections alimentaires collectives (Tiac) inscrites à la liste des maladies à déclaration obligatoire depuis 1986[1], sont définies comme *« la survenue d'au moins 2 cas similaires d'une symptomatologie, en général gastro-intestinale, dont on peut rapporter la cause à une même origine alimentaire »* (Delmas 2010). En considérant l'eau de boisson comme une source alimentaire, une épidémie de gastro-entérite aigüe d'origine hydrique liée à l'ingestion d'eau contaminée répond à la définition d'une Tiac, qu'elle soit de nature infectieuse ou chimique.

Par abus de langage, les termes « épidémie de gastro-entérite aigüe d'origine hydrique » ou « épidémie d'origine hydrique » sont souvent employés car la plupart des épisodes rapportés concernent des cas de gastro-entérite aigüe. Ils englobent la notion de phénomène épidémique d'un symptôme (la gastro-entérite aigüe) pouvant être la conséquence d'une contamination par un agent infectieux à la suite d'ingestion d'eau du robinet. Sauf mention particulière, nous utiliserons ces deux définitions simplifiées dans les parties suivantes.

1.1.2 Critères de classification

Pour compléter ces définitions, plusieurs critères sont proposés pour évaluer le niveau de vraisemblance de l'origine hydrique d'une épidémie. Ces critères sont de nature épidémiologique, microbiologique et liés au traitement ou à la qualité de l'eau. Là encore, il n'existe pas de consensus et deux types de classification sont utilisés :

- la classification adoptée par le Center for Disease Control and Prevention (CDC) aux Etats-Unis qui propose 4 classes en fonction des informations fournies par les données épidémiologiques et les données sur la qualité d'eau - du niveau de preuve maximum (classe 1) au niveau de preuve minimum (classe 4) (Figure 1) (Craun 2010) ;
- la classification développée par le Communicable Disease Surveillance Center (CDSC) au Pays de Galles qui retient 3 niveaux d'association entre l'exposition à l'eau et la survenue de l'épidémie - possiblement, probablement ou fortement associé (Figure 2) (Tillett 1998).

[1] Décret numéro 86-770 du 10 juin 1986

La classification américaine donne un poids plus important aux données épidémiologiques qu'aux données relatives à la qualité de l'eau alors que la classification galloise donne un poids équivalent.

Class	Epidemiologic data	Water quality data
I	Adequate; data provided about exposed and unexposed persons, with relative risk or odds ratio of ≥2 or *P* value of ≤0.05	Provided and adequate; laboratory data or historical information (e.g., reports of a chlorinator malfunction, a water main break, no detectable free-chlorine residual, or the presence of coliforms in water)
II	Adequate	Not provided or inadequate (e.g., laboratory testing of water not conducted and no historical information)
III	Provided but limited Epidemiologic data provided that did not meet the criteria for class I, or claim made that ill persons had no exposures in common besides water but no data provided	Provided and adequate
IV	Provided but limited	Not provided or inadequate

Figure 1 : Classification des épidémies liées à l'eau basée sur la solidité des preuves impliquant l'eau comme véhicule de transmission (source : (Craun 2010))

	Strength of association	
Microbiology	Pathogen identified in patient is also found in water	A
	Indicator organisms and/or water-treatment problem of relevance but outbreak pathogen is not detected in water	B
PLUS	PLUS	
Epidemiology	Analytical epidemiology (case control or cohort) study demonstrates association between water and illness	C
	Descriptive epidemiology suggests that the outbreak is water related and excludes obvious alternative explanations	D
	↓	
Strong association	(A + C) or (A + D) or (B + C)	
Probable association	(B + D) or A only or C only	
Possible association	B only or D only	

Figure 2 : Classification selon les critères du Communicable Disease Surveillance Center (CDSC) au Pays de Galles (source : (Smith 2006)).

1.2 Les données de surveillance dans les pays développés et les tendances épidémiologiques

1.2.1 Sources de données et schéma conceptuel

La surveillance des épidémies d'origine hydrique relève d'une approche intégrée qui nécessite de compiler des données sanitaires (cas de maladies), des données environnementales (indicateurs de pollution de l'eau distribuée), et qu'un lien soit établi entre ces deux types d'information. Les données sanitaires et environnementales peuvent être de natures variées et détenues par différents acteurs. Ces contraintes se traduisent par des stratégies de surveillance variables d'un pays à l'autre en fonction des données disponibles et de l'organisation des agences sanitaires et organismes de santé publique. Plusieurs exemples seront détaillés dans les parties suivantes.

Dans les pays d'Amérique du nord (Etats-Unis, Canada) et d'Europe les sources de données utilisées pour documenter et surveiller les épidémies d'origine hydrique, incluent les signalements de cas groupés de malades par des professionnels de santé ou d'évènements sanitaires inhabituels dans des établissements recevant du public (surveillance passive), les données des laboratoires, les données des systèmes de surveillance syndromique (données médico-administratives), les enregistrements des lignes d'appels téléphoniques d'urgence, les données des ventes de médicaments ou encore des données environnementales témoignant d'une contamination de l'eau (résultats d'analyse, incidents de traitements d'eau, plaintes de consommateurs, etc.) (Figure 3). En dehors des Etats-Unis qui ont développé un système de surveillance dédié aux épidémies d'origine hydrique et produisent des rapports réguliers, la plupart des autres pays réalisent des bilans épidémiologiques ponctuels sur les épidémies d'origine hydrique.

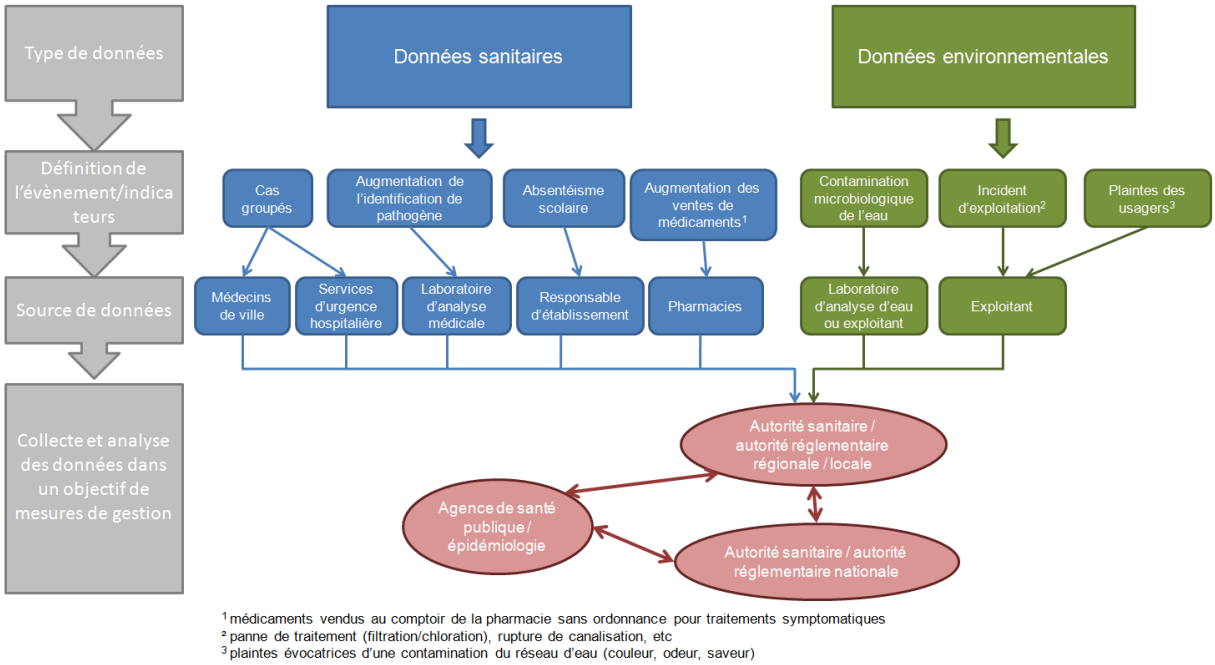


Figure 3 : Exemples de données utilisées, de leur transmission et des acteurs impliqués dans la surveillance des épidémies d'origine hydrique.

1.2.2 Exemples par pays

1.2.2.1 Les Etats-Unis

Les Etats-Unis, pionniers dans ce domaine, enregistrent des données sur la survenue et les causes des épidémies d'origine hydrique depuis les années 1920 (Gorman & Wolman 1939; Eliassen & Cummings 1948; Weibel 1964; Craun & McCabe 1973). Depuis 1971, le CDC, l'U.S. Environmental Protection Agency (US-EPA) et le Council of State and Territorial Epidemiologists (CSTE) collaborent pour coordonner le Waterborne Disease and Outbreak Surveillance System (WBDOSS) et effectuent des

rapports bisannuels (Figure 4). Le dernier bilan publié en 2015 porte sur les années 2011-2012 (Beer 2015). Ce système repose sur la déclaration volontaire par les agences de santé publique des Etats et territoires au CDC d'évènements répondant à des définitions et des critères d'inclusion. Ces derniers sont régulièrement mis à jour depuis la création du système en 1971 (Figure 5). Les informations demandées comprennent les caractéristiques des épidémies, les données épidémiologiques, les résultats des investigations environnementales et d'exploitation de l'eau. Les rapports d'épidémies sont ensuite évalués par l'US-EPA et le CDC pour déterminer si les informations sont suffisantes pour incriminer l'eau comme source de contamination (cf. critères de classification de Craun 2010 – paragraphe Définitions et critères). Les épidémies avec des données environnementales limitées sont inclues dans la base de données pour la surveillance (WBDOSS) mais pas celles dont les données épidémiologiques sont insuffisantes.

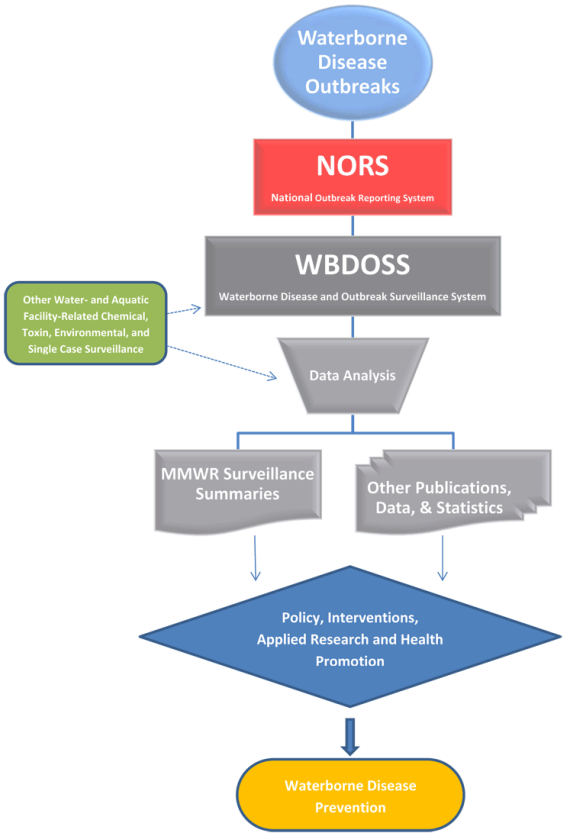


Figure 4 : Les sources de données et les sorties du système de surveillance des épidémies d'origine hydrique (source : http://www.cdc.gov/healthywater/surveillance/)

Reporting period	WBDOSS definitions and inclusion criteria
1971–1972	WBDOSS initiated: "outbreak" defined as two or more cases epidemiologically linked to consumption of water from municipal, semipublic or individual water systems; individual water system defined as wells or springs used exclusively by single residences in areas without municipal systems
1974	Inclusion of single cases of chemical poisoning when drinking water was demonstrated to be contaminated by a chemical
1976	Individual water systems redefined as wells or springs used by a single residence or several residences or by persons traveling outside populated areas
1979	Drinking water systems redefined as community systems, noncommunity systems, and individual systems
1989–1990	Total no. of cases redefined to exclude secondary cases
1991–1992	Specific exclusion of outbreaks due to contamination of water or ice at point of use
1995–1996	Estimated case count used instead of actual case count when the study population was randomly sampled or the estimated count was calculated using the attack rate
1999–2000	Inclusion of outbreaks associated with occupational water; inclusion of water not intended for drinking and bottled water in individual water systems
2001–2002	Inclusion of outbreaks of Legionnaires' disease
2003–2004	Introduction of expanded deficiency classifications that capture point-of-use outbreaks except contamination of ice; removal of water not intended for drinking and bottled water outbreaks from individual water system classification; revision to definition of etiologic agent (multiple etiologies listed when each agent individually represents ≥5% of positive specimens); "unidentified" is now used instead of "AGI" to identify acute gastrointestinal illness of unknown etiology; illness types listed when ≥50% of patients reported a symptom in that category
2005–2006	Deficiency classifications expanded to include a deficiency whereby current treatment is not expected to remove a chemical contaminant; single cases excluded from analyses of outbreaks

Figure 5 : Chronologie des définitions du WBDOSS et des critères d'inclusion concernant l'eau potable (source : (Craun 2010)).

Un bilan de 38 années de données de surveillance issues de ce système permet d'évaluer l'importance des épidémies d'origine hydrique, d'identifier les principaux facteurs de risque, d'analyser les tendances et d'évaluer l'efficacité des actions mises en place (Craun 2012). Dans ce bilan, 733 épidémies sont associées à une contamination de l'eau potable. Ces épidémies totalisent 579 582 malades et 116 décès (dont 403 000 malades et 50 décès pour l'épidémie de Milwaukee) (MacKenzie 1995). Leur nombre décroit significativement après les années 1980 ($p<10^{-4}$, Figure 6) et on observe une saisonnalité, la plupart des épidémies survenant l'été. Parmi les explications possibles de cette tendance décroissante sont cités : la mise en place de la réglementation nationale sur l'eau potable, les changements dans les pratiques et les modes de gestion des systèmes de distribution d'eau, l'amélioration des infrastructures dans le domaine de l'eau potable (en particulier les systèmes de traitement).

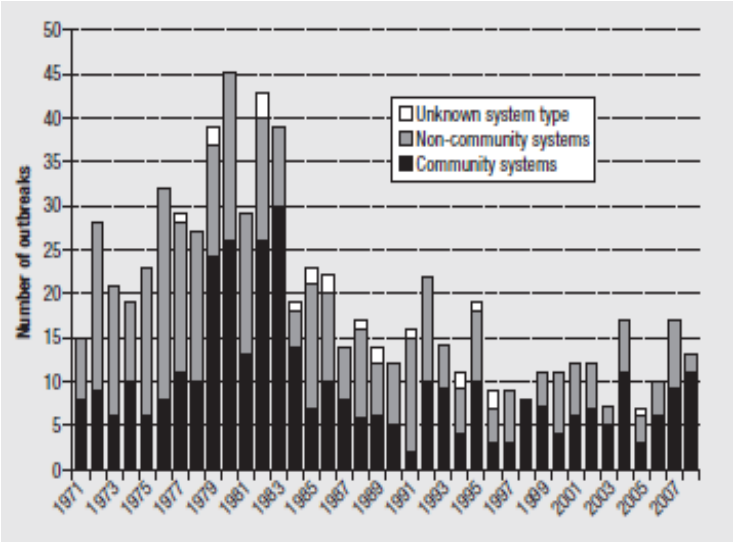


Figure 6 : Nombre d'épidémies d'origine hydrique associées à un réseau d'eau publique aux Etats-Unis, par année de 1971 à 2008 (NB : les systèmes communautaires et non communautaires sont des systèmes publics. Ils sont majoritaires) (source : (Craun 2012)).

Sur les 733 épidémies répertoriées, la majorité (n=660 soit 90% représentant 573 239 cas) sont dues à des agents pathogènes entraînant des infections digestives (n=606 épidémies, 572 767 cas) et des infections respiratoires dues à *Legionella* (n=54 épidémies, 472 cas). Les autres épidémies (n=73 soit 10%) sont associées à une contamination chimique. Les principales maladies rapportées sont des gastro-entérites aigües. Les agents pathogènes identifiés sont par ordre de fréquence décroissante : des parasites – 32% (*Giardia intestinalis* > *Cryptosporidium*), des bactéries hors *Legionella* – 21% (*Shigella* > *Salmonella* > *Campylobacter* > *E. coli*), des virus – 14% (Norovirus > Hépatite A > Rotavirus). Les affections respiratoires aigües dues à *Legionella* sont présentes dans 7% des épidémies rapportées. Les proportions des agents pathogènes identifiés évoluent en fonction du temps (Figure 7). Les principales modifications concernent la période la plus récente (2001-2008) au cours de laquelle on observe une augmentation de la part des *Legionella* en raison de leur introduction dans les critères d'inclusion du système de surveillance à partir de 2001. On observe également une augmentation de la part des virus entériques, une diminution de la part des parasites et une diminution de la part des causes non déterminées.

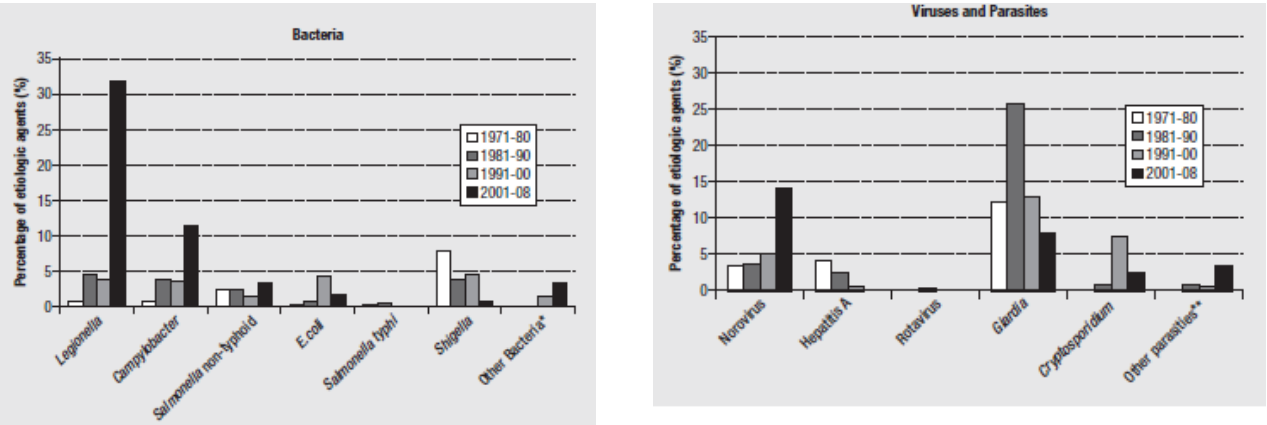


Figure 7 : Agents pathogènes identifiés dans des épidémies d'origine hydrique impliquant des réseaux d'eau publiques aux Etats-Unis entre 1971 et 2008 (source : (Craun 2012)).

La taille moyenne des épidémies entre 1971 et 2008 dans les systèmes de distribution publique est de 280 cas par épidémie (947 cas en incluant l'épidémie de Milwaukee de 1993).

L'évaluation du système de surveillance montre une variabilité dans la qualité des données au cours du temps. Ainsi, les techniques de laboratoire ont probablement permis de réduire la proportion d'étiologies inconnues. *A contrario*, le niveau de vraisemblance de l'origine hydrique établi sur la base des critères du CDC (Figure 1) a diminué entre 1989 et 2008 comparativement à la période 1971 – 1988. Cette diminution pourrait s'expliquer par une baisse des informations recueillies, en particulier celles de nature environnementale qui nécessitent des investigations de terrain de plus en plus difficiles à réaliser à la suite des réductions de moyens humains dédiés. De même, il est indiqué une implication variable des Etats dans cette surveillance, ce qui conduit probablement à une hétérogénéité dans la déclaration de ce type d'épisode entre les Etats. Par conséquent, les données de surveillance ne permettent pas d'estimer le nombre total d'épidémies hydriques et de cas liés. Enfin, la diminution de la taille des épidémies au cours du temps (83 cas en moyenne sur la période 2001-2008 versus 257 cas sur la période 1971-2000) suggère que la détection des petites épidémies a été améliorée. La mise en œuvre de ce système de surveillance nécessite une collaboration continue entre les professionnels de santé publique de l'US-EPA, du CDC et des agences des Etats.

1.2.2.2 Le Canada

Au Canada, la surveillance des épidémies d'origine hydrique n'est pas standardisée ni coordonnée à un niveau national. Elle s'appuie principalement sur le signalement par les laboratoires de l'identification d'agents pathogènes transmissibles par l'eau et sur les signalements des cas groupés de maladies par les médecins. Le circuit de déclaration de ces évènements part du niveau territorial vers le niveau national mais la remontée de l'information sur la source de contamination n'est pas obligatoire et rarement rapportée. Or pour la majorité des épidémies qui ont un lien épidémiologique avec l'eau

potable, il n'y a pas de confirmation microbiologique chez les personnes atteintes. Par conséquent, les données sur les épidémies d'origine hydrique au Canada sont partielles, difficile d'accès, et détenues à différents endroits sous des formats non standardisés. Pour autant, des données existent et plusieurs bilans ont été produits pour permettre de documenter l'impact sanitaire lié à ces évènements et d'identifier les facteurs de risque associés (Schuster 2005; Wilson 2009; Pons 2015). Il s'agit à chaque fois de données de surveillance rétrospectives.

Un bilan publié en 2005 documente les épidémies survenues au Canada sur une période de 28 ans à partir de synthèses annuelles de Santé Canada et Santé Québec, et d'une revue de la littérature (Schuster 2005). Les critères des Gallois (Figure 1) sont utilisés pour classer les épidémies en fonction de leur lien fort, probable ou possible avec l'eau potable. Au final, 288 épidémies liées à l'eau sont comptabilisées entre 1974 et 2001, avec une saisonnalité marquée (pic durant les saisons printemps/été). Les agents pathogènes identifiés sont des parasites (43%), des bactéries (41%) et des virus (16%). Par ordre de fréquence décroissant : *Giardia*, *Campylobacter*, *Cryptosporidium*, Norwalk-like viruses, *Salmonella*, Rotavirus. Les causes environnementales et les facteurs de risque sont également décrits. En revanche, les informations épidémiologiques de base ne sont pas présentes dans ce bilan en raison d'une disponibilité partielle et hétérogène des données (exemple : absence de date et lieu de survenue de l'épidémie, du nombre de personnes malades, etc.). Par conséquent, seul un tiers des épidémies (34%) bénéficient du meilleur niveau de preuve d'association avec l'eau selon les critères gallois (44% ont le plus faible). Les conclusions de cette étude pointent la nécessité de mettre en œuvre un système de surveillance national, standardisé pour améliorer la qualité des données. Elle recommande par ailleurs la mise en place de formations en épidémiologie pour améliorer la collecte d'informations lors des épidémies.

Sur la base de ces conclusions, un deuxième bilan a été réalisé dans le but d'obtenir des informations détaillées et standardisées sur les épidémies d'origine hydrique survenues entre 1993 et 2008 au Canada (Wilson 2009). Les informations ont été récoltées à l'aide d'un questionnaire standardisé rempli lors d'une enquête téléphonique auprès des professionnels en santé-environnement des autorités sanitaires de l'ensemble des provinces du Canada. Ce bilan rapporte 45 épidémies dont la majorité sont identifiées par les patients (35% des épidémies), suivi des laboratoires d'analyse médicale (33%) et des médecins (22%). La gastro-entérite aigüe est le principal symptôme rapporté. Le nombre moyen de cas par épidémie est estimé à 669 cas (médiane = 50) et plus de la moitié des épidémies ont concerné des réseaux d'eau de petite taille (moins de 1 000 personnes desservies).

Comme pour les Etats-Unis, ces deux bilans canadiens sont sujets à des biais de sous-déclaration. Deux raisons principales peuvent expliquer cette sous-déclaration : d'une part le faible taux de déclaration aux autorités locales rapporté au nombre de cas gastro-entérites aigües qui surviennent dans la communauté (1/313 en moyenne, (Majowicz 2005), d'autre part, l'absence de système de surveillance organisée au Canada. Néanmoins, ils mettent en évidence la richesse de l'information disponible à un niveau local sans qu'elle soit standardisée et facilement accessible au niveau national. Ils permettent également d'identifier les principaux facteurs de risque.

Enfin, une revue systématique récente qui s'est focalisée sur les épidémies liées aux réseaux d'eau de petite taille (<1 000 habitants desservis) survenues entre 1970 et 2014 aux Etats-Unis et au Canada a pointé du doigt la diversité des terminologies employées pour décrire les épidémies d'origine hydrique et le manque global de données sur la ressource et le traitement de l'eau (Pons 2015).

1.2.2.3 Les pays européens

En Europe, plusieurs études permettent de dresser un état des lieux des dispositifs de surveillance utilisées pour documenter les épidémies d'origine hydrique et les tendances épidémiologiques (Furtado 1998; Andersson & Bohan 2001; Poullis 2002; Smith 2006; Risebro & Hunter 2007; Beaudeau 2008; Gossner 2015; Guzman-Herrador 2015). Dans la plupart des pays cette surveillance s'appuie sur des systèmes de remontée d'informations (obligatoires ou volontaires) par les professionnels de santé concernant des maladies infectieuses diagnostiquées par un médecin (le plus souvent de type gastro-entérite aigüe) et confirmées ou non par des analyses de laboratoire.

Certains pays comme la Suède collectent des données de surveillance des maladies infectieuses incluant les maladies liées à l'eau depuis les épidémies de choléra au 19ème siècle (entre 1834 et 1874) (Andersson & Bohan 2001). Un bilan historique sur une période de 100 ans (1880-1979) a permis de répertorier 77 épidémies d'origine hydrique, principalement de type fièvre typhoïde et shigelloses, impliquant 26 867 cas de maladie et 789 décès (Andersson 1992). Depuis 1980, le système suédois s'est amélioré en utilisant un questionnaire standardisé pour l'investigation de toute suspicion d'épidémie d'origine hydrique. Un second bilan sur la période 1980-1999 rapporte 116 épidémies totalisant 57 500 malades et 2 décès. Les agents pathogènes les plus souvent identifiés étaient *Campylobacter sp.* et *Giardia lamblia.*

Un bilan récent incluant 4 pays nordiques dont la Suède porte sur la période 1998-2012 (Guzman-Herrador 2015). Avec ce dernier bilan, la Suède cumule ainsi 132 années de données de surveillance

des épidémies d'origine hydrique (1880-2011[2]) ! Dans ce dernier, 59 épidémies, affectant 52 258 personnes sont rapportées pour la Suède, dont deux épidémies majeures à cryptosporidiose : la première en 2010 dans la ville d'Östersund entraînant près de 27 000 cas, ce qui en fait à ce jour l'épidémie d'origine hydrique la plus importante identifiée en Europe ; la deuxième en 2011 touchant environ 20 000 personnes dans la ville de Skellefteå. Pour les autres pays inclus dans cette étude (Danemark, Finlande, Norvège), on dénombre respectivement 4 (660 cas), 59 (22 594 cas) et 53 épidémies (10 483 cas). La plupart des épidémies (76%) impliquent moins de 100 cas et le niveau de vraisemblance du lien avec l'eau augmente avec la taille de l'épidémie selon les critères du CDC (Figure 8).

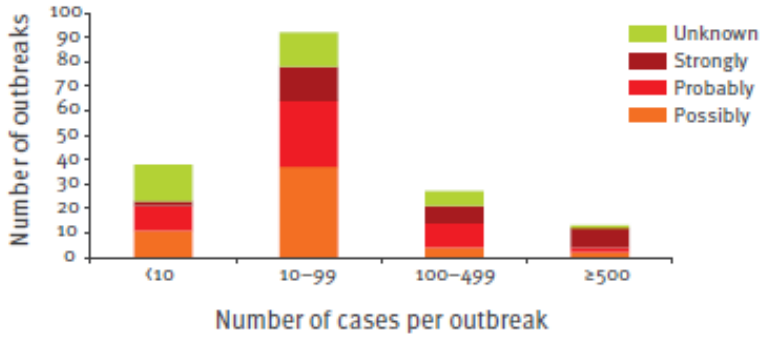


Figure 8 : Répartition de la taille des épidémies d'origine hydrique en fonction de leur niveau d'association avec l'eau dans les pays nordiques (Danemark, Finlande, Norvège, Suède), 1998-2012[3]. (Source : (Guzman-Herrador 2015)). *NB : les critères utilisés sont ceux du Communicable Disease Surveillance Center (Figure 1).*

Les agents pathogènes les plus fréquemment observés dans les épidémies rapportées sont le Norovirus et *Campylobacter*. *Cryptosporidum* est responsable du plus important nombre de cas (Tableau 1).

[2] Dans l'étude de Guzman-Herrador, l'analyse s'est arrêtée à 2011 pour la Suède

Tableau 1 : Répartition des épidémies d'origine hydrique et du nombre de cas en fonction des agents pathogènes responsables, Danemark, Finlande, Norvège, Suède - 1998-2012[3]. (Source : (Guzman-Herrador 2015)).

Year	Number of outbreaks (number of patients involved) by microorganism											
	Caliciviridae	Campylobacter	Cryptosporidium	*Escherichia coli* (pathogenic)	Giardia	Rotavirus	Salmonella	Shigella	Francisella tularensis	Multiple microorganisms	Unknown	Total
1998	2 (2,500)	2 (2,216)	–	1 (unknown)[b]	1 (3)	–	–	–	–	–	1 (13)	7 (4,732)
1999	4 (238)	2 (14)	–	–	–	–	1 (55)	–	–	–	7 (664)	14 (971)
2000	5 (5,944)	4 (1,063)	–	–	1 (37)	–	–	–	–	1 (300)	5 (167)	16 (7,511)
2001	3 (698)	4 (1,069)	–	–	–	–	1 (3)	–	–	–	2 (37)	10 (1,807)
2002	5 (746)	4 (114)	–	–	–	–	–	–	1 (11)	1 (50)	5 (520)	16 (1,441)
2003	7 (291)	1 (3)	–	1 (8)	–	1 (140)	–	–	–	–	3 (101)	13 (543)
2004	3 (259)	3 (13)	–	–	1 (6,000)	–	–	–	–	–	4 (32)	11 (6,304)
2005	1 (45)	2 (300)	–	1 (16)	–	–	–	–	1 (2)	–	5 (144)	10 (525)
2006	1 (150)	2 (45)	–	1 (10)	–	–	–	1 (18)	1 (5)	2 (35)	4 (38)	12 (283)
2007	3 (90)	3 (1,613)	1 (28)	–	1 (13)	–	–	–	3 (27)	2 (6,513)	5 (2,431)	18 (10,715)
2008	1(2,000)	2 (20)[b]	–	1 (20)	1 (2)	–	–	–	–	–	4 (110)	9 (2,152)
2009	4 (436)	2 (210)	–	1 (4)	–	–	–	–	–	–	3 (67)	10 (717)
2010	5 (401)[b]	2 (275)	2 (27,000)[b]	–	–	–	–	–	–	1 (40)	2 (30)	12 (27,746)
2011	5 (57)[b]	3 (56)	1 (20,000)	1 (8)	–	–	–	–	–	1 (27)	2 (15)	13 (20,163)
2012	2 (170)	–	–	1 (15)	–	–	–	–	–	1 (200)	–	4 (385)
Total	**51 (14,025)**	**36 (7,011)**	**4 (47,028)**	**8 (81)**	**5 (6,055)**	**1 (140)**	**2 (58)**	**1 (18)**	**6 (45)**	**9 (7,165)**	**52 (4,369)**	**175 (85,995)**

Dans les quatre pays de cette étude, un système national de déclaration dématérialisé (web-based) est à disposition des autorités municipales pour notifier les épidémies d'origine hydrique. Le signalement initial provient des médecins qui ont obligation de rapporter toute suspicion d'épidémie aux autorités municipales. L'utilisation d'un mode de signalement dématérialisé a amélioré le taux de déclaration. Néanmoins, la qualité des informations épidémiologiques, microbiologiques et environnementales est perfectible car seule une faible proportion des épidémies a pu être classée comme fortement associée à l'eau d'après les critères du CDC (23% - Figure 8).

L'Angleterre et le Pays de Galles se distinguent des autres pays par l'adoption depuis 1992 d'une définition standardisée et de critères de classification pour la surveillance des épidémies d'origine hydrique (cf. partie définition) (Furtado 1998; Smith 2006). Ils permettent ainsi d'identifier et de caractériser les épidémies d'origine hydrique parmi les épidémies signalées au département des gastro-entérites du Centre de surveillance des maladies infectieuses (CDSC) qui coordonne au niveau national la surveillance maladies infectieuses. Les signalements proviennent des acteurs de terrain[3] qui sont sensibilisés à la déclaration (pas de système obligatoire), ou des données de laboratoires lorsqu'une augmentation inhabituelle de la fréquence d'agents pathogènes est détectée à un niveau local. En cas de signalement, un questionnaire est adressé à l'investigateur principal pour renseigner un ensemble d'informations standardisées incluant des données cliniques et biologiques ; et des données sur le système d'approvisionnement en eau et les incidents éventuels.

[3] Les hygiénistes, les épidémiologistes, les agents en santé-environnement, les microbiologistes, les hôpitaux

Un bilan de 12 années de surveillance a permis de recenser 49 épidémies impliquant un réseau d'eau publique ou privé (Smith 2006). L'impact sanitaire de ces épidémies est évalué à 3 399 cas et les principaux agents pathogènes identifiés sont *Cryptosporidium* (55% des épidémies liées à des réseaux d'eau) et *Campylobacter* (24%). Un lien fort avec l'eau a pu être établi pour 20/49 épidémies (41%) selon les critères gallois (Figure 1) et seules un quart (24%) des épidémies ont le lien le plus faible.

Concernant le système de surveillance, plusieurs facteurs sont identifiés comme influençant la détection des épidémies : la sévérité des symptômes, l'accès à un médecin, la réalisation de prélèvements cliniques, la confirmation par les laboratoires. Comme dans les pays nord-américains, une sous-détection est évoquée sans qu'il soit possible d'estimer le nombre d'épidémies non détectées.

Deux études évaluent qualitativement les systèmes de surveillance des épidémies d'origine hydrique dans les pays européens (Poullis 2002; Risebro & Hunter 2007). La première constate que les épidémies hydriques sont souvent enregistrées par des systèmes de surveillance incluant d'autres véhicules possibles comme les aliments. Elle interroge sur la pertinence de mettre en place un enregistrement séparé des épidémies d'origine hydrique. La seconde évoque la nécessité de standardiser les diagnostics de laboratoire et les protocoles de prélèvements pour investiguer les maladies/épidémies, d'encourager la transmission électronique des données de surveillance au sein de l'Europe, d'inclure dans la surveillance des maladies infectieuses des données de géolocalisation des cas (par exemple code postaux). Elle évoque l'importance de la rétro-information vers les personnes qui fournissent des données de surveillance, des relations avec les partenaires et de la communication entre les acteurs (épidémiologistes, autorités sanitaires, exploitants d'eau). Enfin, l'inclusion des données telles que les plaintes groupées de consommateurs font consensus quant à leur utilité car elles représentent un signal précoce mais sont très rarement rapportées aux autorités sanitaires.

1.2.2.4 Le cas de la France

En France, Santé publique France (ex-InVS) est en charge de coordonner la surveillance des épidémies d'origine hydrique au niveau national depuis sa création en 1998. Il n'existe pas de système de collecte organisé spécifique pour la surveillance des épidémies d'origine hydrique, mais un guide méthodologique fournit un cadre national pour les investiguer et recenser les informations utiles à la description des épidémies et des circonstances de survenues (Beaudeau 2007). La surveillance repose principalement sur des systèmes de remontée d'information existants : le signalement volontaire de cas groupés de gastro-entérite aigüe par les professionnels de santé (médecins généraliste, pharmaciens au travers l'augmentation des ventes de médicaments), le système de maladies à déclaration obligatoire incluant les Tiac (surveillance passive) dont la définition permet d'englober les épidémies

d'origine hydrique par ingestion d'eau contaminée (cf. paragraphe Définitions), le contrôle sanitaire de l'eau distribuée ou le signalement de plaintes de consommateurs qui peuvent mettre en évidence une contamination microbiologique. Plus récemment, l'utilisation des données de remboursement de médicaments, permettant de surveiller les cas de gastro-entérite aigüe médicalisés, est venue compléter ce dispositif. Des travaux précédents détaillent la place des données de l'Assurance Maladie dans ce dispositif et les perspectives qu'elles offrent pour la surveillance des épidémies d'origine hydrique (Beaudeau 2012a).

A l'instar des autres pays, un bilan des épidémies d'origine hydrique survenues en France entre 1998 et 2006 a été réalisé à partir de rapports d'investigation existants et de la base de données des Tiac (Beaudeau 2008). Au total, 10 épidémies sont renseignées pour cette période, cumulant plus de 8 500 cas (les estimations imprécises pour 3 épidémies ne sont pas comptabilisées). La moitié d'entre elles ont été identifiées par une analyse non conforme dans le cadre du contrôle sanitaire de l'eau distribuée, l'autre moitié provient du signalement de cas groupés de maladies par les médecins, les particuliers ou les responsables d'établissements (absentéisme scolaire). Concernant les autres sources de données, bien qu'inclues dans la définition d'une Tiac à déclaration obligatoire, les épidémies d'origine hydrique sont exceptionnellement identifiées par ce système de surveillance (la notion d'eau est rapportée dans 1% des Tiac – (Beaudeau 2012a)). Dans ce bilan, 33 Tiac pour lesquelles une origine hydrique était suspectée ont été identifiées mais n'ont pas été retenues pour l'analyse en raison de données épidémiologiques et environnementales jugées inadéquates selon les critères américains (niveau 4 de (Craun 2010)). Parmi les épidémies étudiées, le nombre moyen de cas par épidémie était de 1 200 et les taux d'attaque se situaient entre 5% et 51%. Les agents pathogènes mis en évidence étaient des parasites (*Cryptosporidium*) dans 3 épidémies, des virus (norovirus, 3 épidémies), ou un mélange (rotavirus et *Campylobacter* - 1 épidémie ; *Cryptosporidium* et *Campylobacter* – 1 épidémie).

Le faible nombre d'épidémies rapporté dans ce bilan ainsi que leur taille importante suggère une forte sous-détection qui concerne probablement davantage les épidémies impliquant peu de cas que les épidémies importantes (1 000 cas et plus). Ceci peut s'expliquer en partie par la plus faible sensibilité des déclarants à détecter une petite augmentation de cas de gastro-entérites aigües dans la collectivité et la fréquence des analyses de contrôle de l'eau qui diminue avec la taille des réseaux d'eau (3 analyses par an en moyenne pour les réseaux desservant moins de 2 000 personnes).

Ainsi, la possibilité de détecter les petites épidémies (<100 cas) diminue en fonction du niveau d'agrégation des cas : écoles, établissements regroupant un nombre important de personnes > pharmacies > médecins.

D'un point de vue environnemental, la surveillance de la qualité de l'eau a permis d'identifier 5 épidémies entre 1998 et 2006. Ce nombre est à mettre en regard du nombre de pollutions fécales de l'eau du robinet estimé pour toute la France à 104 093 sur la période 2004-2005 (Beaudeau 2012a). Le ratio nombre d'épidémies détectées / pollutions fécales ($4{,}8.10^{-5}$) rend compte de la faible spécificité de cette surveillance.

1.2.2.5 Autres pays

D'autres pays[4] ont également produit des bilans qui s'appuient sur les critères de classification gallois (Tulchinsky 2000; Sheat & Ball 2007).

1.2.3 Bilan des systèmes de surveillance

Cette revue de la littérature illustre les différentes stratégies mises en place dans les pays développés pour surveiller et détecter les épidémies d'origine hydrique : plusieurs critères et définitions, multiples sources de données, nombre important d'acteurs impliqués. Cette hétérogénéité peut entraîner une variation dans les performances des systèmes au regard des critères définis par le CDC pour évaluer un système de surveillance[5] (Centers for Disease 2001) (Tableau 5).

Globalement, la plupart des systèmes en place sont affectés par une sous détection (manque de sensibilité) sans qu'il soit possible d'estimer le nombre d'épidémies non détectées. Cette hétérogénéité rend donc difficile la comparaison entre les pays de l'impact sanitaire lié aux épidémies d'origine hydrique. Ainsi, les différences observées sur certains indicateurs standardisés comme le nombre de cas par épidémie ou le nombre d'épidémies par an pour 100 000 habitants peuvent être à la fois la traduction d'une réalité épidémiologique, de contextes environnementaux différents, de comportements variables vis-à-vis du risque hydrique et de la consommation des soins (Tableau 2). Les différences peuvent aussi être liées aux différentes sources de données utilisées et aux stratégies de surveillance mises en place, sans qu'il soit possible de quantifier ces différences.

En revanche, tous les bilans convergent vers les mêmes facteurs de risque qui agissent de façon isolée ou cumulée : évènements pluvieux entraînant une pollution et ou inondation de la ressource, incidents d'exploitation (panne de désinfection, incident de filtration, produits de traitement non ajustés), incident de distribution (rupture de canalisation, retour d'eau usée dans le réseau d'eau potable).

---

[4] Israel, Nouvelle-Zélande

[5] simplicité, flexibilité, qualité des données, acceptabilité, sensibilité, valeur prédictive positive, représentativité, réactivité et stabilité

Tableau 2 : Bilan des données sur les épidémies d'infection d'origine hydrique dans plusieurs pays

Pays	Pop (M)	Période	Nb d'épidémies	Nb épidémies par an	Nb épidémies/ an/10^5 habitants	Nb total de cas	Nb cas par épidémie	Nb cas par an	Nb cas/an/10^5 habitants
Angleterre et Pays de Galles (Furtado 1998)	53	1992-1995	19	4,8	8,96E-03	1 638	86,2	409,5	0,8
Angleterre et Pays de Galles (Smith 2006)	53	1992-2003	49	4,1	7,70E-03	3 399	69,4	283,3	0,5
Canada (Schuster 2005)	31	1974-2001	288	10,3	3,32E-02	ND	ND	ND	ND
Canada (Wilson 2009)	31	1993-2008	44	2,8	8,87E-03	29 430	668,9	1 839,4	5,9
Danemark (Guzman-Herrador 2015)	5,4	1998-2012	4	0,3	4,94E-03	660	165,0	44,0	0,8
Etats-Unis (Craun 2012)	247	1971-2008	605	15,9	6,45E-03	572 767	946,7	15 072,8	6,1
Etats-Unis[1] (Craun 2012)	247	1971-2008	604	15,9	6,45E-03	169 767	280,6	4 467,6	1,8
Finlande (Guzman-Herrador 2015)	5,4	1998-2012	59	3,9	7,28E-02	22 594	382,9	1 506,3	27,9
France (Beaudeau 2008)	63,6	1998-2006	10	1,1	1,75E-03	8 400[µ]	1 200[µ]	1 061,1	1,7
Israel (Tulcinsky 2000)	6	1976-1997	130	5,9	9,85E-02	23 787	183,0	1 081,2	18,0
Norvège (Guzman-Herrador 2015)	5	1998-2012	53	3,5	7,07E-02	10 483	197,8	698,9	14,0
Nouvelle Zélande (Sheat & Ball 2007)	4	2001-2005	84	16,8	4,20E-01	724	8,6	144,8	3,6
Suède (Guzman-Herrador 2015)	9,6	1998-2011	59	4,2	4,39E-02	52 258	885,7	3 732,7	38,9
Suède[2] (Guzman-Herrador 2015)	9,6	1998-2011	57	4,1	4,27E-02	23 258	408,0	1661,3	17,3
Suède (Andersson & Bohan 2001)	8,6	1980-1999	116	5,8	6,74E-02	57 500	495,7	2 875,0	33,4
Suède (Andersson 1992)	7	1880-1979	77	0,8	1,10E-02	26 867	348,9	268,7	3,8

[1] sans Milwaukee
[2] sans Ostersund et Skelleftea
ND = non disponible
[µ] pour 7 épidémies

1.3 L'apport de la surveillance syndromique

Ces dernières années, l'utilisation des données de surveillance syndromique incluant les données de consultations aux urgences, les services d'assistance téléphonique, les données de vente de médicaments et l'absentéisme scolaire a montré son utilité pour l'étude rétrospective des épidémies d'origine hydrique (Berger 2006).

1.3.1 Définition et types de données utilisées

La surveillance syndromique est un outil utilisé pour la détection des épidémies aux Etats-Unis depuis le milieu des années 1990 (Heffernan 2004). Le CDC la définit comme *« une approche, dans laquelle les intervenants sont assistés par des procédures d'enregistrement automatiques des données, qui permettent la mise à disposition de données pour le suivi et l'analyse épidémiologique en temps réel ou proche du temps réel. Cela afin de détecter des événements habituels ou inhabituels plus tôt qu'il n'aurait été possible de le faire sur la base des méthodes traditionnelles de surveillance »* (Henning 2004). Son intérêt a augmenté depuis le début du 21ème siècle dans un contexte de risque bio-terroriste

accru. Aujourd'hui, de nombreux pays développés ont adopté ce type de surveillance en complément des systèmes de surveillance traditionnels. De nombreuses sources de données potentielles existent. A titre indicatif, les données les plus utilisées proviennent des services d'urgences (France, Australie, Taiwan, Etats-Unis) (Balter 2005; Muscatello 2005; Wu 2008; Caserio-Schonemann & Meynard 2015), des ventes de médicaments en pharmacies (Canada, Etats-Unis) (Das 2005; Edge 2006), des appels téléphoniques (Royaume-Unis) (Cooper 2002). D'autres sources comme l'absentéisme scolaire ou professionnel sont également citées même si dans ce dernier cas, il n'est pas précisé le mode de remontée des données.

Le terme de surveillance syndromique reste cependant assez large et diversement interprété car pour certains, il s'arrête à la notion de syndrome, pour d'autre il intègre la notion de remontée automatique des données. Ceci le distingue de la surveillance traditionnelle généralement associée à la notion de système déclaratif (obligatoire ou non) par les professionnels de santé.

De par sa capacité à identifier des symptômes plus ou moins spécifiques de la gastro-entérite aigüe selon les sources utilisées (données précliniques[6], cliniques pré-diagnostic), ce type de surveillance peut en théorie apporter une plus-value pour la détection des épidémies d'origine fécale liées à l'eau.

1.3.2 Les données de l'Assurance Maladie : une source de données pertinente pour la détection des épidémies de gastro-entérite aigüe d'origine hydrique en France

En France, la surveillance syndromique des gastro-entérites aigües médicalisées est utilisée depuis plusieurs années pour prévenir le risque infectieux lié à la consommation d'eau du robinet quelle que soit sa nature (endémique, hyper-endémique ou épidémique) (Beaudeau 2012a; Beaudeau 2012b). Les cas de gastro-entérite aigüe médicalisés proviennent du Système national d'information inter-régimes de l'Assurance Maladie (Sniir-AM). Il contient des données individuelles sur les patients (prescriptions médicales, prescripteurs, bénéficiaires). Son accès est autorisé à un nombre restreint d'organisme dont Santé publique France (ex-InVS)[7]. Cette base couvre près de 99% de la population résidant en France quel que soit le régime de l'Assurance Maladie (Tuppin 2010).

L'identification d'un cas de gastro-entérite aigüe à partir du Sniir-AM est réalisée grâce à un algorithme de sélection spécifique (Bounoure 2011). Le critère d'extraction des prescriptions médicales est la présence d'au moins un médicament cible utilisé pour traiter la gastro-entérite aigüe (70 spécialités

[6] Ventes de médicaments au comptoir de la pharmacie, absentéisme scolaire, etc.

[7] Arrêté du 19 juillet 2013 relatif à la mise en œuvre du Système national d'information inter-régimes de l'assurance maladie

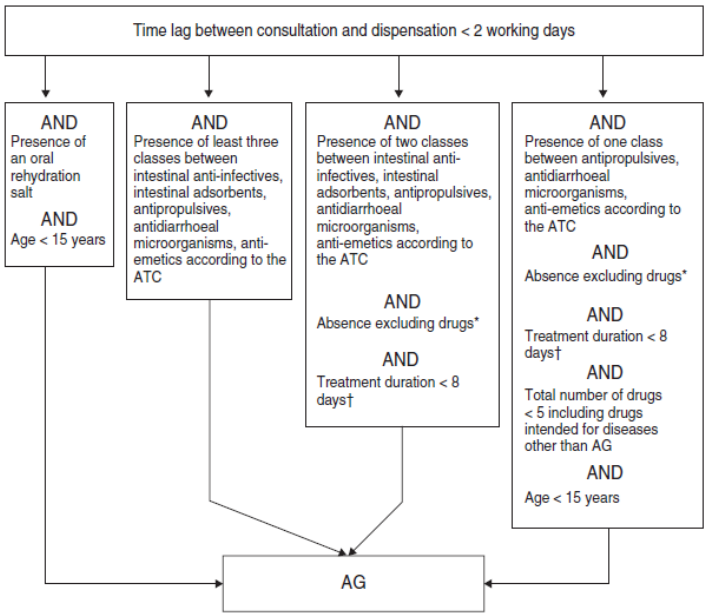
remboursées en 2012 par l'Assurance Maladie) (Tableau 3). Les principaux critères de sélection des cas de gastro-entérite aigüe sont le délai entre la prescription et la délivrance, l'absence de médicaments spécifiques de pathologies chroniques pouvant entraîner des épisodes de gastro-entérites aigüe, la présence d'associations de médicaments spécifiques du traitement de la gastro-entérite aigüe et l'âge (Figure 9).

L'algorithme se caractérise par une sensibilité et une spécificité proches de 89%. Pour chaque cas de gastro-entérite aigüe sélectionné par l'algorithme, les informations suivantes sont disponibles : l'âge, le sexe, la liste des médicaments remboursés, la date de consultation chez le médecin, la date de délivrance à la pharmacie, la commune de résidence, la commune du médecin prescripteur, la commune de la pharmacie.

Tableau 3 : Médicaments utilisés pour le traitement de la gastro-entérite aigüe (source : (Bounoure 2011))

Therapeutic classes	ATC	Trademark	Drug
Intestinal antispamodics*	A03A	Spasfon®	Phloroglucinol
		Duspatalin®	Mebeverine
		Dicetel®	Pinaverium
		Debridat®	Trimebutine
		Meteospamyl®	Alverine
		Meteoxane®	Simeticone
Anti-emetics	A04A	Vogalene®	Metopimazine
	A03F	Motilium®	Domperidone
		Peridys®	Metoclopramide
		Primperan®	
Probiotic antidiarrhoeals	A07F	Ultralevure®	*Saccharomyces boulardii*
		Lacteol®	*S. cerevisiae*
Intestinal antipropulsive	A07D	Imodium®	Loperamide
		Arestal®	
	A07X	Tiorfan®	Racecadotril
Intestinal absorbents	A07B	Carbolevure®	Activated charcoal + *S. cerevisiae*
	No ATC code	Carbosylane®	Activated charcoal
	A02X	Bedelix®	Montmorillonite
		Smecta®	Diosmectite
Intestinal anti-infectious agents	A07A	Ercefuryl®	Nifuroxazide
		Lumifurex®	
		Panfurex®	
Oral rehydration salts	Medical devices	Adiaril® Alhydrate®	
		Fanolyte®	
		Ges 45®	
		Hydrogoz®	
		Picolite®	
		Viatol®	

ATC, Anatomical Therapeutic Classification.
* Not used for the extraction of refund data from NHI database.



*médicaments éliminatoires : « anti-acides/anti-régurgitants, « antibiotiques par voie systémique » et tout médicament entraînant une diarrhée (effets secondaires). † En tenant compte de tous les médicaments présents ; estimé à la fois à partir du contenu (nombre de prises par boîte) et du nombre de boîtes délivrées. ATC : Anatomical Therapeutic Classification

Figure 9 : Algorithme utilisé pour la sélection des cas de gastro-entérite aigüe (AG), basé sur les données de médicaments remboursés (source : (Bounoure 2011)).

La sensibilité de cette source de données pour la détection d'épidémies de gastro-entérite aigüe d'origine hydrique est principalement liée au taux de consultation, estimé à près de 30% pour cette pathologie (Van Cauteren 2012). Parmi les cas qui vont consulter, près de 90% vont voir un médecin généraliste et 1% vont aux urgences. Les habitudes peuvent varier en fonction de l'agent pathogène (Wheeler 1999).

Depuis 2009, les données du Sniir-AM sont utilisées pour décrire *a posteriori* les épidémies de gastro-entérite aigües d'origine hydrique à partir de situations préalablement déclarées et investiguées (Rambaud 2011; Beaudeau 2012a). Elles peuvent également servir à améliorer la surveillance des épidémies d'origine hydrique par la mise en place d'un système de détection automatisé dont l'objectif serait d'identifier des épidémies de gastro-entérites aigües localisées, non détectées par un autre système de surveillance, et dont l'origine hydrique est très probable.

1.4 Synthèse des principales caractéristiques des épidémies d'origine hydrique

Le bilan précédent montre que la probabilité de pouvoir attribuer une épidémie de gastro-entérite à la consommation d'eau augmente avec l'identification de facteurs environnementaux et/ou de données relatives à la distribution de l'eau et l'établissement d'un lien avec la survenue et la répartition des

malades. Cette probabilité est très dépendante des données disponibles et de leur temporalité par rapport à l'évènement.

Il est ainsi possible d'identifier un ensemble de caractéristiques communes aux épidémies d'origine hydrique et des sources de données permettant de les atteindre.

Huit caractéristiques, inspirées des travaux de Poullis *et al.* (Poullis 2005), peuvent être identifiées sur la base du bilan des données de surveillance des épidémies impliquant des gastro-entérites aigües : d'un point de vue environnemental, les épidémies d'origine hydrique sont le plus souvent associées à un incident de traitement de l'eau, de distribution de l'eau et/ou à un évènement de pollution au niveau du bassin versant (caractéristique 1). Ces incidents/évènements peuvent être détectés par une surveillance au niveau de la station de traitement d'eau ou de la qualité de l'eau distribuée. D'un point de vue épidémiologique, la plupart des épidémies d'origine hydrique sont associées à une apparition soudaine et étendue de malades (caractéristique 2) ; une augmentation rapide du nombre de cas avec des symptômes similaires (caractéristique 3) ; une concentration de cas sur une zone desservie par un même réseau d'eau et *a contrario* un plus faible nombre de cas sur une zone desservie par un autre réseau d'eau adjacent (caractéristique 4). Ces caractéristiques peuvent être détectées par des données de surveillance syndromique sous réserve qu'elles contiennent des informations de géolocalisation. La concordance entre les lieux de résidence des cas et le tracé du réseau de distribution d'eau constitue un également argument fort[8] (caractéristique 5). Cette concordance nécessite d'avoir accès à des données de modélisation des réseaux d'eau et des informations précises pour géolocaliser les cas à l'adresse. Une étude épidémiologique descriptive peut également permettre de caractériser une épidémie d'origine hydrique[9] et d'exclure d'autres sources possibles (caractéristique 6). Des investigations de terrain sont le plus souvent nécessaires pour atteindre cette caractéristique. D'un point de vue microbiologique, l'identification de l'agent étiologique dans l'eau par un laboratoire d'analyse caractérise également les épidémies d'origine hydrique (caractéristique 7). Enfin, la mise en évidence d'une association significative entre la survenue de la maladie et la consommation d'eau dans une étude épidémiologique analytique permet également de conforter une épidémie d'origine hydrique (caractéristique 8). L'atteinte de cette caractéristique nécessite la mise en place d'études en population, type cohorte ou cas-témoin. Les caractéristiques et données nécessaires pour détecter et décrire une

[8] Les personnes résidant à proximité du point d'introduction de la pollution sont les plus précocement touchés et ceux habitant les plus loin en dernier

[9] Le profil d'une épidémie d'origine hydrique est influencé par plusieurs paramètres : nature du ou des agents pathogène(s) en cause, durée de l'exposition, habitudes de consommation d'eau dans la population. Leur impact sanitaire est généralement important (entre 100 et 500 cas, fort taux d'attaque) et localisé. Toutes les classes d'âge sont touchées

épidémie d'origine hydrique sont accessibles à des temps différents par rapport à la date de contamination du réseau d'eau (Figure 10).

Prises de façon isolée, ces caractéristiques sont peu spécifiques. En revanche, leur cumul permet d'augmenter la probabilité d'un lien de causalité entre une épidémie détectée et un réseau d'eau contaminé en référence aux critères de classification des épidémies d'origine hydrique évoqués en début de chapitre (Figure 1 et Figure 2) (Poullis 2005). La prise en compte de tout ou partie de ces caractéristiques représente donc un enjeu important dans la construction d'un système de surveillance dédié à ces évènements.

Ainsi, la présence d'indicateurs de contamination de l'eau (caractéristique 1), de cas de gastro-entérite aigüe localisés dans une zone desservie par un même réseau d'eau (et peu de cas dans les zones desservies par des réseaux d'eau adjacents) (caractéristique 4) et d'une localisation des cas compatible avec le tracé du réseau d'eau (caractéristique 5) ; associés à l'apparition soudaine et étendue de malades (caractéristique 2) ou à une augmentation rapide de cas avec des symptômes similaires (caractéristique 3) est une bonne indication pour suspecter/détecter une épidémie d'origine hydrique. Par comparaison, une épidémie d'origine alimentaire peut également entraîner une apparition soudaine et étendue de malades (caractéristique 2), une augmentation rapide de cas avec des symptômes similaires (caractéristique 3) et/ou des cas de gastro-entérite aigüe localisés dans une zone desservie par un même réseau d'eau (et peu de cas dans les zones desservies par des réseaux d'eau adjacents) (caractéristique 4). En revanche, elle ne satisfera pas les caractéristiques 1 ou 5 impliquant un lien avec l'eau (indicateur de contamination de l'eau – caractéristique 1 ou localisation des cas sur le tracé du réseau d'eau – caractéristique 5). Les caractéristiques 6 à 8 (respectivement association suggérée par l'épidémiologie descriptive, identification de l'agent étiologique dans l'eau et mesure de l'association par l'épidémiologie analytique) viennent compléter et affiner la description du lien entre une épidémie détectée et l'origine hydrique. Leurs délais de mise en œuvre sont rarement compatibles avec l'objectif de détection.

En l'absence d'informations environnementales, les caractéristiques 2 à 4 (respectivement apparition soudaine de malades, augmentation rapide de cas avec des symptômes similaires et cas localisés dans une zone desservie par un même réseau d'eau), accessibles à partir de données de surveillance syndromique, permettent de détecter une épidémie possiblement liée à l'eau. La recherche d'indicateurs de contamination de l'eau (caractéristique 1) ou d'une autre source possible (caractéristique 6) permettront de conforter l'origine hydrique de l'épidémie.

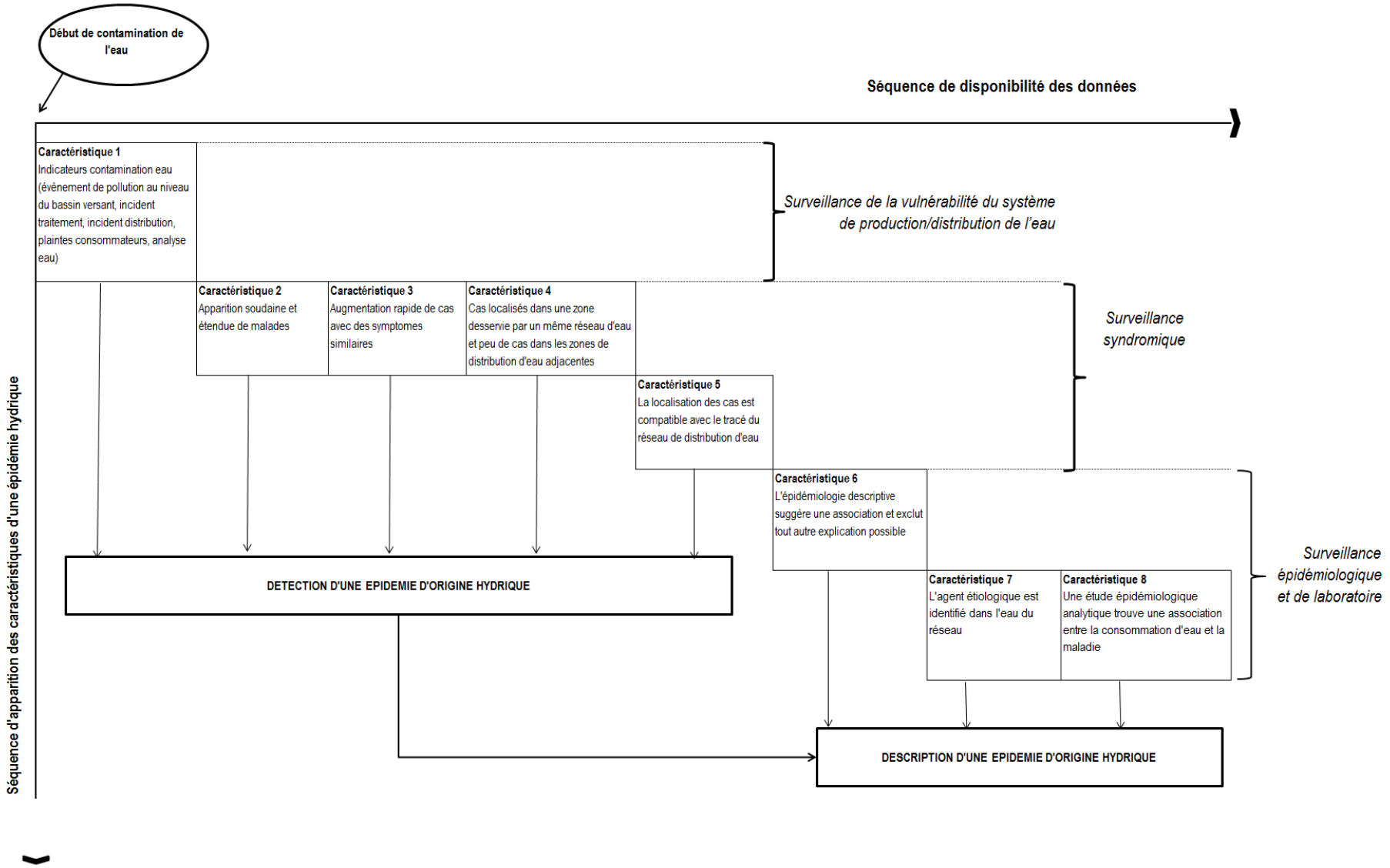


Figure 10 : Caractéristiques et délais d'obtention pour la détection et description d'une épidémie d'origine hydrique (inspiré de (Poullis 2005)).

1.5 Conclusions et perspectives

Les parties précédentes ont permis de montrer :

- Le manque de consensus sur les définitions et critères de classification des épidémies d'origine hydrique ;
- L'hétérogénéité des stratégies mises en place dans les pays pour la surveillance du risque infectieux porté par l'eau du robinet ;
- La convergence sur les facteurs de risque identifiés ;
- La diversité des sources de données existantes et utilisées pour la surveillance des épidémies d'origine hydrique ;
- Une sous-détection des épidémies d'origine hydrique, difficile à quantifier, quelle que soit la stratégie de surveillance mise en place ;
- L'existence de caractéristiques communes aux épidémies d'origine hydrique ;
- La possibilité d'identifier les cas de gastro-entérite aigüe médicalisés en France par jour et par commune à partir des données de l'Assurance Maladie ;
- L'intérêt que représentent les données de l'Assurance Maladie pour la surveillance des épidémies de gastro-entérite liées à l'eau du robinet.

Dans le contexte du risque épidémique, les données de l'Assurance Maladie pourraient répondre à deux objectifs : i) la description d'épidémies de gastro-entérite d'origine hydrique connue ; et ii) la détection automatisée plus systématique des épidémies de gastro-entérites aigües pouvant être liées à l'ingestion d'eau du robinet.

Le Tableau 4 applique les critères d'évaluation d'un système de surveillance à la problématique de la surveillance des épidémies d'origine hydrique et met en perspective l'apport prévisible de l'utilisation des données de l'Assurance Maladie.

Tableau 4 : Critères d'évaluation d'un système de surveillance (German 2001)

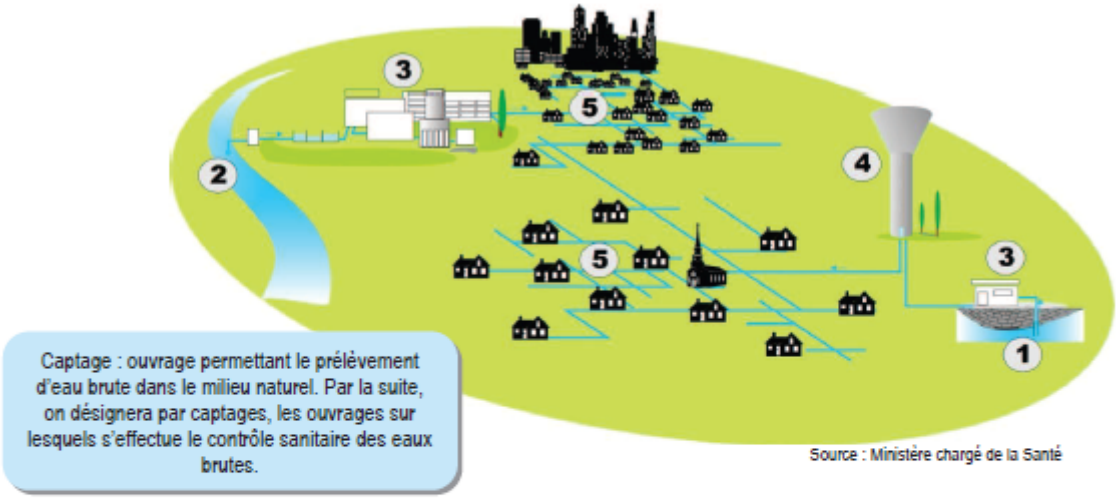
Critères	Définition	Paramètre à prendre en compte pour la surveillance des épidémies d'origine hydrique	Impact prévisible de l'utilisation des données du Sniir-AM pour cette surveillance
Simplicité	La simplicité se réfère à la structure du système de surveillance et à la facilité de sa mise en œuvre. Un système est d'autant plus simple que le nombre d'acteurs impliqués est faible, les évènements à rapporter (définition de cas par exemple) sont clairs, simples et compréhensibles, la transmission des données rapide et standardisée	Peu d'effets sanitaires sont spécifiques d'une pollution d'eau. Peu d'indicateurs de contamination d'eau sont prédictifs d'un risque épidémique. Nombreux acteurs et disciplines impliqués. Peu de standardisation dans la remontée des données.	Peut améliorer la standardisation de la remontée des données. Nécessite une définition de cas spécifique de l'indicateur sanitaire utilisé pour la surveillance. L'automatisation est possible.
Flexibilité/ **Souplesse**	La flexibilité est liée à la capacité d'adaptation du système à des modifications dans les évènements à rapporter et/ou les modalités de transmission. La facilité d'ajout de nouvelles sources de données permet aussi d'apprécier la flexibilité.	La multiplicité des évènements et des sources de données utilisés pour la surveillance des épidémies d'origine hydrique.	Augmentation du nombre d''acteurs (par exemple l'assurance maladie), des contraintes éthiques (confidentialité des données) et des contraintes techniques/informatiques. Possibilité d'ajouter facilement des indicateurs sanitaires si une définition de cas existe.
Qualité des données	Ce critère se réfère à la complétude et la validité des informations transmises.	En l'absence de format standardisé pour la remontée d'information, la qualité des données sera variable et dépendante des acteurs impliqués.	La standardisation de la remontée des données peut améliorer leur qualité. Possibilité d'automatiser le processus de contrôle de la qualité des données.
Acceptabilité	L'acceptabilité dépend de la volonté des acteurs et des organisations à participer au système de surveillance.	Le caractère déclaratif du système de surveillance peu affecter le taux de participation.	Bonne acceptabilité car transparent pour les déclarants (aucun acte déclaratif nécessaire). Adhésion nécessaire des autorités sanitaires pour investiguer les signaux (enquête environnementale)
Sensibilité	La sensibilité est la capacité d'un système à bien détecter les évènements qu'il surveille (notion d'exhaustivité). Elle dépend i) des habitudes de recours aux soins des malades, et ii) de la capacité du système à identifier, transmettre et traiter les informations.	Le premier point est dépendant de la maladie, de facteurs socio-économiques, de facteurs individuels, de l'accès aux soins. Le deuxième point est variable et lié à la simplicité du système.	Sur le premier point, les gastro-entérites aigües médicalisées représentent environ 33% des personnes atteintes en France (Van Cauteren 2012). L'exhaustivité de remontée d'information peut être impactée par un incident technique/informatique. Le risque épidémique pourrait être détecté pour 80% de la population française (Beaudeau 2012a).
Valeur prédictive positive	La valeur prédictive positive (VPP) se définit comme la probabilité qu'un signal détecté corresponde à l'évènement sous surveillance.	Dans le cas des épidémies d'origine hydrique, la prise en compte de l'exposition à l'eau permet d'améliorer la VPP.	Nécessite des études spécifiques pour évaluer la VPP.
Représentativité	La représentativité définit la capacité d'un système à représenter fidèlement un phénomène de santé dans le temps et sa distribution dans la population et l'espace.	Les modalités de recueil des informations (cas médicalisés) peut biaiser la représentativité.	Le système de surveillance est représentatif des cas médicalisés. Des hypothèses peuvent permettre d'estimer l'impact dans la population générale. Bonne couverture de l'assurance maladie
Réactivité	La réactivité représente le délai de transmission entre la survenue de l'évènement et les différents échelons.	La nécessité de collecter des informations sanitaires et environnementales peut augmenter les délais.	Les délais de transmission sont variables en fonction des sources (2 mois pour l'assurance maladie).
Stabilité	La stabilité peut se définir comme la fiabilité des acteurs et des outils et la disponibilité lorsque cela est nécessaire.	La surveillance des épidémies d'origine hydrique qui nécessite une approche multi-disciplinaire implique un nombre important d'acteurs.	L'utilisation des bases de données médico-administratives permet de réduire les interlocuteurs et améliorer la stabilité. L'évolution des stratégies économiques (déremboursements de médicaments utilisés pour identifier un cas de GEA dans le Sniir-AM) peut fragiliser la stabilité

GEA : gastro-entérite aigüe

2 Données d'exposition disponibles pour la prise en compte de l'origine hydrique dans la détection d'épidémies hydriques

2.1 La distribution de l'eau potable en France

En France, plus de 99% de la population est alimentée par un réseau public d'eau potable (Ministère chargé de la santé 2014). D'un point de vue schématique, un système d'adduction en eau potable comprend : un captage d'eau puisant dans une ressource souterraine ou superficielle, une station de traitement d'eau, des systèmes de stockage (réservoirs, châteaux d'eau), et un réseau de distribution d'eau, appelé unité de distribution d'eau (Figure 11). On dénombre en France près de 33 500 captages, 16 300 stations de traitement et 25 300 unités de distribution d'eau (Ministère chargé de la santé 2014). Le nombre d'unités de distribution d'eau varie en fonction des départements de 4 (Paris) à 850 (Isère) (Figure 12). Chaque habitation alimentée par le réseau d'eau public est associée à une unité de distribution.



① Captage d'eau dans une nappe souterraine
② Captage d'eau dans un e ressource superficielle
③ Station de traitement d'eau : selon la qualité de l'eau prélevée, la production d'eau potable peut nécessiter différentes étapes de traitement faisant appel à plusieurs types de procédés
④ Installation de stockage (réservoirs, châteaux d'eau)
⑤ Unité de distribution (UDI) : réseau d'adduction d'eau exploité par la même personne morale, appartenant à la même entité administrative, syndicat ou commune, et où la qualité d'eau est homogène

Figure 11 : Exemple d'organisation d'un système d'adduction en eau potable (source : (Ministère chargé de la santé))

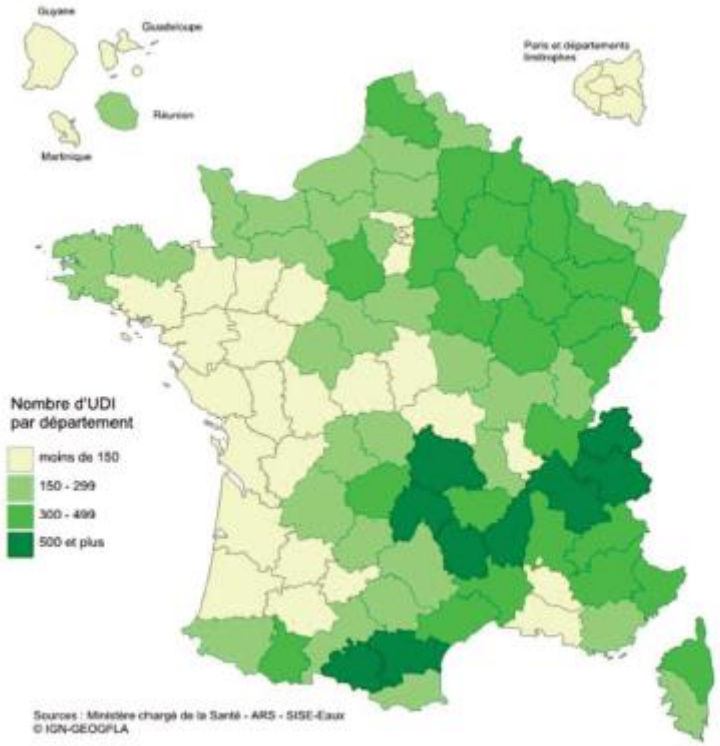


Figure 12 : Nombre d'unités de distribution d'eau potable par département – situation 2012 (source : (Ministère chargé de la santé 2014)).

Une unité de distribution désigne un ensemble de canalisations de distribution de l'eau potable au sein duquel la qualité de l'eau délivrée est considérée comme homogène, hors pollution accidentelle ou dégradation du réseau. La population alimentée par une même unité de distribution est donc semblable du point de vue de l'exposition potentielle aux agents pathogènes véhiculés par l'eau distribuée.

2.2 La base nationale SISE-eaux

Les informations concernant la description des systèmes d'adduction en eau potable et le suivi de la qualité de l'eau sont centralisées depuis 1995 par le Ministère de la Santé dans la base nationale du Système d'Information en Santé-Environnement sur les Eaux d'alimentation (SISE-Eaux). Cette base comporte des données structurelles sur les captages d'eau, sur les installations de traitement et de stockage, sur les unités de distribution et sur le lien entre les unités de distribution et les communes desservies. Elle regroupe par ailleurs l'ensemble des résultats des analyses du contrôle sanitaire de l'eau destinée à la consommation humaine disponibles depuis plus de 20 ans.

Cette base de données est le seul outil national pour estimer le niveau d'exposition des populations aux paramètres chimiques et microbiologiques véhiculés par l'eau destinée à la consommation humaine en fonction du réseau d'eau qui les alimentent. La fréquence d'échantillonnage diffère cependant beaucoup selon la taille du réseau d'eau, variant de 2 à 4 contrôles par an pour les paramètres

microbiologiques (réseaux d'eau de moins de 50 habitants) à plus de 800 contrôles annuels (réseaux d'eau de plus de 625 000 habitants)[10].

2.3 Le choix de l'unité d'exposition à l'eau dans la perspective de détection des épidémies hydriques

L'unité de distribution constitue ainsi un indicateur d'exposition environnemental d'intérêt pour la détection des épidémies d'origine hydrique. Cette entité n'a pas toujours de lien avec le découpage administratif des communes (Figure 13). Ses contours dépendent davantage du relief, des positions des points de captage ainsi que de l'implantation de la population et des activités nécessitant l'utilisation d'eau.

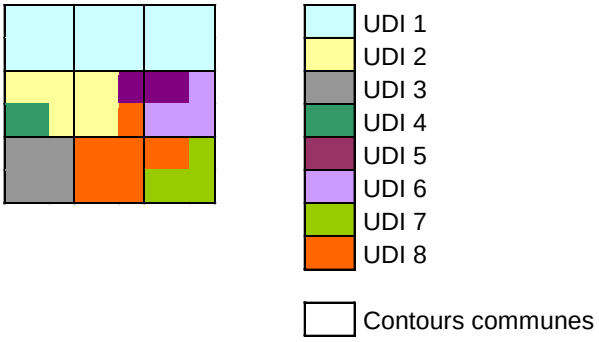


Figure 13 : Exemples de recoupements possibles entre les communes et les unités de distribution (UDI) qui les alimentent.

Pouvoir établir le lien entre des épidémies de gastro-entérite aigüe détectées et l'origine hydrique est un des enjeux majeur du système de détection à mettre en place. Pour évaluer la vraisemblance de l'hypothèse hydrique, deux approches peuvent être envisagées :

- détecter des agrégats de cas gastro-entérite aigüe à l'échelle de la commune et tester *a posteriori* le lien avec l'eau ;
- intégrer le lien avec l'eau en amont de la détection d'agrégats en tenant compte du contour des unités de distribution d'eau pour faire émerger des signaux cohérents avec une origine hydrique.

[10] Arrêté du 11 janvier 2007 relatif au programme de prélèvements et d'analyses du contrôle sanitaire pour les eaux fournies par un réseau de distribution, pris en application des articles R. 1321-10, R. 1321-15 et R. 1321-16 du code de la santé publique. NOR : SANP0720202A

Ces deux approches se traduisent notamment par des zones cibles testées distinctes : la commune (niveau d'agrégation de l'indicateur sanitaire) ou l'unité de distribution d'eau (niveau d'agrégation de l'indicateur d'exposition) (Tableau 5).

Tableau 5 : Présentation de deux approches possibles pour l'analyse de détection spatio-temporelle

	Approche « par commune »	Approche « par UDI »
Principe	Les données de l'assurance maladie étant agrégées à l'échelle de la commune, l'approche la plus naturelle consisterait donc à appliquer une méthode de détection d'agrégats spatio-temporels sur les données à la commune puis à évaluer *a posteriori* la pertinence de l'hypothèse hydrique pour l'épidémie constatée.	Le facteur d'exposition (consommation d'eau du robinet), étant agrégé à l'échelle de l'UDI, une approche alternative pourrait être d'appliquer une méthode de détection d'agrégats spatio-temporels sur les cas agrégés à l'UDI.
Prise en compte de l'aspect hydrique	L'hypothèse hydrique voit sa vraisemblance évaluée en aval de la détection de cas groupés. Il faut pour cela estimer si les agrégats identifiés sont compatibles avec l'emprise géographique d'une UDI.	L'hypothèse hydrique est maximisée en amont de la détection. En effet, le fait de considérer directement l'ensemble des communes alimentées par une même UDI permet d'interpréter chaque agrégat constaté comme pouvant être d'origine hydrique.
Mise en forme des données / Implémentation	La mise en forme des données ne demande pas de travail particulier, les cas étant déjà répertoriés au niveau communal.	Les données sanitaires sont disponibles à la commune, il y a donc un travail de mise en forme à réaliser pour tenir compte de l'adéquation entre les UDI et les communes.

UDI : unité de distribution d'eau

Le point de vue centré sur les unités de distribution pour appliquer les méthodes de détection d'agrégats paraît pertinent au regard de notre problématique, cependant il demande la construction de plusieurs objets et le choix de plusieurs paramètres pour déterminer le regroupement de communes le plus pertinent pour chaque unité de distribution.

3 Méthodes statistiques adaptées à la détection de cas groupés

3.1 Nature des données, facteurs influant et critères de choix

Le recours à une méthode statistique est fortement conditionné par la nature des données étudiées et les objectifs visés. Ce paragraphe traite des spécificités des données de l'Assurance Maladie, des facteurs à prendre en compte et *in fine* propose des critères qu'une méthode statistique devra remplir dans la perspective de son utilisation pour la détection d'agrégats assimilables à des épidémies de gastro-entérites aigües d'origine hydrique.

3.1.1 Caractéristiques des données de l'Assurance Maladie

3.1.1.1 Aspects temporels

Le pas de temps journalier des données de l'Assurance Maladie permet de définir des périodes de temps très précises. D'après la littérature, la plupart des recherches menées sur la détection d'épidémies s'appuient sur des données journalières (Dematteï 2006; Buckeridge 2007; Gaudart 2007; Unkel 2011; Pelat 2012)

Les données étant consolidées après un délai d'environ 2 mois, l'identification des épidémies est effectuée de façon rétrospective, ce qui constitue une des principales limites de cette source de données. Sur ce point, la démarche envisagée diverge de celle de la majorité des études menées depuis le début des années 2000 à partir de données de surveillance syndromique pour prévenir le risque bio-terroriste et les risques émergents (grippe aviaire, …). Le fait de disposer des données avant et après une valeur critique dans le cadre d'une détection rétrospective (pic de l'épidémie) présente certains avantages et inconvénients. Les avantages sont la possibilité d'envisager des méthodes de détection avec un degré de paramétrage supplémentaire (inclusion possible de co-facteurs tel que l'effet jour de la semaine) ; et la capacité à identifier des épidémies répétées sur une même zone de distribution d'eau, ce qui constitue un argument intéressant pour conforter l'origine hydrique. Un des inconvénients est lié au fait que le profil d'une épidémie d'origine hydrique se caractérise généralement par un fort accroissement du nombre de cas jusqu'à l'atteinte d'un pic. En revanche, la fin d'une épidémie peut présenter différentes allures en termes de décroissance avec la présence possible de pics secondaires et des cinétiques variables selon notamment la source de l'épidémie et les comportements des individus.

3.1.1.2 Aspects géographiques

Les données sont disponibles à l'échelle communale. L'ensemble de l'information pourra être ainsi portée par un point représentant la commune, par exemple son centroïde de la figure géométrique que constitue le contour de la commune. Ce point peut servir dans un calcul de matrice des distances par exemple. Néanmoins, la position des habitants dans un village peut être assez éloignée du centre de gravité, notamment dans le cadre d'une commune répartie sur plusieurs hameaux.

3.1.2 Facteurs influant sur la détection

3.1.2.1 Facteurs comportementaux

L'identification d'agrégats spatiaux ou spatio-temporels pose l'hypothèse d'une uniformité de comportement des individus. Des disparités peuvent néanmoins exister vis-à-vis de l'exposition (plusieurs facteurs peuvent influencer la consommation de l'eau du robinet, (Beaudeau 2003) ou du recours aux soins. Les individus malades, à symptômes identiques, ne vont pas tous consulter leur médecin avec la même probabilité. L'âge, la sévérité et la durée de la maladie, la facilité d'accès aux structures de soins influencent le taux de consultation (Van Cauteren 2012). Il est également possible que des populations desservies par un réseau d'eau présentant des pollutions chroniques achètent préventivement des médicaments contre la gastro-entérite aigüe ou développent une immunité vis-à-vis de certains agents pathogènes. Dans cette hypothèse, le taux de médicalisation pourrait être plus faible. La possibilité d'inclure des variables permettant d'ajuster l'analyse sur certains facteurs comportementaux peut être considérée.

3.1.2.2 Facteurs géographiques

Les communes présentent certaines disparités qui peuvent faire varier le signal à détecter : effectif et densité de population, taux de résidences secondaires, structures de soins, activités professionnelles. Ces différences peuvent par exemple jouer sur l'ampleur d'une épidémie (nombre de cas médicalisés) et son emprise géographique (des personnes résidentes de communes adjacentes peuvent être atteintes si elles exercent une activité professionnelle sur la commune avec le réseau d'eau impacté).

3.1.2.3 Facteurs temporels

La gastro-entérite aigüe peut résulter classiquement d'une contamination de personne à personne, d'une intoxication alimentaire ou d'une contamination environnementale. Il s'agit d'une pathologie dont l'incidence varie fortement selon la saison, avec une épidémie hivernale d'environ 2 à 3 mois entre décembre et avril avec un pic qui se situe généralement en janvier/février. Cette épidémie hivernale se cumule au « bruit de fond » que constitue le nombre de cas observés tout au long de l'année. En cas d'épidémie d'origine hydrique, la variation d'incidence peut être potentiellement plus difficile à détecter en période hivernale que le reste de l'année. La prise en compte des variations saisonnières est un point prépondérant dans le choix d'une méthode de détection.

Un autre facteur important lié à la nature des données de l'Assurance Maladie est l'effet jour de la semaine : en effet on distingue de fortes variations entre les jours ouvrés et les jours chômés (week end

ou jours fériés) durant lesquels l'offre de soins est réduite (Tableau 6). Les vacances scolaires peuvent également jouer un rôle.

Tableau 6 : Répartition du nombre de cas de gastro-entérites aigües médicalisés par jour (Auvergne, données 2009-2013, source : Données Santé publique France-DSE à partir de données de l'Assurance Maladie, SNIIRAM)

Jour	Nombre Cas GEA	Pourcentage
Lundi	97 414	24%
Mardi	76 427	19%
Mercredi	62 836	15%
Jeudi	67 485	16%
Vendredi	65 798	16%
Samedi	30 875	7%
Dimanche	12 015	3%
Total	**412 850**	**100%**

GEA : gastro-entérite aigüe

3.1.3 Critères de choix pour identifier une méthode statistique adaptée

Les réflexions précédentes permettent de lister certains critères pour guider le choix d'une méthode de détection adaptée à la problématique des épidémies de gastro-entérites aigüe d'origine hydrique (Tableau 7).

Tableau 7 : Critères de choix d'une méthode de détection adaptée à la problématique des épidémies de gastro-entérite aigüe d'origine hydrique.

Type de critères	Critères
Temporels	Données sanitaires journalières Démarche rétrospective Peu de recul historique (depuis 2010) Adaptation à une possible tendance Prise en compte de phénomènes périodiques : saison, jour de la semaine, … Prise en compte de cofacteurs : vacances scolaires, …
Géographiques	Données sanitaires agrégées par commune Données d'exposition agrégées à l'unité de distribution Prise en compte de l'inadéquation commune / unité de distribution Prise en compte de cofacteurs : distance avec les structures de soins, …
Population	Prise en compte des petits effectifs Prise en compte de cofacteurs : âge, catégorie socioprofessionnelle, …
Opérationnels	Facilité d'implémentation de l'analyse Simplicité de la méthode Paramétrisation / flexibilité

La nature même de certaines données, où le lieu d'habitation d'un individu ayant contracté une gastro-entérite aigüe est la commune, induit le recours à une méthode opérant sur des données agrégées. Au-

delà de cette restriction liée aux données, le choix d'une méthode de détection d'agrégats sera principalement guidé par la problématique de l'étude.

La détection d'agrégats a fait l'objet de nombreuses recherches depuis les années 60 (Elderer 1964). Historiquement, les dimensions temporelles, spatiales et spatio-temporelles ont été étudiées successivement. La prise en compte simultanée des dimensions spatiale et temporelle semble être l'approche la plus pertinente dans la détection d'épidémies localisées comme celles d'origine hydrique pour lesquelles ni la période, ni la zone ne sont a priori fixées (Tableau 8). L'intérêt de la détection spatio-temporelle par rapport à la détection spatiale est de prendre en compte le temps de survenue des cas. L'analyse spatio-temporelle est donc généralement plus intéressante que la détection purement spatiale (Le Strat 2015).

Les autres approches peuvent présenter un intérêt si la zone à étudier ou la période sont fixées. Dans le premier cas, si un réseau d'eau est identifié comme potentiellement à risque, une méthode d'étude de série temporelle peut permettre de mettre en évidence des phénomènes particuliers. L'approche spatiale peut également être utile lorsqu'une période d'intérêt est ciblée (par exemple à la suite de fortes précipitations) et qu'il s'agit d'identifier et localiser des évènements anormaux ayant pu survenir.

Tableau 8 : Principe de classification des méthodes de détection d'agrégats

	Zone connue	**Zone inconnue**
Période connue	Sans objet	Détection d'agrégats spatiaux
Période inconnue	Etude de séries temporelles	Détection d'agrégats spatio-temporels

3.2 Typologie des méthodes de détection d'agrégats spatio-temporels

3.2.1 Principe général

L'analyse statistique spatio-temporelle est la généralisation d'une analyse spatiale en incluant la dimension temporelle comme dimension supplémentaire. Elle peut être utilisée aussi bien pour une étude rétrospective en utilisant des données historiques, que pour la surveillance prospective, où l'analyse est répétée selon une périodicité définie (par exemple chaque jour, semaine, mois ou année).

3.2.2 Approches existantes

L'approche spatio-temporelle, plus complexe et développée plus récemment que les autres, bénéficie de moins d'outils déjà éprouvés : la plupart des publications consultées sont récentes et font état de progrès qui leur pourraient être appliqués (Demattei 2006; Unkel 2011; Pelat 2012). Néanmoins des

méthodes reconnues existent (scan de Kulldorff, régression,…). Elles permettent notamment de gérer la question des tests multiples qui se pose dès lors que de nombreux tests statistiques (comparaisons) sont effectués sur le même jeu de données. En l'absence de leur prise en compte, le risque de première espèce associé à un test statistique (alpha), qui est la probabilité de conclure à tort au rejet de l'hypothèse nulle H_0, augmente en fonction du nombre de tests effectués. En d'autres termes, lorsqu'on teste la significativité d'un grand nombre d'évènements simultanément, on augmente la probabilité que des évènements détectés soient de nature aléatoire (sans significativité au sens statistique).

3.2.3 La statistique de balayage spatio-temporelle de Kulldorff

Parmi les méthodes de détection d'agrégats spatio-temporels, la méthode développée par Kulldorff (Kulldorff 2005) est apparue comme la plus pertinente au regard des critères établis précédemment pour répondre à la problématique de détection des épidémies de gastro-entérite aigüe d'origine hydrique. Il s'agit de la méthode de référence dans le monde à l'heure actuelle. Elle bénéficie notamment d'une bonne gestion des aspects temporels (prise en compte de l'épidémie hivernale), d'une bonne prise en compte des tests multiples, d'une spécificité élevée et de l'existence d'un logiciel gratuit[11]. Elle offre par ailleurs la possibilité d'inclure des cofacteurs dans l'analyse. Bien que potentiellement intéressantes, les autres méthodes n'ont pas été retenues en premier choix en raison notamment d'une moins bonne adaptation à la détection localisée d'agrégats, une moins bonne prise en compte des petits effectifs et une mise en œuvre plus complexe (Tableau 9).

Les utilisations prenant en compte les dimensions spatiales et temporelles sont nombreuses. Par exemple, Mostashari a développé un système de surveillance du virus West Nile à partir des signalements d'oiseaux morts en l'adaptant à des incidences calculées sur fenêtres glissantes (Mostashari 2003). Assunção et Correa ont construit un système de surveillance en l'appliquant à la statistique de Shiryaev-Roberts (Assuncao & T. 2009). Heffernan et Balter ont exploité des données d'activité des urgences dans la surveillance syndromique de la gastro-entérite aigüe, de la fièvre et des infections respiratoires (Heffernan 2004; Balter 2005). Kleinman a apporté un grand nombre d'éléments de discussion sur la méthode de détection d'agrégats spatio-temporels de Kulldorff, notamment sur les covariables pouvant être intégrées au modèle. Les commentaires de Kleinman sur la méthode sont effectués à la suite de simulations sur des données quotidiennes et communales dans la détection de l'Anthrax. A quantité fixée, il observe les variations du comportement de la méthode pour des zones temporelles et spatiales de situation et dimensions variées (Kleinman 2005).

[11] http://www.satscan.org

Tableau 9 : Evaluation des méthodes de détection spatio-temporelle selon des critères adaptés à la problématique de la détection des épidémies de gastro-entérite aigue d'origine hydrique. *NB : l'importance des critères communs à toutes les méthodes (en colonne) a été attribuée arbitrairement au regard de la problématique. De la même façon, l'appréciation du niveau d'adaptation de chaque méthode pour chaque critère a été estimée arbitrairement au regard des caractéristiques des méthodes et de la problématique.*

Méthode / Critères		TEMPOREL				SPATIAL			POPULATION		OPERATIONNEL		
	Détection d'agrégats spatio-temporels localisés	Faible historique de données (< 2 ans)	Adaptation à une possible tendance	Prise en compte de périodicités : épidémie hivernale, jour de la semaine, ...	Prise en compte de cofacteurs : vacances scolaires, ...	Données agrégées à la commune	prise en compte de l'inadéquation commune / UDI	Prise en compte de cofacteurs : type de région, distance commune - pharmacie	Prise en compte des petits effectifs	Prise en compte de cofacteurs : âge, catégorie socio-professionnelle, ...	Simplicité de l'utilisation de la méthode / interprétation des résultats	Paramétrisation / flexibilité	Score
Importance du critère	3	3	3	3	2	3	3	2	3	2	3	2	
Scan spatio-temporel de Kulldorff (Kulldorff 2005)	3	2	2	2	2	1	0	2	3	2	2	2	61
Méthode de régression "classique" (Nordin 2005)	0	2	2	2	3	2	0	2	1	3	0	3	49
Approche simplifiée de Zhou-Lawson (Zhou et Lawson 2008)	0	2	2	2	3	2	0	2	1	3	0	2	47
Approche simplifiée de Kleinman (Kleinman 2004)	0	2	2	2	3	2	0	2	1	3	0	2	47
Modèle markovien caché (Lawson 2003)	0	2	2	2	1	2	1	2	1	1	0	2	42
Approche ponctuelle (Clark et Lawson 2006)	0	2	2	2	2	2	0	2	1	3	0	1	43

Très adapté	3
Bien adapté	2
Peu adapté	1
Mal adapté	0

3.2.3.1 Principe du scan spatio-temporel

Kulldorff a d'abord créé un scan spatial qui évalue si des zones géographiques circulaires présentent des valeurs anormalement élevés pour une variable donnée par rapport à leur complémentaire (Kulldorff & Nagarwalla 1995). Il a par la suite adapté son scan spatial au cadre spatio-temporel (Kulldorff 2001), ce qui permet de détecter des épidémies non plus uniquement géographiquement, mais également dans le temps.

L'analyse peut être réalisée à partir d'un cylindre à base circulaire (ou elliptique) avec une hauteur correspondant au temps (Figure 14). La base circulaire est définie exactement comme pour l'analyse spatiale, alors que la hauteur correspond à la période de recherche des agrégats potentiels dans le temps. Le cylindre est ensuite déplacé dans l'espace et dans le temps, de sorte que pour chaque zone géographique et pour chaque taille d'agrégat possible, toutes les périodes de temps possibles soient testées. On obtient ainsi un nombre important de superpositions de cylindres de tailles et de formes différentes, couvrant toute la zone d'étude, où chaque cylindre constitue lui-même un agrégat potentiel. Les cylindres qui présentent un nombre anormalement élevé de cas sont identifiés en utilisant le test du rapport de log-vraisemblance. Ce test de détection globale est conçu pour des données de nature individuelle ainsi que des données agrégées.

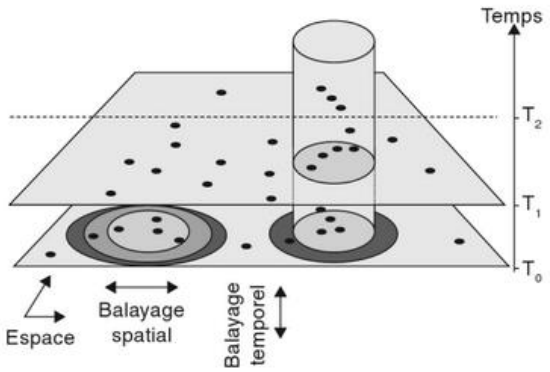


Figure 14 : Principe du scan spatial et spatio-temporel de Kulldorff, les points noirs représentant des cas (source : (Texier 2011))

Cette méthode est disponible pour plusieurs types de distributions et permet notamment de considérer des données agrégées avec une matrice de voisinage pour définir les liens entre unités sur lesquelles on regroupe les valeurs. Les améliorations de 2005 et 2007 (Kulldorff 2005; Kulldorff 2007) apportées par Kulldorff à sa méthode permettent, par un jeu de permutations sur les cas de s'affranchir des données relatives aux populations, et ainsi d'effectuer une détection d'agrégat sur des unités spatiales prédéfinies sans avoir connaissance de la répartition de la population. La méthode traite alors

uniquement des cas et non pas de l'incidence. Le scan spatio-temporel de Kulldorff permet d'intégrer des covariables au modèle. L'intérêt est d'ajouter des facteurs supposés jouer un rôle dans la répartition des cas mais qu'on ne souhaite pas voir ressortir (par exemple le jour de la semaine pour les données de l'Assurance Maladie).

3.2.3.2 Description de la méthode du scan de permutations spatio-temporelles de Kulldorff

- Le scan spatial

La fenêtre est placée successivement au sein de différents centroïdes de la zone étudiée. Un agrégat est détecté lorsque le risque à l'intérieur de la fenêtre est significativement supérieur à celui en dehors de cette fenêtre.

Les notations utilisées

$\boldsymbol{G}$ **:** espace d'étude

$\boldsymbol{\mu(G)}$ **:** nombre total d'individus dans la population à risque dans G

$\boldsymbol{n_G}$ **:** nombre total de cas observé dans G

Pour toute zone géographique A de G délimitée par une frontière, les notations sont similaires :

$\boldsymbol{\mu(A)}$ **:** nombre d'individus à l'intérieur de la zone A

$\boldsymbol{n_A}$ **:** nombre de cas observé à l'intérieur de la zone A

Les zones parcourues sont des fenêtres mobiles de forme circulaire. Pour chaque fenêtre Z :

$\boldsymbol{\mu(Z)}$ **:** nombre d'individus à l'intérieur de la fenêtre Z

$\boldsymbol{n_Z}$ **:** nombre de cas observé à l'intérieur de la fenêtre Z

$\boldsymbol{p}$ **:** risque observé à l'intérieur de la fenêtre Z

$\boldsymbol{q}$ **:** risque observé à l'extérieur de la fenêtre Z

Les hypothèses de travail

Soit Ω_0 une collection de fenêtre Z de G : $\Omega_0 = \{ Z_i \in G, i = 1, \ldots, k \}$. Cet ensemble désigne l'ensemble des fenêtres spatiales parcourues par l'algorithme. Le nombre de fenêtres ainsi que l'amplitude maximale peuvent être limitées pour ne pas surcharger l'algorithme.

L'hypothèse nulle d'absence d'agrégat à tester est la suivante :

H0 : $\boldsymbol{p = q}$ i.e. pour toute zone A de G, $nA \sim Bin\ (\mu(A), p)$

L'hypothèse alternative traduit la présence d'un agrégat.

H1 : $\boldsymbol{p > q}$ i.e. pour toute zone A de Z_i, $nA \sim Bin\ (\mu(A), p)$ et pour tout A dans $\overline{Z_i}$, $nA \sim Bin\ (\mu(A), q)$

La statistique de test permettant de détecter si l'agrégat le plus probable est significatif repose sur le rapport du maximum de vraisemblance :

$$\lambda = \frac{sup_{Z\epsilon\Omega_0, p>q} L(Z,p,q)}{sup_{Z\epsilon\Omega_0, p=q} L(Z,p,q)} = \frac{sup_{Z\epsilon\Omega_0} L(Z)}{L_0}$$

Où $L(Z) = sup_{p>q}\ L(Z,p,q)$ et $L_0 = sup_{p=q}\ L(Z,p,q)$ pour Z fixé.

La statistique calculée est un rapport du maximum de vraisemblance reposant sur le modèle de distribution adapté à l'évènement étudié (distribution de Bernoulli ou de Poisson).

L'inférence de Monte-Carlo

La significativité est testée par l'inférence de Monte-Carlo, la distribution n'étant pas connue.

Il s'agit pour chaque zone Z de Ω_0 d'effectuer un nombre M de réplicats simulés de l'échantillon sous H_0 pour lequel la statistique λ est calculée. La p-valeur associée est le nombre de réplicats dont la statistique est supérieure à celle de l'échantillon initial divisé M + 1. Un nombre de réplicats de 999 ou 9999 est généralement considéré.

- Le scan de permutation spatio-temporelle

Pour prendre en compte la dimension temporelle, Kuldorff a étendu la méthode de scan spatial à chaque étape mesurée dans le temps (Kulldorff 2001). La méthode est construite à partir de fenêtres cylindriques de base circulaire et dont la hauteur correspond à la dimension temporelle. La méthode procède de la même manière que la version spatiale pour déterminer l'agrégat le plus vraisemblable. En 2005, Kulldorf propose un modèle de permutations spatio-temporelles (Kulldorff, 2005). La différence avec le modèle précédent porte sur le calcul du nombre de cas attendus dans chaque localisation, dépendant des probabilités marginales observés.

Le nombre total de cas attendus $a_{z,t}$ dans une région Z à l'instant t dépend du nombre de cas observés sur Z dans l'ensemble de la période, ainsi que du nombre de cas observés pour l'ensemble de la région à l'instant t :

$$a_{Z,t} = \frac{1}{N}\sum_{Z'} n_{Z't} \sum_{t'} n_{Zt'}$$

Où $n_{Z't}$ (ie. $n_{Zt'}$) désigne le nombre de cas dans la région Z' (ie. Z) à l'instant t (ie. t') et N le nombre total de cas observés sur toute la région durant toute la période.

Le nombre de cas attendus pour une fenêtre cylindrique $\mathcal{F}$ est la sommation des cas attendus sur chaque région et période de la fenêtre.

$$a_{\mathcal{F}} = \sum_{(Z,t)\in\mathcal{F}} a_{Zt}$$

L'ajustement de la méthode sur une ou plusieurs covariables

Pour le scan spatial, il est possible d'ajuster le modèle sur plusieurs covariables catégorielles. L'ajustement se fait sur le calcul du nombre de sujets attendus.

Dans le cas du scan spatial, par exemple, le nombre de sujets attendus (sous l'hypothèse H_0) dans une fenêtre Z est calculé comme la proportion de cas sur l'ensemble de la région multipliée par le nombre d'individus de la population à risque à l'intérieur de la fenêtre :

$$a_Z = \mu(Z) \times \frac{n_G}{\mu(G)}$$

Considérant une covariable à k modalités, on peut associer le nombre d'individus correspondant pour chaque modalité i de la variable : ${a_Z}^i, \mu(Z)^i, {n_G}^i, \mu(G)^i$. Avec l'introduction d'une covariable, le nombre de sujet attendus est donc égal à la somme du nombre de sujets attendus de chaque modalité :

$$a_Z = \sum_{i=1}^{k} {a_Z}^i = \sum_{i=1}^{k} \mu(Z)^i \times \frac{{n_G}^i}{\mu(G)^i}$$

Lorsque plusieurs covariables sont prises en compte, l'ajustement se fait sur chaque combinaison des modalités de chaque covariable.

Le principe est le même pour le scan spatio-temporel, la formule étant plus complexe du fait que le nombre de sujets attendus est calculé au regard de la dimension temporelle en plus. Dans ce cadre-là,

l'ajustement par covariable a un intérêt s'il y a une interaction temporo-spatiale en relation avec cette covariable plutôt que de l'évolution sous-jacente de la maladie.

La fenêtre de parcours

Pour la dimension spatiale du scan de permutation spatio-temporelle, les fenêtres de parcours sont des cercles placés au centroïde de chaque unité spatiale. La méthode est d'autant plus efficace lorsqu'un agrégat possède effectivement une forme circulaire mais selon la réalité géographique, il serait utile de s'affranchir de l'utilisation de la distance euclidienne pour la détection des agrégats potentiels.

Dans le cas de données groupées par unité spatiale, il est donc possible de définir par le biais d'une matrice de voisinages la fenêtre de parcours au niveau spatial afin de ne considérer que des fenêtres de parcours présentant une vraisemblance géographique.

Il s'agit pour chaque unité spatiale i de lister l'ensemble de ces voisins (n_i) en établissant un ordre de priorité pour le parcours de chaque unité spatiale voisine. Pour chacune des unités spatiales i, la fenêtre de scan Z ne pourra inclure au maximum que n_i unités spatiales, le calcul du maximum de vraisemblance n'étant pas modifié. Les voisins peuvent être triés par une distance le long d'un système de distribution d'eau (Kulldorff 2010).

L'implémentation de la méthode

La méthode originelle, ainsi que d'autres méthodes de détection sont implémentées sous le logiciel Satscan™ développé par Martin Kulldorff et l' « Information Management Services Inc. Financial », en accès libre (Kulldorff 2010).

Ce logiciel à la particularité de présenter une interconnexion possible avec les autres logiciels utilisés communément en statistiques (R, SAS) favorisant l'automatisation d'un processus de détection.

3.2.3.3 Avantages et limites pour la détection des épidémies d'origine hydrique

La méthode de Kulldorff permet une utilisation rétrospective, sur des données journalières. Elle ne nécessite pas de disposer de données historiques et son fonctionnement lui permet de s'adapter à des tendances et/ou des périodicités, comme la saisonnalité de l'incidence de la gastro-entérite aigüe.

En outre, l'utilisation de covariables est aisée et permet de stratifier sur des cofacteurs temporels (jours de la semaine) ou caractérisant la population (âge, etc.).

Le problème de l'inadéquation entre l'exposition à l'eau du robinet qui suit la logique spatiale des unités de distribution d'eau et le géocodage des cas à la commune peut être traité en utilisant la matrice de voisinage.

Enfin, la méthode est bien documentée et beaucoup d'utilisations sont référencées. Son implémentation dans le logiciel SaTScan, et son adaptation possible à différents logiciels de statistiques facilitent son utilisation.

Deuxième partie : démarche scientifique pour améliorer l'étude et la surveillance des épidémies de gastro-entérite aigüe d'origine hydrique

Le développement d'une approche pour améliorer l'étude et la surveillance des épidémies de gastro-entérite aigüe d'origine hydrique a été réalisé en trois étapes successives, donnant chacune lieu à la production d'un article scientifique publié ou en cours de publication. Les articles, présentés dans cette partie, avaient pour objectifs respectifs :

- D'étudier la capacité des données de l'Assurance Maladie à décrire des épidémies de gastro-entérites d'origine hydrique connues (article 1) ;
- De développer une méthode intégrée pour la détection automatisée des épidémies d'origine hydrique à partir des cas de gastro-entérite aigüe médicalisés provenant des données de l'Assurance Maladie, en tenant compte du contour des unités de distribution d'eau provenant de la base nationale SISE-eaux (article 2) ;
- D'évaluer, sur la base d'une étude de simulation, les performances et les facteurs influençant la détection des épidémies d'origine hydrique (article 3).

1 Etude de la capacité des données de l'Assurance Maladie à décrire des épidémies de gastro-entérite aigüe d'origine hydrique connues

Article 1 : Description de deux épidémies d'origine hydrique en France : étude comparative à partir de données d'études de cohorte et de bases de données médico-administratives

Auteurs : Mouly D.[1], Vincent N.[1], Vaissière E.[1], Van Cauteren D. [1], Beaudeau P.[1], Ducrot C.[2], Gallay A.[1]

[1] : Institut de veille sanitaire, Saint Maurice ; [2] Inra – Unité d'épidémiologie animale, Theix

Epidemiology and Infection: 144, 591-601 (2016).

Présentation synthétique du travail réalisé et des résultats de cet article

La sous détection des épidémies de gastro-entérites aigües d'origine hydrique dans les pays développés comme la France encourage à rechercher des nouvelles sources de données pour améliorer leur détection et leur surveillance. Depuis plusieurs années, les données de l'Assurance Maladie constituent une source de données pertinente pour la surveillance du risque infectieux d'origine hydrique au travers la surveillance syndromique de la gastro-entérite aigüe. Un algorithme spécifique précédemment développé permet d'identifier les cas de gastro-entérite aigüe médicalisés par jour et par commune en France.

L'objectif de cette étude était d'évaluer la capacité des données de l'Assurance Maladie pour décrire de façon rétrospective des épidémies connues de gastro-entérites aigües d'origine hydrique.

Deux épidémies, survenues dans la région Auvergne en juin 2010 et avril 2012 et ayant fait l'objet d'études de cohorte, ont été utilisées comme support pour notre étude. Les données des études de cohortes étaient issues de travaux précédents (Daures 2011; Mouly 2013) et les données de l'Assurance Maladie ont été extraites du Système national d'information inter-régime (Sniir-AM) pour cette étude. Trois définitions de cas ont été utilisées en fonction de la source de données utilisée : i) un « cas cohorte » défini comme une personne résidant dans les communes impactées au moment de la pollution et ayant présenté au moins 3 selles en 24h ou des vomissements, ii) un « cas cohorte avec consultation médicale » défini comme un « cas cohorte » ayant consulté un médecin, iii) un « cas Sniir-AM » défini comme une personne résidant dans les communes impactées et identifiée par l'algorithme de sélection comme un cas de gastro-entérite aigüe avec une date de consultation médicale dans les 3 semaines qui ont suivi la pollution. Les deux épidémies ont été analysées de façon indépendante. Les deux sources de données ont été comparées sur des indicateurs épidémiologiques : impact sanitaire,

courbe épidémique, description des cas. La corrélation des courbes épidémiques a également été testée en faisant varier le niveau d'agrégation temporel des cas cohortes et des cas Sniir-AM.

Les résultats sont contrastés avec pour l'épidémie de 2010 à *Campylobacter* une bonne corrélation temporelle journalière des cas cohortes et des cas Sniir-AM (Figure 1, article ci-après) et pour l'épidémie de 2012 à Norovirus une faible superposition des courbes épidémiques (Figure 2, article ci-après). Il a été montré que la corrélation était maximisée en agrégeant les cas sur plusieurs jours (3 jours pour l'épidémie de 2010 et 5 jours pour celle de 2012) (Figure 3 & 4, article ci-après).

Quelle que soit l'épidémie, l'impact sanitaire estimé à partir des données de cohorte est très supérieur aux données de l'Assurance Maladie. Il varie d'un facteur 5 pour l'épidémie de 2010 (254 « cas cohorte » et 54 « cas Sniir-AM ») à 17 pour celle de 2012 (458 « cas cohorte » et 26 cas « Sniir-AM »). Ces différences peuvent en partie s'expliquer par des biais identifiés qui tendent à surestimer l'impact dans les études de cohorte en raison des modalités d'interrogation des personnes (auto-questionnaire déposé dans les boites aux lettres) et à le sous-estimer dans les données de l'Assurance Maladie pour lesquelles plusieurs déterminants peuvent influencer le recours aux soins des malades. Parmi ces déterminants, certains sont indépendants de l'origine hydrique de l'épidémie (distance par rapport aux médecins et pharmacies, gravité de la maladie, automédication), d'autres peuvent être expliqués par des habitudes de la population en fonction de leur niveau de connaissance et de sensibilisation au risque lié à l'eau. Ainsi, l'épidémie de 2012 pour laquelle on observe une différence importante entre le nombre de « cas cohortes » et le nombre de « cas Sniir-AM », a touché une commune qui a déjà connu des épisodes de pollution d'eau et des épidémies d'origine hydrique. Dans cette situation, le signal à détecter dans les données de l'Assurance Maladie est bien moindre.

Il est également montré que les taux de cas médicalisés dans la population impactée (c'est-à-dire la proportion de cas de gastro-entérite aigüe issus des données de l'Assurance Maladie) étaient faibles, de l'ordre de 1,5 à 2% (cf. Table 2, article ci après). Enfin, les deux épidémies sont caractérisées par un taux de cas médicalisés plus important chez les enfants de moins de 15 ans que chez les adultes. Le taux le plus faible étant observé chez les personnes âgées (65 ans ou plus).

En conclusion, les résultats de cette étude montrent que pour deux épidémies d'origine hydrique, les données de l'Assurance Maladie peuvent décrire avec une précision variable ce type d'évènement. Elles permettent également de générer un signal épidémique avec un profil plus ou moins comparable à celui obtenu avec une étude de cohorte. Il est notamment montré que l'agrégation des cas de gastro-entérite aigüe provenant de l'Assurance Maladie sur plusieurs jours (au minimum 3 jours) génère un signal épidémique plus proche de celui des études de cohorte. Ainsi, l'existence d'un signal sanitaire

dont l'emprise géographique est cohérente avec les contours des réseaux d'eau potable impactés justifie l'emploi des données de l'Assurance Maladie pour la détection rétrospective des épidémies de gastro-entérites aigües d'origine hydrique.

Epidemiol. Infect., Page 1 of 11. © Cambridge University Press 2015
doi:10.1017/S0950268815001673

Description of two waterborne disease outbreaks in France: a comparative study with data from cohort studies and from health administrative databases

D. MOULY[1]*, D. VAN CAUTEREN[1], N. VINCENT[1], E. VAISSIERE[1], P. BEAUDEAU[1], C. DUCROT[2] AND A. GALLAY[1]

[1] *French Institute for Public Health Surveillance, Saint-Maurice, France*
[2] *INRA, Epidemiology Animal Unit, Clermont-Ferrand – Theix, France*

Received 29 January 2015; Final revision 1 June 2015; Accepted 29 June 2015

SUMMARY

Waterborne disease outbreaks (WBDO) of acute gastrointestinal illness (AGI) are a public health concern in France. Their occurrence is probably underestimated due to the lack of a specific surveillance system. The French health insurance database provides an interesting opportunity to improve the detection of these events. A specific algorithm to identify AGI cases from drug payment reimbursement data in the health insurance database has been previously developed. The purpose of our comparative study was to retrospectively assess the ability of the health insurance data to describe WBDO. Data from the health insurance database was compared with the data from cohort studies conducted in two WBDO in 2010 and 2012. The temporal distribution of cases, the day of the peak and the duration of the epidemic, as measured using the health insurance data, were similar to the data from one of the two cohort studies. However, health insurance data accounted for 54 cases compared to the estimated 252 cases accounted for in the cohort study. The accuracy of using health insurance data to describe WBDO depends on the medical consultation rate in the impacted population. As this is never the case, data analysis underestimates the total number of AGI cases. However this data source can be considered for the development of a detection system of a WBDO in France, given its ability to describe an epidemic signal.

Key words: Investigation, outbreaks, surveillance system, waterborne infections.

INTRODUCTION

Waterborne disease outbreaks (WBDO) are a public health concern in France because of the proportion of people affected when contamination of drinking water occurs. Almost all WBDO result in outbreaks of acute gastrointestinal infection (AGI) and for most of these, the attack rate in an exposed population reaches 20–50% in France [1]. Children and people with low immunity are usually the most affected. To date, detection of these events is mainly based on the reporting of clusters of AGI by general practitioners (GPs) to health authorities. Consequently, the number of WBDO is probably underestimated in France due to the absence of a specific surveillance system. Improving the detection of infections caused by contaminated drinking water regarding public health, is a challenge to improving the knowledge of risk factors, identifying the drinking water networks with high risk, and proposing appropriate preventive

* Author for correspondence: Mr D. Mouly, InVS-Dcar-Cire Midi-Pyrénées, 10 chemin du raisin 31050 Toulouse, Cedex 9, France. (Email: damien.mouly@ars.sante.fr)

measures. In this context, the French Institute for Public Health Surveillance is exploring the possibility of using the health administrative databases from the French Health Insurance to develop a national automated detection system of WBDO.

Healthcare administrative databases, which collect data for management and medical purposes, are increasingly used for epidemiological surveillance in developed countries. Several studies using these types of databases have already highlighted their strengths and weaknesses with respect to accurate disease surveillance [2]. In France, an algorithm was specifically developed to identify AGI cases in 2011. It uses data on reimbursement for payment of prescribed drugs from the French National Health Insurance Information System (SNIIRAM; Système national d'information inter régimes de l'Assurance maladie) database [3]. The SNIIRAM database covers 98% of the French population and collects both administrative and individual medical information [4]. Therefore analysis of this data source constitutes one possible approach to develop a detection system of WBDO resulting in AGI. From this perspective, the ability of the SNIIRAM database to describe a WBDO has first to be evaluated.

The benefits of syndromic surveillance to describe WBDO, compared to pharmacy over-the-counter sales data, emergency department visits and even epidemic curves related to AGI have been well documented [5–9]. Nevertheless, to date no comparative study has been published in France to evaluate the use of the SNIIRAM database for the description of WBDO resulting in AGI.

The primary purpose of this study was to compare the SNIIRAM data with a classic epidemiological approach (population-based cohort study) for the description of WBDO. This comparison would improve our knowledge of benefits and limits of SNIIRAM data to describe WBDO with the aim of developing an automated system for their detection with this data source (in process).

MATERIAL AND METHODS

Two different WBDO which occurred in France in 2010 and 2012 were selected for this comparison. For each WBDO, retrospective cohort studies were conducted during both outbreaks and institutional reports (in French) were edited [10, 11]. In the present study, data collected from SNIIRAM will be compared to data previously collected during cohort studies. The comparison focused on the epidemic curves, the number of cases, individual characteristics (age group, gender), and the extent of the outbreaks.

The two selected AGI WBDO occurred in three municipalities located in the Auvergne region in France, in June 2010 ('WBDO A') and April 2012 ('WBDO B'). The main characteristics of both outbreaks and affected populations are summarized in Table 1.

Data from cohort studies

Two cohort studies were conducted in the population of the municipalities served by the polluted drinking water network (40% of the total municipal population in WBDO A, i.e. 1067 inhabitants, and 100% in WBDO B, i.e. 1753 inhabitants). Three weeks after the beginning of each WBDO, self-administered questionnaires were distributed in the mailboxes of all households served by the contaminated water networks. One overall 'household' questionnaire and four 'individual' questionnaires were distributed to each household. An information letter and a self-addressed, stamped return envelope were also distributed. Data were collected on individual characteristics (age, gender), clinical symptoms (dates of symptom onset, nature and duration), the use of healthcare (medical consultation, date of consultation) and consumption habits of tap water.

For WBDO A, the circumstances that may have led to contamination of the drinking water system included 3 consecutive days of heavy rain, flooding of the system's drinking water borehole and of the mechanical chlorination system (the only treatment mechanism in place). For WBDO B, an incident with the system's sand filter followed by a malfunction of the turbidity alarm was responsible for the introduction of polluted raw water (river) into the drinking water system.

Data from SNIIRAM

SNIIRAM aims at evaluating beneficiaries' healthcare consumption and associated expenditures. It covers more than 98% of the French population and records all reimbursements to patients for out-of-pocket medical procedures, medications and payments to professionals for consultations [4]. AGI medications are included in this database if they are reimbursable, prescribed by a GP and dispensed in a pharmacy. The identification of AGI cases in the two WBDO above

Table 1. *Description of study sites, population and criteria for waterborne disease outbreaks A and B, France (own data, not previously published, available in institutional reports [10, 11])*

Type of waterborne disease outbreak	Waterborne disease outbreak 2010 (WBDO A) [10]	Waterborne disease outbreak 2012 (WBDO B) [11]	
Municipalities impacted	Pérignat les sarliève	Pleaux; Barriac les Bosquet	
Municipal population (Insee, 2010)	2696 inhabitants	1753 inhabitants	
Distribution by age group (years)	*N* (%)	*N* (%)	*P*†
0–5	134 (5%)	44 (3%)	$<10^{-3}$
6–14	393 (15%)	133 (8%)	$<10^{-3}$
15–64	1753 (65%)	968 (55%)	$<10^{-3}$
⩾65	416 (15%)	608 (35%)	$<10^{-3}$
Distribution by gender			
Male	1341 (49·7%)	864 (49·3%)	
Female	1355 (50·3%)	889 (50·7%)	
Drinking water supply of municipalities			
Number of drinking water networks impacted by the pollution	2/4	3/3	
Number of people supplied by polluted drinking water (% of all inhabitants)	1067 people (39·6%)	1753 people (100%)	
Type of source	Mountain spring + borehole in alluvial aquifer	Borehole in surface river	
Type of treatment	Disinfection (Cl_2)	Pre-oxidation + clarification + filtration + disinfection (ClO_2)	
Occurrence of pollution at the time of waterborne disease outbreak			
Circumstances of occurrence of pollution	Heavy rains Flooding of the borehole Cessation of chlorination	Heavy rains River pollution Operating incident in treatment plant + alarm malfunction	
Date of the pollution intrusion into the drinking water network	17 June 2010	7 April 2012	
Supposed duration of exposure to polluted tap water*	8 days	5 days	
Faecal contamination indicators in drinking water network	>100 c.f.u. per100/ml *E. coli*	>100 c.f.u. per100/ml *E. coli*	

c.f.u., Colony-forming units.
* Delay between pollution intrusion and restrictions on water consumption.
† *P* value of similar population in both municipalities.

required two consecutive steps: (i) data extraction from the SNIIRAM database and (ii) using the AGI algorithm developed by Bounoure *et al.* [3] for selecting AGI cases. The criterion for the data extraction step was the reimbursement for at least one prescribed target drug used to treat AGI† bought by people living in the impacted municipality. The criteria for the AGI discriminative algorithm were: the delay between the prescription and delivery of drugs (<24 h), the number of different AGI-specific drugs prescribed, the treatment duration (<8 days), and the co-prescription of non-AGI specific drugs (e.g. anti-cancer drugs). Information on age, gender, date of consultation and place of residence was available for each AGI case.

† Antiemetic drugs – ATC classification: A04A, A03 F; anti-diarrhoea drugs – A07X, A07D; intestinal adsorbents drugs – A07B, A02X and oral rehydration salts.

Case definitions

In cohort studies, a case of waterborne AGI was defined as any person in the population exposed to contaminated drinking water, with ⩾3 stools in a 24-h period or vomiting [12] within 3 weeks following contamination of the water system. These cases were defined as 'cohort cases'. Of these cases, those consulting a GP were defined as 'cohort cases with GP consultation'.

Using the SNIIRAM data, people living in the impacted municipalities who consulted a GP within 3 weeks after contamination and who then went to a pharmacy to buy medications prescribed to treat AGI, were defined as 'SNIIRAM cases'.

Data comparison

Description of WBDO

Several epidemiological parameters were used for the description of cohort studies: the *attack rate* in the population was estimated using the ratio between cohort cases and the total number of respondents of the cohort studies. The attack rate was used to estimate the total number of AGI cases in the general population. The *consultation rate* was defined as the ratio between cohort cases who consulted a GP and the total number of cohort cases. Finally, the number of AGI cases in the general population who consulted a GP was estimated from cohort studies by applying age-based consultation rates to the number of AGI cases estimated in the general population.

For SNIIRAM data, the *medication rate* was estimated by comparing SNIIRAM cases with the total population of municipalities impacted (2696 people in WBDO A and 1753 in WBDO B).

The total number of cases assessed from cohort studies and the number of SNIIRAM cases were compared.

Additional comparisons between both data sources included:

- The duration of the epidemic, which was arbitrarily defined as the period covering 90% of the cohort or SNIIRAM cases and starting with the day when at least 5% of the cases had already occurred.
- The delay between the contamination of the water system and the peak of the epidemic curve.
- The distribution of gender and age groups (by applying Fisher's exact test).

Analysis of the correlation between SNIIRAM data and cohort studies

With the view of using SNIIRAM data for the detection of WBDO we tested the similarity of both signals (SNIIRAM *vs.* cohort) by variation of two parameters: (i) the temporal window of aggregation of AGI cases from 1 to 7 days, and (ii) the lag of the two series of AGI cases (SNIIRAM *vs.* cohort) from 0 to 7 days. A correlation coefficient between the two time series was estimated for each pair of values (aggregation level, lag).

RESULTS

Characteristics of outbreak cases

General data

The WBDO A cohort study identified 74 cases (attack rate = 18·1%) in 408 respondents (response rate = 38·2%). Of these, 27 people had consulted a GP (consultation rate = 36·5% [13]) (Table 2). The number of AGI cases in the affected population was estimated from the cohort study at 252, of whom 97 had consulted a GP. The ratios between the number of SNIIRAM cases ($n = 54$) and respectively, the number of AGI cases in the affected population who had consulted a GP (cohort-based estimation), and total AGI cases (cohort-based estimation), were 0·56 (95% confidence interval (CI) 0·42–0·81) and 0·21 (95% CI 0·16–0·31). The pathogen agent identified in the cohort study for WBDO A was *Campylobacter jejuni* in 2/12 patients' stools [13].

In WBDO B, the attack rate estimated in the cohort study was 25·4% (171 cases) for 674 respondents (response rate = 38·4%) [13]. Of these, 50 people had consulted a GP (consultation rate 29·2%). The number of AGI cases in the population was estimated at 458, of whom 123 had consulted a GP. The ratios of cases estimated (see for WBDO A above) were 0·21 (95% CI 0·17–0·28) and 0·06 (95% CI 0·05–0·07). In WBDO B, pathogen agent identified was norovirus genogroup 2 in 4/5 patients' stools [13].

By gender

Men and women were equally affected in both outbreaks and both data sources. The sex ratio (female/male) in both WBDO A and B, respectively, was 1·6 and 0·9 in the cohort studies [13], and 1·0 and 1·4 using SNIIRAM data.

By age group

In both outbreaks, the age groups most affected included children aged <15 years: those aged 6–14 years in the cohort studies [13] (attack rate = 43·1% in WBDO A and 42·9% in WBDO B) and those aged 0–5 years using SNIIRAM data (medication rate = 9·0% in WBDO A and 2·3% in WBDO B) (Table 2).

Table 2. *Description of outbreak cases from cohort studies and Health Insurance data, France, June 2010 and April 2012*

	Cohort study*				SNIIRAM data		
	WBDO A (408 respondents); WBDO B (674 respondents)				Cases with medical consultation followed by reimbursed purchase of drugs (SNIIRAM cases)		
	Cohort cases		Cohort cases with medical consultation				
	N (%)	Attack rate† % (95% CI)	*N* (%)	Consultation rate in cohort‡ % (95% CI)	*N* (%)	Medication rate in population§ % (95% CI)	*P*‖
WBDO A							
Gender							
Male	29 (39·2)	15·4% (10·3–20·6)	10 (37·0)	34·5% (17·2–51·8)	27 (50·0)	2·0% (1·3–2·8)	0·279
Female	45 (60·8)	20·5% (15·2–25·9)	17 (63·0)	37·8% (23·6–51·9)	27 (50·0)	2·0% (1·2–2·7)	
Age group (years)							
0–5	4 (5·4)	14·1% (0·2–28·0)	2 (7·4)	50·0% (1·0–99·0)	12 (22·2)	9·0% (4·1–13·8)	0·006
6–14	16 (21·6)	43·1% (28·9–57·2)	9 (33·3)	56·3% (31·9–80·6)	21 (38·9)	5·3% (3·1–7·6)	0·048
15–64	44 (59·5)	23·3% (17·7–28·8)	14 (51·9)	31·8% (18·1–45·6)	19 (35·2)	1·1% (0·6–1·6)	0·008
⩾65	10 (13·5)	4·9% (0·9–8·9)	2 (7·4)	20·0% (0·0–44·8)	2 (3·7)	0·5% (0·0–1·1)	0·071
Total cases	**74 (100·0)**	**18·1% (14·4–21·9)**	**27 (100·0)**	**36·5% (25·5–47·5)**	**54 (100·0)**	**2·0% (1·5–2·5)**	
Total cases estimated#	**252 (175–328)**		**97 (67–128)**				
WBDO B							
Gender							
Male	88 (51·5)	27·2% (22·3–32·0)	24 (48·0)	27·3% (18·0–36·6)	11 (42·3)	1·3% (0·5–2·0)	0·408
Female	83 (48·5)	23·7% (19·3–28·2)	26 (52·0)	31·3% (21·3–41·3)	15 (57·7)	1·7% (0·8–2·5)	
Age group (years)							
0–5	4 (2·3)	23·4% (3·3–43·5)	1 (2·0)	25·0% (0·0–67·4)	1 (3·8)	2·3% (0·0–6·7)	0·511
6–14	17 (9·9)	42·9% (28·5–57·4)	4 (8·0)	23·5% (3·4–43·7)	2 (11·5)	2·3% (0·0–4·8)	1·00
15–64	91 (53·2)	27·8% (23·1–32·5)	16 (32·0)	17·6% (9·8–25·4)	13 (61·5)	1·7% (0·8–2·5)	0·834
⩾65	59 (34·5)	19·9% (15·1–24·8)	29 (58·0)	49·2% (36·4–61·9)	3 (23·1)	1·0% (0·2–1·8)	0·022
Total cases	**171 (100·0)**	**25·4% (22·1–28·7)**	**50 (100·0)**	**29·2% (22·4–36·1)**	**26 (100·0)**	**1·5% (0·9–2·0)**	
Total cases estimated#	**458 (355–561)**		**123 (94–152)**				

CI, Confidence interval.

* Own data, not previously published, available in institutional reports [10, 11].

† The attack rate was estimated for respondents (408 in WBDO A and 674 in WBDO B).

‡ The consultation rate was estimated for AGI cases in cohort studies (74 in WBDO A and 171 for WBDO B).

§ The medication rate was estimated for total population of municipalities impacted (2696 in WBDO A and 1753 in WBDO B).

‖ The *P* value compare the distribution of gender and age groups in the cohort study versus the SNIIRAM cases.

Estimation of total cases in the population impacted from cohort studies (1067 people in WBDO A and 1753 in WBDO B).

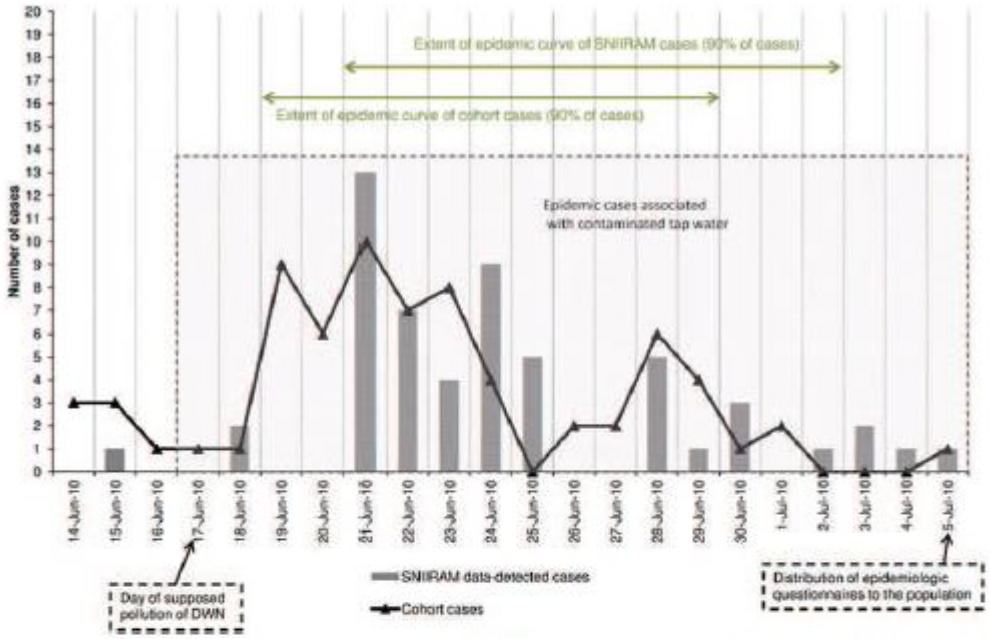


Fig. 1. Description of daily numbers of cohort cases and SNIIRAM cases 14 June 2010 to 5 July 2010 for WBDO A, Pérignat les Sarliève, France, June 2010. In the WBDO A cohort study, missing data existed for ten cases (14%) regarding the date of onset of symptoms, and consequently could not be represented. Cohort data were collected during the outbreak (own data, not previously published, available in institutional reports [10]). DWN, Drinking water network.

In both outbreaks and with both data sources, persons aged >64 years were the least affected age group. Estimated proportions of cases in this age group were from 0·5% to 1·0% using the SNIIRAM data and from 4·9% to 19·9% in the two cohort studies [13]. This age group was also characterized by a different rate of GP consultations compared to other age groups. The lower rate was observed in WBDO A (20%) and the higher in WBDO B (49%), irrespective of the estimation means used.

The 15–64 years age group represented the intermediate age group for the estimated cohort rates and for medication in the SNIIRAM data.

Comparison of epidemic curves

In WBDO A, the temporal distribution of SNIIRAM cases was similar to the distribution of cohort cases (Fig. 1). The duration of the epidemic using SNIIRAM data was 12 days (21 June to 2 July 2010), peaking on 21 June 2010. These results were similar with the cohort data [13]: an epidemic duration of 11 days (19 June–29 June 2010) with an epidemic peak on June 21 (Fig. 1), 4 days after the contamination of the water system. A secondary peak was observed for both data sources on 28 June 2010.

In WBDO B, temporal distribution of SNIIRAM cases was different from the distribution of cohort cases, no large increase in the number of cases nor the epidemic peak being observed (Fig. 2). Using the cohort data [13], the duration of the epidemic was estimated at 14 days (8 April–21 April 2012) with an epidemic peak on 12 April 2012, 5 days after the contamination of the water system (Fig. 2).

Correlation between SNIIRAM data and cohort studies

The aggregation of cases over 3 days in WBDO A (lag = 1 day) and 5 days in WBDO B (lag = 5 days) is associated with the highest correlation coefficient (respectively 0·83 and 0·94) between epidemic curves from SNIIRAM data and cohort studies (Figs 3 and 4).

DISCUSSION

Our study evaluated the possibility of using the SNIIRAM database to describe WBDO, by comparing the results from the former's data with

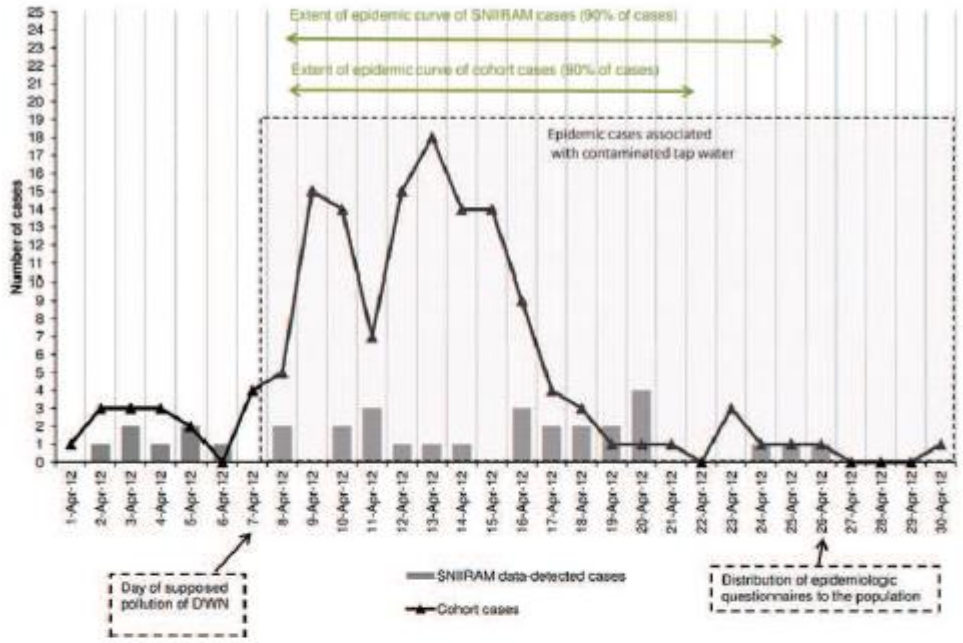


Fig. 2. Description of daily numbers of cohort cases and SNIIRAM cases 1 April 2012 to 30 April 2012 for WBDO B, Pleaux, France, April 2012. In the WBDO B cohort study, 39 (23%) cases had missing data regarding the date of onset of symptoms of cases, and consequently could not be represented. Cohort data were collected during the outbreak (own data, not previously published, available in institutional report [11]). DWN, Drinking water network.

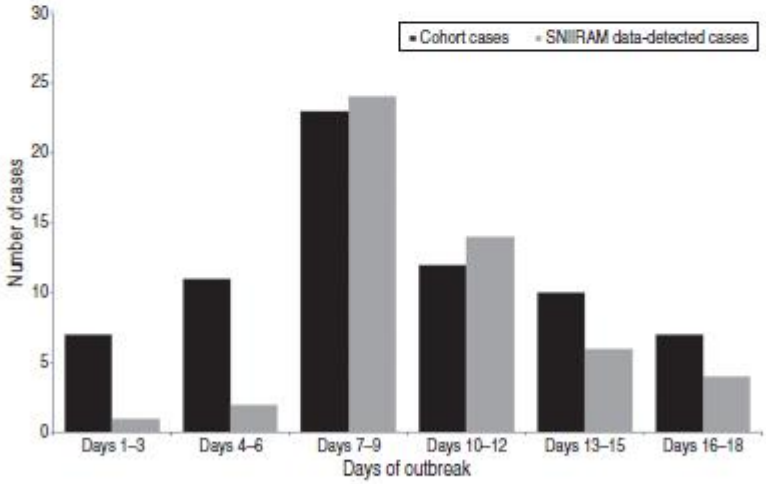


Fig. 3. Distribution of cohort cases and SNIIRAM cases aggregated over 3 days and applying a lag of 1 day on SNIIRAM cases – WBDO A, Pérignat les Sarliève, France, June 2010.

results from two population-based cohort studies. Results of our comparative study point out the benefits and limits of SNIIRAM data for their use in an automated detection system for WBDO as discussed below.

Interpretation of data comparison between SNIIRAM data and cohort studies

The comparison of the two epidemic curves created using data from the cohort studies and from

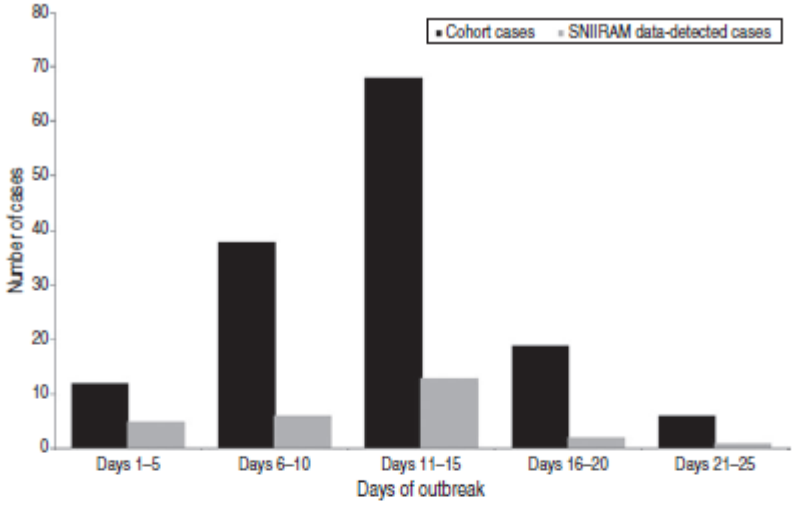


Fig. 4. Distribution of cohort cases and SNIIRAM cases aggregated over 5 days and applying a lag of 5 days on SNIIRAM cases – WBDO B, Pleaux, France, April 2012.

SNIIRAM showed an accurate representation of the epidemic by the SNIIRAM data in WBDO A. The duration of the epidemic was similar in both curves, the peak of the epidemic occurring the same day in both data sources. In France, it is estimated that more than nine out of 10 AGI cases consult within 3 days of illness onset [14]. To be detected in the SNIIRAM database the delay between the GP visit and drugs delivery in a pharmacy had to be <24 h. Therefore a delay between 0 and 4 days was expected between the cohort cases (date of illness onset) and the SNIIRAM cases (date of GP consultation) for both the outbreak duration and outbreak peak. This delay was not observed in WBDO A and may be explained by the fact that 19–20 June 2010 was on Saturday–Sunday, leading to fewer GP consultations, and reduced or no healthcare utilization. Correlation analysis shows that an aggregation of cases over 3 days allows the optimization of the epidemic signal with data from SNIIRAM (highest coefficient between SNIIRAM and cohort).

For WBDO B no peak was observed using SNIIRAM data. This may be explained by the following factors: first, WBDO B occurred the first day of the Easter weekend (7–9 April 2012). During this period health services were closed and therefore healthcare utilization was limited, with few cases being identified in the SNIIRAM data analyses. Second, school holidays continued for 2 weeks following the Easter weekend (7–22 April 2012). Third, alternative healthcare utilization (e.g. a home visit by a nurse where no prescription was written) cannot be excluded given the 3-day closure period of medical services. Finally people reported previous episodes of drinking water pollution in WBDO B. The knowledge of risk by inhabitants of the municipality (repeated pollution) may have led to more use of the family medicine chest or over-the-counter drugs without medical consultation.

However, by aggregating cases over 5 days in WBDO B, we improved the correlation level (highest coefficient) between SNIIRAM and cohort.

Overall sensitivity of the SNIIRAM data – detected cases

The SNIIRAM data accounted for 21% and 6% of all AGI cases estimated from cohort studies in the population during WBDO A and WBDO B, respectively. These proportions are lower than consultation rates observed in a national population-based study (33%) [14]. Nevertheless, the number of total cases from cohort studies could be overestimated. Indeed, it is possible that ill people were more likely to participate in cohort studies, because of the procedure involved for interviewing people (i.e. the use of a voluntary, self-administered questionnaire). This may have constituted a source of selection bias, leading to an overestimation of the number of AGI cases in the population [15, 16].

Using the attack rate usually associated with WBDO in France (from 20% to 50%) [1], the expected health impact from the SNIIRAM data analysis – percentage of medical cases – would lie between 1% and 10% of people exposed to polluted drinking water. This sensitivity could affect the capacity for detection of WBDO based on the use of SNIIRAM data, especially when a small population is served by contaminated drinking water.

Factors influencing the sensitivity of SNIIRAM indicator for AGI

The overall sensitivity of SNIIRAM data for the description of WBDO may have been influenced by algorithm discrimination of AGI cases, healthcare-seeking behaviour for AGI and access to health services, age and the nature of pathogen.

AGI algorithm

The selection of a case of AGI using SNIIRAM data was dependent on the AGI definition case implemented in the algorithm [3]. The algorithm was cross-validated on a national level with data from the National GP Sentinel Network [17] and with data from a population-based national study [14]. Results showed a good representativeness of the seasonality when using SNIIRAM data, compared to the Sentinel Network, and an estimated annual incidence rate equivalent to that obtained from the national study [14]. Furthermore, the intrinsic sensitivity and specificity of the algorithm were evaluated, each reaching almost 90% [3]. In the context of localized AGI outbreaks such as WBDO, the number of AGI cases selected using SNIIRAM data may be sensitive to the proportion of older people (>65 years) involved. For this age group, we set the selection algorithm more towards specificity, as many treatments, including anti-diarrhoeal medications, are prescribed for reasons other than AGI.

Healthcare-seeking behavior for AGI and access to health services

Several determinants of treatment for AGI in the population may affect the sensitivity of the SNIIRAM indicator as we observed in our study. Although 76% of AGI cases in France use medication, most utilize the family medicine chest (42%) [14]. For these cases, the SNIIRAM data source is blind because of the absence of consultations and prescriptions. Only cases who consulted a physician for AGI (33%) were registered in the SNIIRAM data: 31% of these consulted a GP, 1% a paediatrician and 1% visited the hospital. Alternative healthcare, such as home visits by a nurse, is not visible in SNIIRAM data because of the absence of a drug prescription. Neither does SNIIRAM data take into account over-the-counter medicines bought at a pharmacy without prescription. However, other data collecting over-the-counter information exist in France. Despite the quick availability of over-the-counter data, we considered this data source to be less appropriate for surveillance of WBDO than SNIIRAM data, because of its lack of specificity and the fact that its spatial resolution (pharmacy) did not overlap with the drinking water distribution system [18].

Age and the nature of the pathogen

Deciding to consult a GP is dependent on the person's age and the nature of the pathogen (causative agent). In both WBDO, we observed that younger people (<15 years) were those most affected by the disease (higher attack rate), irrespective of the pathogen in question (*Campylobacter* sp. in WBDO A and norovirus WBDO B). In WBDO B, the particularly high attack rates in young children can be explained by the greater sensitivity of children to contract AGI, and by the causative agent (virus) of the disease in which secondary transmission plays a more important role than in adults. The GP consultation rate was also higher in younger people. Consequently, relative to the cohort studies, a higher proportion of younger cases occurred using SNIIRAM data (children aged <15 years accounted for 60% of cases from SNIIRAM data in WBDO A *vs.* 27% from the cohort study). This implies improved sensitivity of the SNIIRAM data for AGI for these age groups.

Consultation rates following WBDO A and B were consistent with those found in a published study showing that the frequency of visits to a GP was more often associated with bacterial than viral infections [19]. Similar trends resulting from waterborne disease outbreaks of AGI have been highlighted in other cohort studies reporting behavioural differences in outbreak situations (e.g. consultation rate = 52% in Gourdon [20]).

Counting waterborne AGI cases

Waterborne outbreak cases, i.e. AGI cases resulting from the consumption of contaminated tap water,

were defined in our study as any AGI case occurring after the day polluted water was introduced into the network. This definition does not distinguish between individual AGI cases due to contaminated drinking water and the baseline of AGI cases. Taking into account the size of the respective municipalities involved, and weekly incidence of AGI reported by the National GP Sentinel Network [17], the number of cases not associated with drinking polluted water during the two WBDO would be 0·75 for WBDO A, and 1·14 for WBDO B. During annual winter outbreaks of AGI with mainly person-to-person transmission, description of a WBDO would necessitate removing cases directly related to the winter outbreak.

Implication for waterborne disease detection

Several studies which deal with the question of the implications for syndromic surveillance of AGI illness or WBDO detection have been published previously [5, 9]. A recent study has compared the ability of three sources of syndromic data (telephone triage, over-the-counter sales, web queries) for the detection of local outbreak signals [6]. Nine outbreaks, which each involved more than 100 cases, were selected. The authors concluded that four out of nine point-source outbreaks were validated in the telephone triage of AGI and two in over-the-counter sales. The three largest outbreaks detected were associated with drinking water contamination and reported between 2400 and 27 000 AGI cases.

Unlike our study, the size and duration of the detected outbreak in Andersson *et al.* [6] were much higher than WBDO A or B. Furthermore, indicators for AGI were established from pre-clinical data, i.e. without medical consultation (which was a prerequisite for SNIIRAM cases). Therefore, one can assume that telephone triage and over-the-counter sales are more sensitive and more readily available than SNIIRAM data, despite their lower specificity.

Furthermore, the challenge of WBDO detection addressed in published studies [5, 6, 9] highlights the difficulty of detecting short outbreaks involving fewer than 100 cases. For this purpose, information collected for cases has to have sufficient temporal (ideally the day) and spatial (municipality could be sufficient) resolution to allow the detection of local outbreak signals like WBDO. Correlation analysis in our study suggests taking into account the aggregation of cases over several days (e.g. 3 and 5 days in WBDO A and WBDO B, respectively) to optimize the detection of the epidemic signal.

In addition, syndromic surveillance is useful to estimate the size, duration and health impact of detected outbreaks, as we know the consultation rate in the impacted population. This estimation should take into account factors influencing the consultation rate, in particular age and access to health services as shown in our study.

From a public health point of view, detected epidemic signals from SNIIRAM data should be followed by a set of operational measures, including field investigation. These were conducted to validate and describe the outbreak, and to understand the origin and mechanisms involved in case diffusion in order to influence decision-making for public health prevention.

CONCLUSION

We evaluated the ability of SNIIRAM data to describe a WBDO. Our work helped to provide parameters for the description of WBDO of AGI using data from SNIIRAM. It also identified benefits and limits of syndromic surveillance for the detection of WBDO. However, the results of this study, based on two well-documented WBDO, cannot be extrapolated to all WBDO situations and could only be confirmed and complemented by other comparative studies. Nevertheless, the results do allow us to conclude that the use of SNIIRAM data could improve the detection of AGI WBDO with respect to the current surveillance system which is mainly based on GP voluntary reporting. Finally, taking tap water sources of exposure into account in the method of detection of AGI WBDO requires the development of an integrated approach which ensures that data on administrative delimitation of municipalities (aggregation area of AGI cases) and delimitation of the drinking water distribution units (ecological unit of exposure to tap water) can be overlapped.

ACKNOWLEDGEMENTS

The authors thank National Health Insurance for access to health administrative data. We also thank Magali Corso and Grégoire Falq from InVS for the preparation of case data of acute gastroenteritis from health administrative databases.

DECLARATION OF INTEREST

None.

REFERENCES

1. **Beaudeau P, *et al.*** Lessons learned from ten investigations of waterborne gastroenteritis outbreaks, France, 1998–2006. *Journal of Water and Health* 2008; **6**: 491–503.
2. **Saint-Laurent D, Grémy I, Therre H.** Contribution of health administrative databases to epidemiology and surveillance: crossed views from France and Quebec. *Bulletin Epidemiologique Hebdomadaire* 2013; **19**: 1–58.
3. **Bounoure F, *et al.*** Syndromic surveillance of acute gastroenteritis based on drug consumption. *Epidemiology and Infection* 2011; **139**: 1388–1395.
4. **Tuppin P, *et al.*** French national health insurance information system and the permanent beneficiaries sample. *Revue Epidemiologique de Sante Publique* 2010; **58**: 286–290.
5. **Edge VL, *et al.*** Syndromic surveillance of gastrointestinal illness using pharmacy over-the-counter sales. A retrospective study of waterborne outbreaks in Saskatchewan and Ontario. *Canadian Journal of Public Health* 2004; **95**: 446–450.
6. **Andersson T, *et al.*** Syndromic surveillance for local outbreak detection and awareness: evaluating outbreak signals of acute gastroenteritis in telephone triage, web-based queries and over-the-counter pharmacy sales. *Epidemiology and Infection* 2014; **142**: 303–313.
7. **Proctor ME, Blair KA, Davis JP.** Surveillance data for waterborne illness detection: an assessment following a massive waterborne outbreak of *Cryptosporidium* infection. *Epidemiology and Infection* 1998; **120**: 43–54.
8. **Buckeridge DL.** Outbreak detection through automated surveillance: a review of the determinants of detection. *Journal of Biomedical Informatics* 2007; **40**: 370–379.
9. **Berger M, Shiau R, Weintraub JM.** Review of syndromic surveillance: implications for waterborne disease detection. *Journal Epidemiology and Community Health* 2006; **60**: 543–550.
10. **Daures M.** Investigation of waterborne gastroenteritis outbreak, Pérignat-lès-Sarliève, France, June-July 2010 [in French]. Saint-Maurice: Institut de veille sanitaire, 2011, 46 pp.
11. **Mouly D.** Outbreak of viral gastroenteritis due to contaminated drinking water in Pleaux, France, April 2012 [in French]. Saint-Maurice: Institut de veille sanitaire, 2013, 44 pp.
12. **Majowicz SE, *et al.*** A common, symptom-based case definition for gastroenteritis. *Epidemiology and Infection* 2008; **136**: 886–894.
13. **Mouly D, *et al.*** Authors' data, not previously published, available in institutional reports in references [10, 11]).
14. **Van Cauteren D, *et al.*** Burden of acute gastroenteritis and healthcare-seeking behaviour in France: a population-based study. *Epidemiology and Infection* 2012; **140**: 697–705.
15. **Goldberg M, Luce D.** Selection effects in epidemiological cohorts: nature, causes and consequences. *Revue d'Epidemiologique et de Santé Publique* 2001; **49**: 477–492.
16. **Hernan MA, Hernandez-Diaz S, Robins JM.** A structural approach to selection bias. *Epidemiology* 2004; **15**: 615–625.
17. **Sentinelles Network.** (https://websenti.u707.jussieu.fr/sentiweb/?lang=en). Accessed 10 November 2014.
18. **Beaudeau P, *et al.*** Automated detection of gastroenteritis waterborne outbreaks from the data of sale or drug reimbursement [in French]. Saint-Maurice: Institut de veille sanitaire, 2006, 40 pp.
19. **Wheeler JG, *et al.*** Study of infectious intestinal disease in England: rates in the community, presenting to general practice, and reported to national surveillance. The Infectious Intestinal Disease Study Executive. *British Medical Journal* 1999 **17**; 318: 1046–1050.
20. **Gallay A, *et al.*** A large multi-pathogen waterborne community outbreak linked to faecal contamination of a groundwater system, France, 2000. *Clinical Microbiology and Infection* 2006; **12**: 561–570.

2 Développement d'une méthode intégrée pour la détection automatisée des épidémies d'origine hydrique

Article 2 : Détection des épidémies d'origine hydrique : une approche intégrée utilisant les données de l'Assurance Maladie

Auteurs: Coly S. [1], Vincent N.[2], Vaissière E.[2], Charras-Garrido M. [1], Gallay A.[2], Ducrot C. [1] Mouly D.[2]

[1] Inra – Unité d'épidémiologie animale, Theix ; [2] Institut de veille sanitaire, Saint Maurice.

Journal of Water and Health: accepté le 12 juin 2016 (en cours d'édition)

Présentation synthétique du travail réalisé et des résultats de cet article

Aucune approche de détection d'épidémies de gastro-entérite aigüe actuellement développée en France n'est spécifique de la cause hydrique et ne cherche à la prendre en compte en amont de la recherche d'épidémies. Certaines régions comme l'Auvergne sont particulièrement concernées par cette problématique, comme en témoignent deux épidémies de gastro-entérite aigüe d'origine hydrique impliquant à chaque fois plusieurs centaines de cas survenues en 2010 et 2012 (cf. article 1).

L'objectif de ce travail est de développer une méthode intégrant l'origine hydrique en amont de la détection des épidémies de gastro-entérite aigüe à partir des données de l'Assurance Maladie. L'approche doit prendre en compte le niveau d'agrégation des données de l'Assurance Maladie (à la commune) et le niveau d'agrégation de l'exposition (l'unité de distribution d'eau). La contrainte est que ces deux zones spatiales ne sont pas systématiquement superposables (cf. paragraphe « Données d'exposition disponibles pour la prise en compte de l'origine hydrique dans la détection d'épidémies », page 43).

La démarche a été tout d'abord d'élaborer un algorithme décisionnel pour définir les regroupements de communes à opérer en fonction des unités de distribution d'eau qui les alimentent. Les données de population et de structure des unités de distribution présentes dans la base nationale SISE-Eaux ont été utilisées. Les critères de décision de l'algorithme étaient basés sur la proportion de la population desservie par une unité de distribution d'eau dans chaque commune. Les nouvelles unités géographiques pouvaient correspondre aussi bien à une seule commune (cas d'une commune ne partageant pas d'unité de distribution avec les communes voisines) qu'à un groupe de plusieurs communes (cas de communes partageant une même unité de distribution d'eau). Dans cette situation, de nouvelles coordonnées géographiques étaient attribuées au regroupement créé et les cas de gastro-entérite aigüe de chaque commune du regroupement étaient agrégés. La méthode de détection d'agrégats, le scan spatio-temporel de Kulldorff, a ensuite été appliquée sur les nouveaux

regroupements. Les signaux identifiés avaient en commun le fait de partager la même exposition à l'eau du robinet. La vraisemblance de l'origine hydrique a été confortée en comparant le profil des signaux détectés aux profils connus des épidémies d'origine hydrique ainsi qu'à la présence d'indicateurs environnementaux évocateurs d'une pollution d'eau. Les logiciels R et SaTScan ont été utilisés pour automatiser l'ensemble de la démarche. La méthode a été appliquée à la région Auvergne (analyse par département) sur la période 2009 – 2012 (analyse par année). Parmi les agrégats de gastro-entérite aigüe significatifs ($p<0,05$), ont été sélectionnés ceux dont la durée était compatible avec une épidémie hydrique (>6jours) et qui présentaient un excès de cas supérieur à 10 et qui correspondaient à un risque épidémique (rapport entre le nombre de cas observés et nombre de cas attendus ou RR > 3).
Sept cent quatorze nouvelles unités géographiques ont été créées en effectuant des regroupements parmi les 1 310 communes et 1 706 unités de distribution d'eau que contient la région Auvergne (Figure 5, article ci-après). En appliquant la méthode de détection sur ces nouvelles unités géographiques, 50 agrégats de cas de gastro-entérite aigüe partageant la même exposition à l'eau étaient significatifs ($p<0,05$), et onze correspondaient aux critères de sélection (durée supérieure à 6 jours et excès de cas supérieure à 10 et RR supérieur à 3). Pour dix d'entre eux, les investigations environnementales ont permis d'identifier au moins un évènement ayant pu conduire à une pollution microbiologique de l'eau distribuée dans les jours qui ont précédé les signaux épidémiques détectés (Table 2, article ci-après).

L'algorithme de regroupement de communes présente un intérêt particulier pour les unités de distribution d'eau desservant une ou plusieurs communes (c'est à dire 69,2% de la population dans la région Auvergne), car il crée une unité cohérente avec l'exposition à l'eau et augmente la puissance de détection en augmentant l'effectif de la population. En revanche, il ne permet pas d'améliorer la prise en compte de l'origine hydrique pour une commune desservie entièrement par plusieurs unités de distributions, non partagées avec des communes voisines (30,8% de la population pour la région Auvergne). En effet, les données de l'Assurance Maladie ne permettent pas de descendre à un niveau infra-communal. Cette proportion de la population peut varier d'un endroit à l'autre du territoire en fonction de l'adéquation entre le contour des unités de distribution et celui des communes. La (Figure 12 page 44) montre que des situations très contrastées peuvent apparaître en fonction des départements et que le relief influence fortement l'organisation de la distribution de l'eau en France.

Sur l'ensemble des signaux détectés dont l'origine hydrique a été confortée, seulement deux avaient été identifiés par le contrôle sanitaire au moment de leur survenue et avaient fait l'objet d'investigations par Santé publique France (ex-InVS) (les épidémies de 2010 et 2012 de l'article 1). Par ailleurs, la méthode a permis de détecter une épidémie en 2010 associée au même réseau d'eau qui était en cause dans l'épidémie de 2012 de l'article 1. Ce second signal conforte l'hypothèse de pollutions

chroniques évoquées par les participants lors de l'enquête de cohorte de l'épidémie de 2012. D'après ces résultats, la méthode de détection développée et testée sur la région Auvergne pourrait augmenter d'un facteur 5 la sensibilité du dispositif de surveillance de ces évènements. Il n'est en revanche pas possible d'estimer la spécificité de la méthode à partir de données réelles. Les résultats montrent également que pour dix agrégats identifiés sur onze, des arguments environnementaux en faveur d'une origine hydrique étaient présents. Enfin, les taux de médicalisation estimés pour les agrégats identifiés étaient en moyenne de 1,8% et s'étalaient de 0,7% à 4,8%. Ces taux sont cohérents avec ceux observés dans les épidémies décrites dans l'article 1 (respectivement 1,5 et 2%). Enfin, la mise en œuvre de cette méthode nécessite un contrôle préalable de la qualité des données de population de la base SISE-eaux.

Journal of Water and Health

DETECTION OF WATERBORNE DISEASE OUTBREAKS: AN INTEGRATED APPROACH USING HEALTH ADMINISTRATIVE DATABASES

--Manuscript Draft--

Manuscript Number:	JWH-D-15-00273
Full Title:	DETECTION OF WATERBORNE DISEASE OUTBREAKS: AN INTEGRATED APPROACH USING HEALTH ADMINISTRATIVE DATABASES
Article Type:	Research Paper
Corresponding Author:	Damien MOULY Institut de Veille Sanitaire FRANCE
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Institut de Veille Sanitaire
Corresponding Author's Secondary Institution:	
First Author:	Sylvain Coly
First Author Secondary Information:	
Order of Authors:	Sylvain Coly
	Nicolas Vincent
	Emmanuelle Vaissière
	Myriam Charras-Garrido
	Christian Ducrot
	Anne Gallay
	Damien MOULY
Order of Authors Secondary Information:	
Abstract:	Hundreds of cases of water-borne disease outbreaks of Acute Gastroenteritis (AGI) due to contaminated tap water are reported in France each year. Such outbreaks are probably under-detected. The aim of our study was to develop an integrated approach to detect and study clusters of AGI in geographical areas with homogeneous exposure to drinking water. Data for the number of AGI cases are available at the municipality level and exposure to tap water depends on drinking water networks (DWN). These two geographical units do not systematically overlap. This study proposed to develop an algorithm which would match the most relevant grouping of municipalities with a specific DWN, in order that tap water exposure can be taken into account when investigating future disease outbreaks. Space-time detection method was applied to the grouping of municipalities. Seven hundred and fourteen new geographical areas (groupings of municipalities) were obtained compared with the 1 310 municipalities and the 1 706 DWN. Eleven potential WBDO were identified in these groupings of municipalities. For ten of them, additional environmental investigations identified at least one event that could have caused microbiological contamination of DWN in the days previous to the occurrence of a reported WBDO.

DETECTION OF WATERBORNE DISEASE OUTBREAKS: AN INTEGRATED APPROACH USING HEALTH ADMINISTRATIVE DATABASES

Authors:

S. COLY[1], N. VINCENT[2], E. VAISSIERE[2], M. CHARRAS-GARRIDO[1], A. GALLAY[2], C. DUCROT[1], D. MOULY[2*].

Institutional addresses:

[1] INRA, UR346 - Unité d'Épidémiologie Animale, Centre de recherche de Clermont-Ferrand, 63122 Saint Genès Champanelle, France

[2] French Institute for Public Health Surveillance, 12 rue du Val d'Osne, 94 415 Saint-Maurice Cedex, France

***Author for correspondence:**

Mr D. MOULY

Address:

InVS-Dcar-Cire Midi-Pyrénées, 10 chemin du raisin 31050 Toulouse, Cedex 9.

Email: damien.mouly@ars.sante.fr; dmouly@gmail.com

Phone: +33534302518 - Cell: +33679601042

Fax: +33534302532

Short running head: Detection of waterborne disease outbreak: an integrated approach

Abstract

Hundreds of cases of water-borne disease outbreaks of Acute Gastroenteritis (AGI) due to contaminated tap water are reported in France each year. Such outbreaks are probably under-detected. The aim of our study was to develop an integrated approach to detect and study clusters of AGI in geographical areas with homogeneous exposure to drinking water. Data for the number of AGI cases are available at the municipality level and exposure to tap water depends on drinking water networks (DWN). These two geographical units do not systematically overlap. This study proposed to develop an algorithm which would match the most relevant grouping of municipalities with a specific DWN, in order that tap water exposure can be taken into account when investigating future disease outbreaks. Space-time detection method was applied to the grouping of municipalities. Seven hundred and fourteen new geographical areas (groupings of municipalities) were obtained compared with the 1 310 municipalities and the 1 706 DWN. Eleven potential WBDO were identified in these groupings of municipalities. For ten of them, additional environmental investigations identified at least one event that could have caused microbiological contamination of DWN in the days previous to the occurrence of a reported WBDO.

Keywords

Environmental studies; Health administrative data; Integrated approach; Public health surveillance ; Space-time detection; Waterborne disease outbreaks.

ABBREVIATIONS

AGI	Acute gastrointestinal infection
DWN	Drinking water network
GP	General Practitioner
InVS	Institut de veille sanitaire (French Institute for Public Health Surveillance)
SISE-eaux	Système d'Information en Santé-Environnement sur les Eaux d'alimentation
SNIIRAM	Système national d'information inter-régimes de l'Assurance maladie (French National Health Insurance Information System)
WBDO	Waterborne disease outbreak

Introduction

Waterborne disease outbreaks (WBDO) are a public health concern in France because of the large proportion of people potentially affected when contamination of drinking water occurs (Beaudeau 2010). To date, detection of these events has mainly occurred through general practitioners' (GPs) reporting of clusters of acute gastrointestinal infection (AGI) to health authorities. The absence of a designated surveillance system suggests therefore that the number of WBDO is probably underestimated in France. In public health terms, increasing the detection of infections caused by contaminated drinking water contributes to improving the following factors: knowledge of risk factors, identification of high risk drinking water networks, and development of appropriate preventive measures. In this context, the French Institute for Public Health Surveillance (InVS) is exploring the possibility of using the health administrative databases from the French Health Insurance to develop a national automated detection system for WBDO.

Searching for a link between a health indicator ,(e.g. a WBDO), and associated environmental exposure factors (e.g. drinking water consumption), is a frequent subject of study in the field of epidemiological surveillance (Chaput 2002; Klassen 2005).

To date, most related studies have considered health and environmental data separately (Mostashari 2003; Hayran 2004; Osei & Duker 2008), by first attempting to detect spatial or spatiotemporal areas in which clusters of cases occur, and mapping environmental exposure factors. Then the locations of case clusters and factors linked to the environmental area of exposure are compared (Fukuda 2005). Other studies in the literature have successfully considered both health concerns and environmental factors together. Most of these have taken a common approach (Patil & Taillie 2004) whereby a statistical method is first applied to detect the occurrence of a cluster of cases. Then a validation test of the detected clusters is performed, followed by identification *a posteriori* of the environmental factors related to each cluster. However, results of tests to validate and explain detected clusters have not been conclusive in most of these studies (D'Aignaux 2002; Odoi 2004).

In our study, within the framework of the detection of WBDO, the environmental exposure factor considered was the Drinking Water Network (DWN). By definition, a DWN delivers water of homogenous quality to consumers, i.e. any persons connected to the same DWN are similar from the

point of view of water quality. For this reason, DWN was considered as the environmental factor of interest for the detection of WBDO. Therefore the area of the DWN was considered as the spatial unit of interest to study the clusters of WBDO. The health data considered in the present study came from cases of AGI identified from the French National Health Insurance Information System (SNIIRAM: Système national d'information inter régimes de l'Assurance maladie) using a specific algorithm previously published (Bounoure 2011).

The aim of the study presented in this article was to develop an integrated approach to detect and study clusters of AGI in geographical areas with homogeneous exposure to drinking water, for which data on the human population and cases of AGI were both available.

This newly developed approach was tested on real AGI data in France. Further work will evaluate the capacity of detection of this approach using simulated data (forthcoming publication).

Methods

The integrated approach needs to handle the absence of systematic overlapping between exposure data area (DWN) and cases of AGI data area (municipality). This is a two-step approach as follows: first, the creation of new geographical units taking into account *a priori* drinking water exposure and aggregation of cases of AGI; second, the application of a space-time detection method of clusters of AGI in these geographical units by using a method previously published (Kulldorff 2005).

Three types of data were used: health data, geographical and population data, environmental data.

Health data and case definition

The health indicator used in our study for the detection of WBDO was medicalized cases of AGI.

In France, an algorithm was specifically developed to identify AGI cases by using data on reimbursement for payment of prescribed drugs from the SNIIRAM database (Bounoure 2011). The SNIIRAM aims at evaluating beneficiaries' healthcare consumption and associated expenditures. It covers more than 98% of the French population and records all reimbursements to patients for out-of-pocket medical procedures and medications and payments to professionals for consultations (Tuppin 2010). AGI medications are included in this database if they are reimbursable, prescribed by a GP and dispensed in a pharmacy. The identification of AGI cases required two consecutive steps: (i) data

extraction from the SNIIRAM database and (ii) using the AGI algorithm developed by Bounoure *et al.* (2011) to select AGI cases. The criterion for the first step was as follows: reimbursement for at least one prescribed target drug used to treat AGI[1]. The criteria for the AGI discriminative algorithm were as follows: a delay of <24 hours between the prescription and delivery of drugs, the number of different AGI-specific drugs prescribed, treatment duration (less than 8 days), and the co-prescription of non-AGI specific drugs (e.g. anti-cancer drugs). The sensibility and specificity of the AGI algorithm is 90% (Bounoure 2011). Data on age, gender, date of consultation, and municipality of residence were available for each case of AGI and cases were aggregated by municipality of residence.

The AGI database used for analysis contained the number of new cases for each day and for each municipality of residence.

Geographical and population data

The data describing the municipalities' coordinates and population numbers came from the national geographic institute (Institut National de l'information Géographique et forestière, 2013)

Environmental data

Environmental indicator for exposure used is DWN.

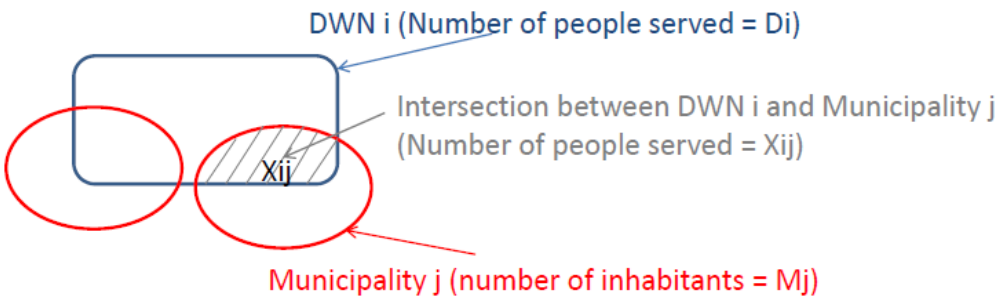
To take account of the environmental exposure factor (DWN), we used the French Health-Environment Information System on Water Delivery database (SISE-eaux: "Système d'Information en Santé-Environnement sur les Eaux d'alimentation") managed by the French Ministry of Health. This database contains the list of all DWN in France and the list of the municipalities served. For each DWN, technical information of installations is also available (e.g. water treatment plant, tank). Moreover, population data correspond to the number of people served by each DWN, the number of inhabitants in each municipality and the number of people served at the overlap of DWN and served municipalities.

Data extraction was performed by selecting the following variables: French county number, DWN code and name, zip codes and names of municipalities served by DWN, population numbers for DWN (Di in

[1] Antiemetics drugs - ATC classification: A04A, A03F; anti-diarrhea drugs - A07X, A07D; intestinal adsorbents drugs - A07B, A02X and oral rehydration salts

Figure 1), number of inhabitants of each municipality (Mj in Figure 1) and population number at the overlap of DWN and municipalities served (e.g. Xij in Figure 1).

Figure 1: population number for DWN, Municipalities and intersection DWN/Municipalities



Study area and period

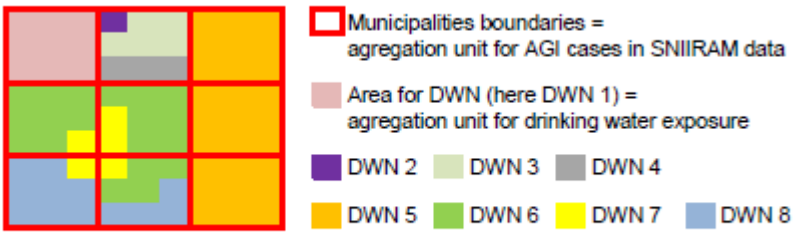
The study area was an administrative region in the center of France (Auvergne). This area was chosen because of the existence of environmental and sanitary signal associated with tap water: the occurrence of frequent microbiological contaminations of DWN and of WBDO (Mouly 2015).

Health data were collected for the period between January 1th 2009 and December 31th 2012.

Description of an algorithm used to define area with homogeneous tap water exposure in WBDO detection system

The geographic areas for aggregation of cases of AGI (municipality level) and for exposure factor (DWN) do not always overlap (Figure 1). There are four configurations to represent the correspondence between DWN geographical limits and municipality boundaries in France (Figure 2): (i) 1 municipality for 1 DWN (e.g DWN 1 in Figure 2). This is the perfect overlapping configuration. The aggregated unit of health data (cases of AGI) corresponds exactly to a single aggregated unit of tap water exposure (i.e. one single DWN); (ii) 1 municipality = n DWN indicates that the population of the municipality is served by different DWN so the population of a same municipality may be exposed to heterogeneous tap water quality (e.g. DWN 2-4); (iii) m municipalities = 1 DWN indicates that only one DWN serves several municipalities (e.g. DWN 5). The whole population of these municipalities drinks water of the same quality; (iv) m municipalities = n DWN is the most complicated configuration, because there is no direct relationship whatsoever between exposure by aggregation unit (i.e. DWN) and health data aggregation unit (i.e. municipalities) (e.g. DWN 6-8).

Figure 2 – Several possible configurations for overlapping between DWN area and Municipalities boundaries



Based on the different possible configurations between DWN and municipalities, the algorithm contained the following phases (Figure 3):

1. Listing the municipalities served by each DWN and evaluating the corresponding population size (step 1, Figure 3). For each DWN, all the municipalities partially or globally served were recorded. The following variables were specified:

- total number of people served by the DWN (Di, Figure 1) ;
- total number of people living in each municipality (Mj, Figure 1) ;
- total number of people served by the DWN in each municipality (Xij, Figure 1).

*2. Listing all the possible configurations between DWN and municipalities (step 2, Figure 3).*Two situations were distinguished:

1st situation: All the municipalities were served by only one DWN (configurations "1 municipality = 1 DWN" and "m municipalities = 1 DWN"). In this case it was supposed that all the municipalities were fully exposed when pollution occurs. Consequently, they were all systematically included in the possible configurations associated with this DWN.

2nd situation: Municipalities were partially served by a DWN (configurations "1 municipality = n DWN" and "m municipalities = n DWN"). In this case, including or excluding this municipality in the grouping which best matched the DWN had to be decided. If k is the number of municipalities partially served by the DWN, 2^{k-1} is the possible number of groupings of municipalities which must be considered for inclusion or not.

3. Computing indicators for each configuration (step 3, Figure 3).

For each DWN, two indicators were built for each possible grouping of municipalities:

- The "**exposure – municipalities ratio**". This is the ratio of the population size served by a DWNi in a grouping of municipalities' (ΣiXij) to the total population of all the concerned municipalities (Σ Mj). The higher the ratio, the greater the capability of the statistical methods employed to detect low-intensity WBDO (i.e. higher power of detection).
- The "**exposure - DWN ratio**". This is the ratio of the population size served by a DWNi in a grouping of municipalities (ΣiXij) to the total population served by the DWN (Di). The higher this ratio, the stronger the plausibility that a potential outbreak of AGI in this grouping is due to exposure to contaminated drinking water (likelihood of a WBDO).

4. Selection rule based on these indicators to determine the final configuration for each DWN (step 4, Figure 3).

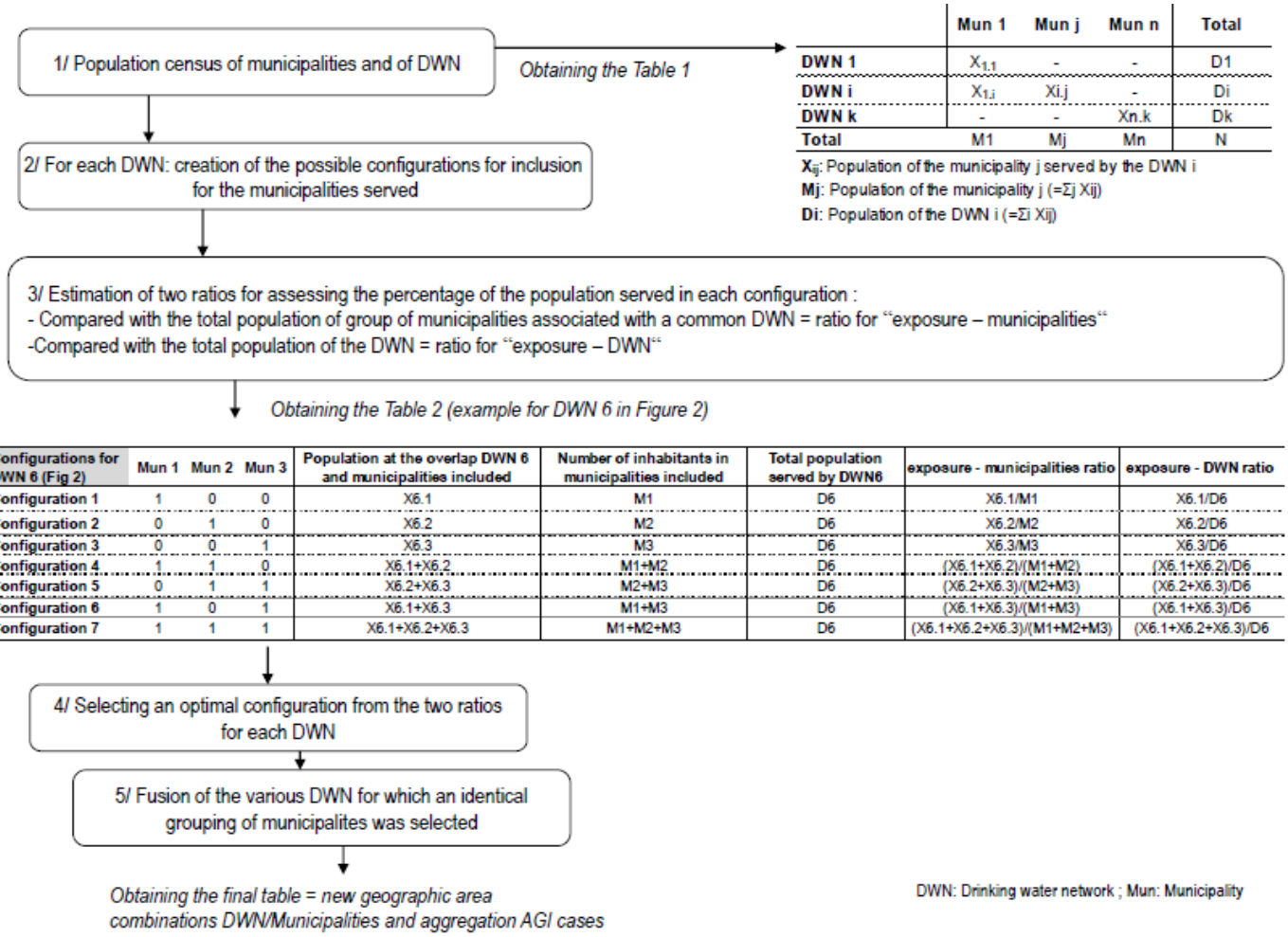
The selection of the optimized configuration (grouping of municipalities) for each DWN was made by minimizing the Euclidian distance between (exposure-municipalities ratio, exposure-DWN ratio) and (1, 1): $M = (1 - \text{exposure-municipalities ratio})^2 + (1 - \text{exposure-DWN ratio})^2$ *(equation 1)*

The grouping of municipalities chosen was the one associated with the smallest value of M.

5. Merging the DWN corresponding to the same grouping of municipalities (step 5, Figure 3). After the 4 previous steps, certain DWN corresponded to the same grouping of municipalities. We merged these to avoid any problems of repetition.

Cases of AGI were then aggregated over the new geographical area created by algorithm before the cluster detection process.

Figure 3 – Pattern of new geographical areas based on population size data in DWN and municipalities



	Mun 1	Mun j	Mun n	Total
DWN 1	$X_{1,1}$	-	-	D1
DWN i	$X_{1,i}$	Xi.j	-	Di
DWN k	-	-	Xn.k	Dk
Total	M1	Mj	Mn	N

X_{ij}: Population of the municipality j served by the DWN i
Mj: Population of the municipality j (=Σj Xij)
Di: Population of the DWN i (=Σi Xij)

Configurations for DWN 6 (Fig 2)	Mun 1	Mun 2	Mun 3	Population at the overlap DWN 6 and municipalities included	Number of inhabitants in municipalities included	Total population served by DWN6	exposure - municipalities ratio	exposure - DWN ratio
Configuration 1	1	0	0	X6.1	M1	D6	X6.1/M1	X6.1/D6
Configuration 2	0	1	0	X6.2	M2	D6	X6.2/M2	X6.2/D6
Configuration 3	0	0	1	X6.3	M3	D6	X6.3/M3	X6.3/D6
Configuration 4	1	1	0	X6.1+X6.2	M1+M2	D6	(X6.1+X6.2)/(M1+M2)	(X6.1+X6.2)/D6
Configuration 5	0	1	1	X6.2+X6.3	M2+M3	D6	(X6.2+X6.3)/(M2+M3)	(X6.2+X6.3)/D6
Configuration 6	1	0	1	X6.1+X6.3	M1+M3	D6	(X6.1+X6.3)/(M1+M3)	(X6.1+X6.3)/D6
Configuration 7	1	1	1	X6.1+X6.2+X6.3	M1+M2+M3	D6	(X6.1+X6.2+X6.3)/(M1+M2+M3)	(X6.1+X6.2+X6.3)/D6

DWN: Drinking water network ; Mun: Municipality

The algorithm was implemented using R software (versions 2.14 and 2.15).

Space-time detection of cluster of AGI

Several published methods for cluster detection are available in the literature (Mostashari 2003; Kulldorff 2005; Takahashi 2008; Assuncao & Correat 2009; Cucala 2009). We selected the space-time detection method, developed by Kulldorff (Kulldorff 2005), because of the consideration of seasonality during winter epidemics of AGI and the simplicity of the method's application with the SatScan software. This method has been widely used in the literature and would appear to be the reference method for cluster detection in epidemiological surveillance.

Space-time permutation of Kulldorff (Kulldorff 1997, 2005) allows areas with excess cases of AGI to be identified in terms of space and time. Applying the method to the algorithm-created geographical units (see previous section) consists in performing a scan of the whole study area, by moving a sliding window located successively at the central point of each geographic unit. Each window is compared with the outer window (which constitutes the entire geographic area under study). For space-time detection, the cylindrical window then travels in time and space so that all geographic units, sizes and durations are successively considered. This results in a great number of windows, and each is a candidate for an AGI cluster. A cluster is detected when the number of case of AGI within the window

is significantly higher than that outside this window. The test statistic is based on the likelihood ratio, i.e. the ratio of the likelihood calculated under the alternative hypothesis (the risk within the window is greater than that outside), and the calculated likelihood in the null hypothesis of equal risks. The window with the highest likelihood ratio defines the most likely cluster, i.e. the cluster least likely to occur by chance.

Covariates may be used in the detection process.

SatScan v9.3 software was used to implement the Kulldorff method- The following parameters were defined: time aggregation unit for AGI cases (day), aggregation duration for analysis (day), analysis type (retrospective space-time analysis with a space-time permutation model), and finally, type of inference (Monte Carlo inference with 999 replications).

The analysis was performed for the study area each year from 01/01/2009 to 31/12/2012.

Selection of clusters and validation of the waterborne origin

The clusters obtained were analyzed to select those whose characteristics most reflected the characteristics found for a given WBDO, according to epidemiological knowledge already available regarding that WBDO. Criteria for the selection of clusters were: duration of the signal for more than 6 days, size of the outbreak with more than 10 excess AGI cases , ratio between observed and expected cases of AGI higher than 3 and a p-value lower than 0.05.

Finally, the selected clusters were analyzed with the local health authorities to investigate whether specific environmental factors could have pointed to a microbiological contamination of the targeted DWN in the days before clusters appeared. These factors were as follows: results of sanitary control on fecal indicators (*Escherichia Coli* and fecal *streptococci*), heavy rains, an incident in the water treatment plant or in the DWN, cessation of disinfection. Furthermore, we checked for the existence of WBDO notification to authorities at the time of the occurrence.

Selected clusters were described using several epidemiological and environmental parameters. The former included the starting date and the duration of the period associated with the cluster, the number of observed and expected cases associated with the cluster, the observed/expected case ratio, the AGI case attack rate (estimated by the ratio between the number of observed cases and the size of population). Environmental parameters included checking for the existence of a microbiological

pollution during the cluster, the percentage of non-microbiological compliance with Ministry of Health fecal indicators during the study period (2009-2012), and the existence of other environmental risk factors.

Results

Description of configurations of inclusion of municipalities and drinking water network

The region of Auvergne contains 1 343 964 people living in 1 310 municipalities. The biggest municipality (regional capital, Clermont-Ferrand) contains 139 000 inhabitants. Fifteen municipalities have more than 10 000 inhabitants each, and 81% percent of municipalities have less than 1 000 inhabitants, accounting for 27% of the global regional population.

In Auvergne, 543 of the region's 1 706 DWN serve 20 people or fewer. Indeed, DWN serving 100 people or fewer account for the majority of DWN (62%), but serve only 2.5% of the whole population. Combined, the 10.6% of DWN which each serve more than 1 000 people serve 86.6% of the global population. Only four DWN in Auvergne each serve more than 30 000 people.

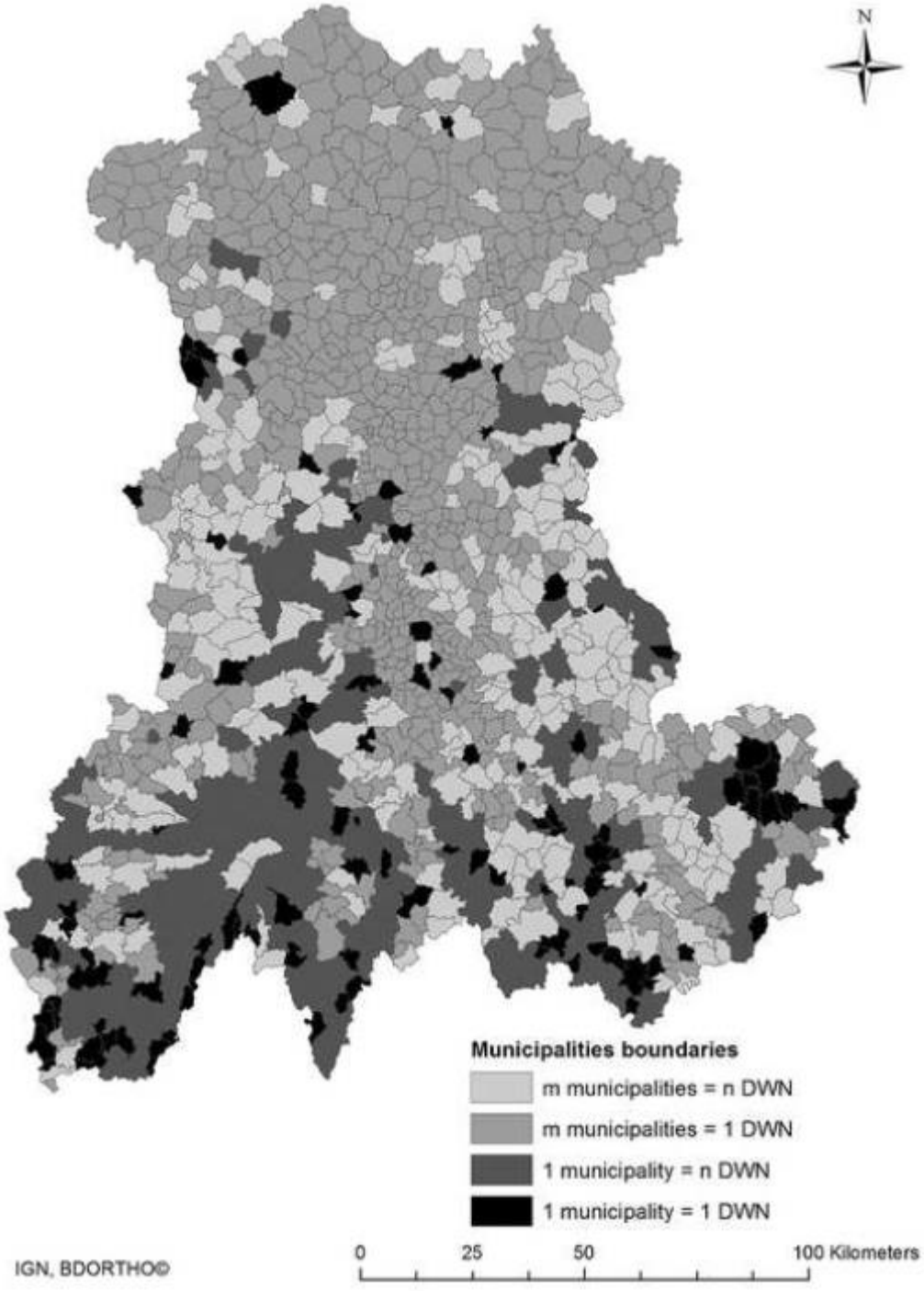
The four different matching configurations of DWN and municipalities for our study area are summarized in Table 1 and Figure 4

Table 1: Configurations of inclusion of municipalities and DWN in Auvergne, including number of corresponding municipalities and population size.

	Configurations of inclusion Municipalities/DWN				
	1 Municipality = 1 DWN	m Municipalities = 1 DWN	1 Municipality = n DWN	m Municipalities = n DWN	Total
Municipalities					
N	114	659	241	296	1,310
Percentage	**8.7%**	**50.3%**	**18.4%**	**22.6%**	**100.0%**
Population					
N (inhabitants)	151 447	524 630	293 074	374 813	1,343,964
Percentage	**11.3%**	**39.0%**	**21.8%**	**27.9%**	**100.0%**

DWN: Drinking Water Network;

Figure 4: Map of the configurations of inclusion of municipalities and DWN in Auvergne.

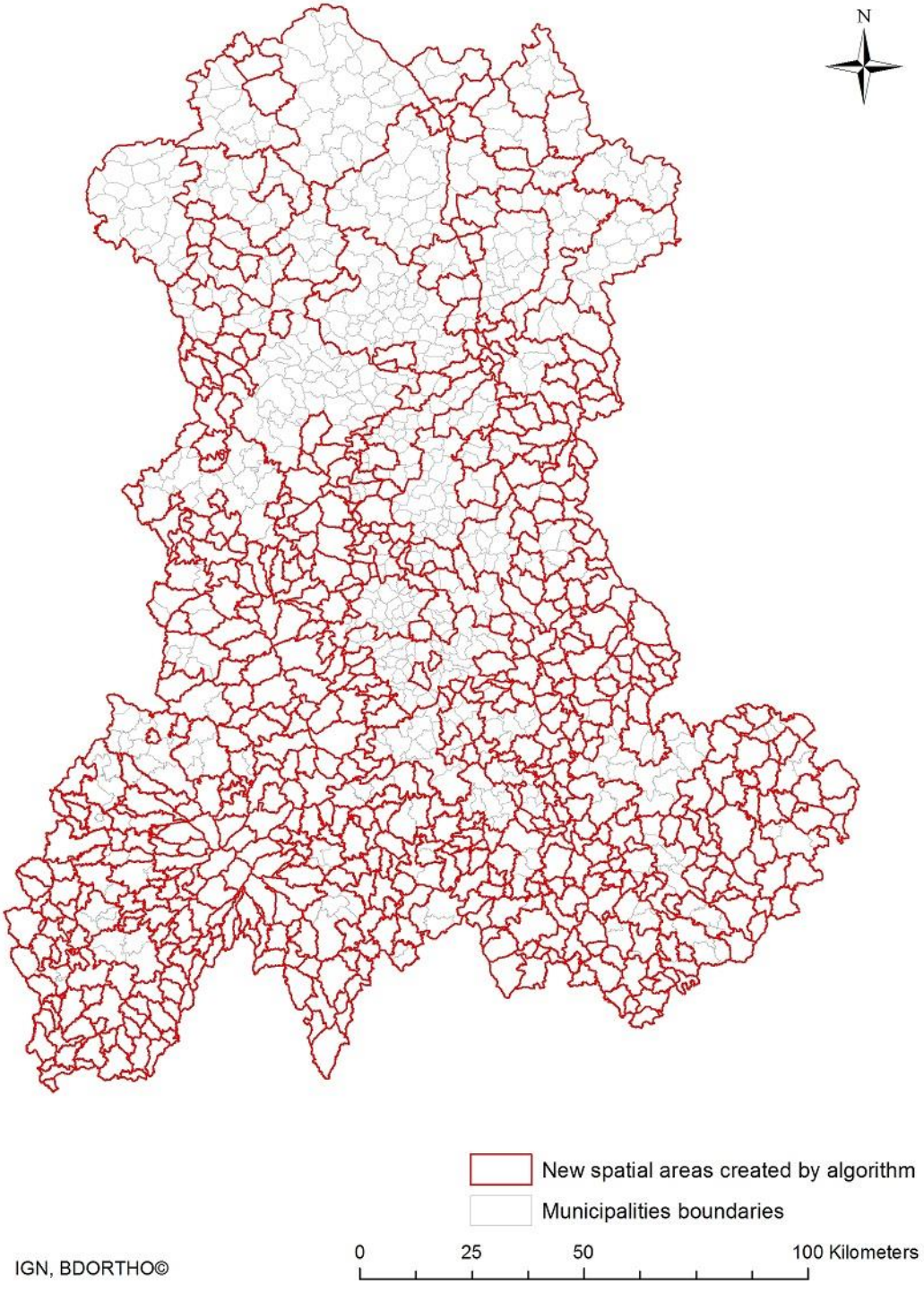


Source: Sise-Eaux, Ministère chargé de la santé ; DWN : Drinking Water Network

Description of the new areas obtained by the algorithm

After applying the algorithm, 714 new geographical areas (Figure 5) were created which grouped together the 1 310 municipalities and 1 706 DWN in Auvergne. The average population size of these areas was 1 891 people. .

Figure 5 –Spatial delimitation of the new geographical area.



Most of the new areas contained only one municipality (n=573, 80%). However, 12% were associated with at least 3 municipalities, accounting for 53% of all the municipalities. These areas were much more concentrated in lowland areas (topographic data not presented). Only 3% of the new areas contained at least 10 DWN. No new geographical area simultaneously included more than one DWN and more than one municipality.

Finally, all municipalities were included at least once in the composition of the resulting areas. Approximately 91% of the municipalities were associated with only one new area, 9% with at least two areas. Only one municipality was included in 4 new areas.

Description of clusters

Among all the detected clusters (50 clusters with $p<0.05$), 11 were consistent with possible WBDO according to the selection criteria above (Table 2). The impacted grouping of municipalities defined by the algorithm numbered between 500 and 5000 inhabitants each. Between 20 and 60 cases of AGI were involved in each cluster. The medication rate in the impacted population, was approximately 1.4% (median) and varied between 0.7% and 4.8%. The total duration of all the clusters was 177 days of cumulated WBDO, and the longest cluster was 35 days. For 2 of the 11 selected clusters, fecal pollution of DWN during the outbreak (Clusters ID 4 and 8 in Table 2) was detected. For the same 2 clusters, a notification of WBDO had been made to the local health authority. Moreover, one geographic area (Group ID 207 in Table 2) was concerned by 2 clusters respectively in 2010 (cluster 5) and in 2012 (cluster 8). The sanitary control of fecal indicators in drinking water highlighted the

occurrence of several non-microbiological compliance between 2009 and 2012 for 9 of the 11 clusters detected (mean=5.6%; min=0%; max=14.3%).

Environmental risk factors of pollution of DWN were identified for all selected clusters when the information was available (64% of clusters). These included heavy rains in the days before the start of the outbreak (clusters 2,3,4,6, 8, 9), the flooding of the drinking-water borehole causing a cessation of chlorination (clusters 4 and 8) and water pipes' breakage in a DWN (cluster 1).

Finally, except for cluster 7, at least one environmental factor or water treatment/distribution incident was associated with clusters selected as WBDO.

Table 2 – Description of the 11 clusters of AGI most probably related to contamination of drinking water network - Auvergne region, 2009 - 2012. Clusters presented in the table were selected with the following criteria: cluster duration < 7 days, excess cases > 10, ratio observed/expected cases of AGI > 3, p-value < 0.05.

Cluster ID	Year	Area ID	Number of Municipalities	Number of DWN	Population served (inhabitants)	Start date	Duration (days)	Observed cases of AGI	Expected cases of AGI	Obs/ Exp	Medication rate in population[ϕ]	Microbiological pollution during cluster	% of non-microbiological compliance[£]	Other environmental factors	Notification of WBDO to the health authority	p-value
1	2009	707	1	2	4910	26/11/2009	7	67	13,9	4,8	1,4%	No	0,0%	YES	NO	1,0E-17
5	2010	385	1	3	1563	19/03/2010	19	33	9,59	3,4	2,1%	No	12,2%	YES	NO	6,0E-04
4	2010	155	1	1	501	31/03/2010	12	24	3,25	7,4	4,8%	No	7,1%	YES	NO	2,4E-08
2	2010	638	1	3	2650	21/06/2010	7	42	4,8	8,8	1,6%	Yes*	1,5%	YES	YES	1,0E-17
6	2010	207	1	4	1549	16/08/2010	35	21	5,29	4,0	1,4%	No	8,5%	NA	NO	4,8E-02
3	2010	88	12	1	5500	09/09/2010	20	72	21,88	3,3	1,3%	No	14,3%	YES	NO	4,0E-12
9	2012	31	8	1	4752	15/02/2012	8	34	9,76	3,5	0,7%	No	0,0%	NA	NO	2,3E-04
10	2012	207	1	4	1549	23/03/2012	28	31	9,48	3,3	2,0%	Yes[¥]	8,5%	YES	YES	5,2E-03
11	2012	452	1	4	2411	27/03/2012	15	23	6,16	3,7	1,0%	No	5,7%	YES	NO	3,0E-02
7	2012	53	6	1	1933	03/12/2012	12	44	8,62	5,1	2,3%	No	2,1%	NA	NO	1,5E-12
8	2012	673	1	4	3628	03/12/2012	14	48	13,25	3,6	1,3%	No	1,2%	NA	NO	2,3E-08
Total					**30946**		**177**	**439**								

* 900 E.coli UFC/100 mL - 21/06/10

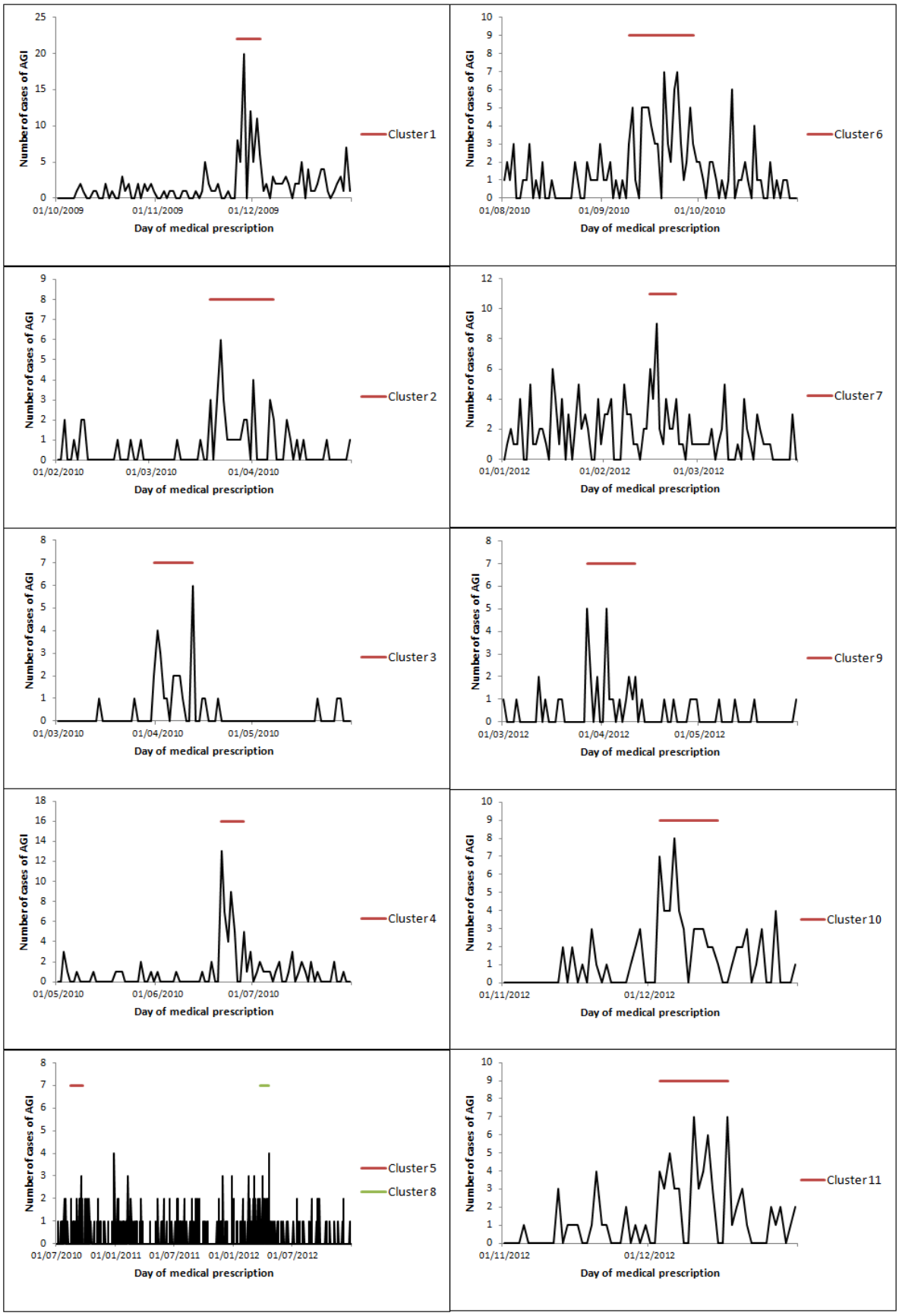
¥>100 E.coli UFC/100 mL - 10/04/12

£ percentage of analysis > 1 E.coli and/or fecal streptococci for the period 2009-2012

ϕ The medication rate was estimated for the total population of municipalities impacted

NA: not available

Figure 6 – Description of the number of cases of AGI according to the day of medical prescription and selected clusters



Discussion

Several factors can explain the occurrence of clusters of AGI cases and their increased incidence. The most commonly documented factors are the ingestion of contaminated food (foodborne disease), person-to-person transmission (in particular in children and older populations), and WBDO. The integrated and approach developed in this article for the detection of WBDO using health administrative databases identified 11 AGI clusters in the Auvergne region between 2009 and 2012 where a link with the consumption of contaminated tap water was likely. However, although the integrated approach optimizes the reliability of this link (by taking into account, exposure to water tap prior to the detection), all identified clusters have to be analyzed and investigated individually to increase the accuracy of determination.

Validation of selected clusters as WBDO

Several criteria (statistical, epidemiological and environmental) pointed to the existence of WBDO for the 11 selected clusters:

First, cases of AGI which shared the same DWN and therefore had homogenous drinking water quality were aggregated into newly created geographic areas (by the algorithm) before the application of space-time detection method. This ecological approach helped highlight any link between health signal (cluster of AGI) and exposure factor (DWN) as much as possible. Nevertheless, as seen in Table 1, 21.8% of the study population lives in municipalities served by more than 1 DWN (configuration 1 municipality = n DWN). For this configuration, the unit of aggregation of cases of AGI is the municipality. Health data do not enable us to geo-localize cases of AGI at an infra-municipality level. For the 7 clusters selected where one municipality was served by more than one DWN, additional investigation is needed to identify the impacted DWN. This would include checking for incidents in water treatment processes and in the distribution networks.

For epidemiological evidence, we used results from past investigations of WBDO (Beaudeau 2008) in impacted populations to identify several criteria for selecting clusters as follows: they usually last 1 to 3 weeks (clusters over 6 days were selected here), at least a few dozen cases are involved (clusters

with more than 10 cases of AGI were selected here), the relative risk presented is greater than 3 (the same value was used here). Finally, a p-value<0.05 was also chosen. Moreover, a recent comparative study for the description of two WBDO by using two data sources (cohort study and health administrative database) highlighted a low medication rate in the population (1.5% and 2% respectively for both WBDO) (Mouly 2015). The medication rate observed for selected clusters in the present study was between 0.7% and 4.8%) (Table 2). Moreover, the application of epidemiological criteria enable us to exclude other origins of localized outbreak of AGI, for example foodborne origins, usually characterized by an outbreak duration between 1 and 7 days, and by fewer than 10 cases most of the time.

In addition to these epidemiological criteria, we looked for environmental factors for each selected cluster. The occurrence of WBDO is often associated with rains (Beaudeau 2010), particularly in rural areas with small DWN exposed to fecal discharges from livestock farms. Furthermore, the boreholes of small DWN are often poorly protected compared with DWN in urban areas. For example in cluster 3, a hydrological report indicated that heavy rains fell September 7th 2010, i.e. two days before the start of detected cluster. For clusters 2 and 10, two WBDO were investigated and described in detail in the literature (Mouly 2015).

The combination of the integrated approach which takes into account exposure to DWN before the detection of clusters of AGI and the application of selection criteria of cluster detected based on epidemiological knowledge allow us to improve the overall specificity of our detection method. Moreover, the occurrence of several clusters of AGI at different times, focused on the same DWN (e.g. 2 clusters for area 207 in 2010 and 2012, Table 2), provides strong evidence of a WBDO.

Finally, additional environmental investigation for DWN associated with selected clusters will be necessary to identify the circumstances and the origin of the contamination of tap water.

Benefits and limits of the algorithm in the context of an integrated detection system for WBDO

Increased likelihood of WBDO detection and additional investigations required

Health data were available for municipalities. Drinking water exposure data depended on the individual DWN. We created an algorithm to take into account exposure to drinking water to use in tandem with an existing method for detecting outbreaks of AGI. AGI clusters detected by the combined system

have good specificity with respect to the individual water supply. The links between clusters and drinking water must still be confirmed by environmental investigations (rainfall) and the search for possible incidents in drinking water processes and distribution networks on the date of AGI clusters.

Because of the definition of a DWN - whereby homogenous water quality is assumed-, if pollution is introduced somewhere into the network, it spreads throughout the DWN concerned. However, such a hypothesis does not consider the state of individual pipes, differences in flow rates, and stagnant water which occurs when water is not drawn for a long period of time. A potential bias may also occur when the network is contaminated by waste water reflux. All of these situations imply a great deal of heterogeneity in the water quality, depending on the position of the water-treatment plant.

The advantage of using the algorithm-based method is clear when several municipalities are served by a single DWN. In our study, all the AGI cases occurring in the same DWN were considered together. The corresponding configurations, i.e. "m municipalities = 1 DWN" and "m municipalities = n DWN", concerned 49.7% of the population and 72.9% of the municipalities (Table 1). On the contrary, the creation of the new geographical areas did not help to determine which DWN was involved when a municipality served by several DWN was concerned by an AGI outbreak. In our study 41.1% of the municipalities in the new geographical areas were associated with two or more DWN. Nevertheless the exposure-municipalities and exposure-DWN ratios provided information which helped us focus further investigations on a specific DWN and to confirm the hydric origin. For the configuration "1 municipality = n DWN", in the case of a disease outbreak, it is not possible to identify the DWN responsible, as AGI cases are counted in a municipality which is bigger than the corresponding DWN. Nevertheless, merging the DWN associated with the same municipality (or grouping of municipalities) helps decrease the number of occurrences of the municipalities in the dataset.

Improving power of detection

The algorithm created 714 new geographical areas. This number is much lower than the number of municipalities and DWN (respectively, 1 310 and 1 706), which implies a shorter computation time of spatiotemporal outbreak detections, because fewer geographical units need to be tested. Moreover, the average population size of the new geographical units was much greater (1 891 inhabitants) than for municipalities (1 031 inhabitants) and DWN (791 inhabitants). In turn, this implies improved power of detection of clusters using the algorithm over the standard approach.

Algorithm characteristics

Over the course of developing the algorithm, several methods seemed suitable. First, we considered that maximizing the number of potential AGI cases (to increase the power of detection) was as important as taking into account the corresponding population's exposure to tap water. So we gave the same importance to the exposure-municipalities and exposure-DWN ratios. Second, we chose to minimize the Euclidian distance between combined ratios (1,1) and (exposure-municipalities ratio, exposure-DWN ratio) to select the grouping of municipalities which best matched the specific DWN. These two choices did not greatly influence the results. We then performed sensitivity tests (not presented) which showed that the selected methods minimized the repetition of municipalities.

The set of new geographic areas constituted a territory whose characteristics (population size, global incidence and number of AGI cases) were very close to those of the Auvergne region. Accordingly, any repetition of municipalities had an insignificant impact in the incidence evaluation.

Conditions to apply the algorithm

The study area is characterized by a particularly hilly landscape. The relationship between DWN and municipalities is very complicated, and most of the region is rural. Accordingly, one can suppose that the algorithm can be used in other less topographically-complex territories as part of an integrated approach for the detection of WBDO.

The health and environmental data used in the algorithm are available for all French regions, so the integrated approach developed here for the detection of WBDO can be applied to other regions in France.

SISE-Eaux database quality

The SISE-Eaux database is maintained at a regional and departmental level. The reliability of this database is essential to obtain accurate matching of drinking water exposure and AGI cases. These data are very reliable in the area studied (Auvergne), in particular the population size counted at the overlap of municipalities and the DWN, an element which is of crucial importance when applying our algorithm.

Space-time detection method and setting

The space-time detection method used to detect clusters of AGI sharing the same DWN (Kulldorff 2010) was selected both because of the consideration of seasonality and the simplicity of its application with SatScan software. With respect to the former, selected clusters after analysis were as numerous during the winter season (5/11 clusters between January and March) as the rest of the year. While high incidence of AGI is common in European countries during winter, the space-time detection method does not appear to have been influenced by this phenomenon. For time aggregation, we used "days" whereas most of retrospective studies use "weeks" or "months" (Dematteï 2006). This decision was based on the high incidence of AGI compared with other infectious or chronic diseases. Day-based aggregation time ensures day time precision for detected cluster duration.

Implication for waterborne disease detection

The challenge of WBDO detection addressed in published studies (Edge 2004; Berger 2006; Andersson 2014) highlights the difficulty to detect short outbreaks involving fewer than 100 cases. For this purpose, information collected for cases has to have sufficient temporal (ideally at the day level) and spatial resolution (municipality level may be sufficient) to enable the detection of local outbreak signals like WBDO. Unlike our study, each cluster in published studies selected count for less than 100 cases of AGI.

In addition, syndromic surveillance is useful to estimate the size, duration and health impact of detected outbreaks, as it provides us with the consultation rate in the impacted population. Any such estimation should take into account influencing factors on consultation rate, in particular age and access to health services, as shown in our study, and described elsewhere (Mouly 2015).

From a public health point of view, detected epidemic signals from SNIIRAM data should be followed by implementing a set of operational measures, including field investigation. These should be conducted to validate and describe the outbreak, and to understand the origin and mechanisms involved in case diffusion. In turn this information can inform decision-making for public health prevention.

Conclusion

We implemented an algorithm to create new geographical areas which matched health data and environmental exposure levels (i.e. in drinking water), despite complicated associations between

municipalities and drinking water networks. The 714 new geographical areas/units accounted for all the drinking water networks and municipalities in the Auvergne region. The new geographic areas were bigger than DWN and municipalities, both in terms of surface and population sizes. Creating these areas resulted in greater power of detection of potential future outbreaks. The application of a space-time detection method on the new geographical areas for the Auvergne region between 2009 and 2012 identified 11 potential WBDO.

Accordingly, the relevance of this approach needs to be strengthened by analyzing other datasets (as described in this article) and by evaluating by simulation approach (forthcoming publication).

Acknowledgements

The authors thank the National Health Insurance for access to health administrative data. We also thank Pascal Beaudeau, Henriette De Valk and Yann Le Strat from the InVS for their advice and recommendations; and Magali Corso for the preparation of case data of acute gastro-enteritis from health administrative databases.

REFERENCES

Andersson, T., Bjelkmar, P., Hulth, A., Lindh, J., Stenmark, S. & Widerstrom, M. 2014 Syndromic surveillance for local outbreak detection and awareness: evaluating outbreak signals of acute gastroenteritis in telephone triage, web-based queries and over-the-counter pharmacy sales. *Epidemiol.Infect.* **142**, 303-313.

Assuncao, R. & Correat, T. 2009 Surveillance to detect emerging space-time clusters. *Comput Stat Data Anal.* 53,

Beaudeau, P., De Valk, H., Vaillant, V., Mannschott, C., Tillier, C., Mouly, D. & Ledrans, M. 2008 Lessons learned from ten investigations of waterborne gastroenteritis outbreaks, France, 1998-2006. *J.Water Health.* 6, 491-503.

Beaudeau, P., Valdes, D., Mouly, D., Stempfelet, M. & Seux, R. 2010 Natural and technical factors in faecal contamination incidents of drinking water in small distribution networks, France, 2003-2004: a geographical study. *J.Water Health.* 8, 20-34.

Berger, M., Shiau, R. & Weintraub, J. M. 2006 Review of syndromic surveillance: implications for waterborne disease detection. *J.Epidemiol.Community Health.* 60, 543-550.

Bounoure, F., Beaudeau, P., Mouly, D., Skiba, M. & Lahiani-Skiba, M. 2011 Syndromic surveillance of acute gastroenteritis based on drug consumption. *Epidemiol.Infect.* **139**, 1388-1395.

Chaput, E. K., Meek, J. I. & Heimer, R. 2002 Spatial analysis of human granulocytic ehrlichiosis near Lyme, Connecticut. *Emerg Infect Dis.* 8, 943-948.

Cucala, L. 2009 A flexible spatial scan test for case event data. *Comput Stat Data Anal.* 53, 2843.

D'Aignaux, J. H., Cousens, S. N., Delasnerie-Laupretre, N., Brandel, J. P., Salomon, D., Laplanche, J. L., Hauw, J. J. & Alperovitch, A. 2002 Analysis of the geographical distribution of sporadic Creutzfeldt-Jakob disease in France between 1992 and 1998. *Int J Epidemiol.* **31**, 490-495.

Edge, V. L., Pollari, F., Lim, G., Aramini, J., Sockett, P., Martin, S. W., Wilson, J. & Ellis, A. 2004 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.* 95, 446-450.

Fukuda, Y., Umezaki, M., Nakamura, K. & Takano, T. 2005 Variations in societal characteristics of spatial disease clusters: examples of colon, lung and breast cancer in Japan. *Int J Health Geogr.* **4**, 16.

Hayran, M. 2004 Analyzing factors associated with cancer occurrence: A geographical systems approach. *Turkish Journal of Cancer.* 36, 4.

Klassen, A. C., Kulldorff, M. & Curriero, F. 2005 Geographical clustering of prostate cancer grade and stage at diagnosis, before and after adjustment for risk factors. *Int J Health Geogr.* **4**, 1.

Kulldorff, M. 2010 StaScan User Guide for version 9.0. 110.

Kulldorff, M., Feuer, E. J., Miller, B. A. & Freedman, L. S. 1997 Breast cancer clusters in the northeast United States: a geographic analysis. *Am J Epidemiol.* **146**, 161-170.

Kulldorff, M., Heffernan, R., Hartman, J., Assuncao, R. & Mostashari, F. 2005 A space-time permutation scan statistic for disease outbreak detection. *PLoS.Med.* **2**, e59.

Mostashari, F., Kulldorff, M., Hartman, J. J., Miller, J. R. & Kulasekera, V. 2003 Dead bird clusters as an early warning system for West Nile virus activity. *Emerg Infect Dis.* **9**, 641-646.

Mouly, D., Van Cauteren, D., Vincent, N., Vaissiere, E., Beaudeau, P., Ducrot, C. & Gallay, A. 2015 Description of two waterborne disease outbreaks in France: a comparative study with data from cohort studies and from health administrative databases. *Epidemiology and Infection.*

Odoi, A., Martin, S. W., Michel, P., Middleton, D., Holt, J. & Wilson, J. 2004 Investigation of clusters of giardiasis using GIS and a spatial scan statistic. *Int J Health Geogr.* **3**, 11.

Osei, F. B. & Duker, A. A. 2008 Spatial dependency of V. cholera prevalence on open space refuse dumps in Kumasi, Ghana: a spatial statistical modelling. *Int J Health Geogr.* **7**, 62.

Patil, G. P. & Taillie, C. 2004 Upper level set scan statistic for detecting arbitrarily shaped hotspots'. *Environ Ecol Stat.* **11**, 15.

Takahashi, K., Kulldorff, M., Tango, T. & Yih, K. 2008 A flexibly shaped space-time scan statistic for disease outbreak detection and monitoring. *Int J Health Geogr.* **7**, 14.

Tuppin, P., De Roquefeuil, L., Weill, A., Ricordeau, P. & Merliere, Y. 2010 French national health insurance information system and the permanent beneficiaries sample. *Rev.Epidemiol.Sante Publique.* **58**, 286-290.

3 Etude des performances et des facteurs influençant la détection des épidémies d'origine hydrique

Article 3 : Détection des épidémies d'origine hydrique à partir des données de l'Assurance Maladie : une étude de simulation

Auteurs: D Mouly[1], S Goria[1], M Mounié[2], L Rambaud[1], P Beaudeau[1], A Gallay[1], C Ducrot[3], Y Le Strat[1].

[1] : Institut de veille sanitaire, Saint Maurice ; [2] Inserm-UMR 1027, Toulouse ; [3] Inra – Unité d'épidémiologie animale, Theix

Soumis à PlosOne: le 4 septembre 2016 (accepté le 25 octobre 2016, en cours de révision)

Présentation synthétique du travail réalisé et des résultats de cet article

Dans cet article, il s'agissait de déterminer plus précisément les performances de la méthode de détection présentée dans l'article 2 en s'appuyant sur une étude de simulation. Les objectifs secondaires étaient d'identifier les déterminants de la détection et de quantifier leurs poids respectifs.

L'étude de simulation s'est inspirée de la méthode de Noufaily. Les données de l'Assurance Maladie ont été utilisées comme données de référence pour la simulation de la ligne de base de la gastro-entérite aigüe. Deux départements ont été utilisés comme support pour l'analyse : le Puy-de-Dôme qui était inclut dans l'article 2 et l'Isère, un département de la région voisine avec un grand nombre d'unités de distribution d'eau et ayant connu plusieurs épisodes de gastro-entérite aigüe d'origine hydrique. Le processus de simulation a consisté dans un premier temps à générer l'incidence de base de la maladie (bruit de fond). La moyenne journalière des cas a été estimée à un niveau départemental à partir des données réelles (Assurance Maladie) puis répartie entre les communes. Pour la simulation du bruit de fond (un bruit de fond simulé pour chaque épidémie), une régression de Poisson, associée à une fonction spline, a été utilisée pour estimer la moyenne journalière (nombre de cas attendus). Une distribution binomiale négative a ensuite permis de générer aléatoirement les cas à la commune (Figure 2, article ci-après). Dans un deuxième temps, 2 000 épidémies dont le profil était compatible avec des épidémies d'origine hydrique, ont été générées aléatoirement à l'aide d'une distribution log-normale. Les contraintes mises sur les paramètres des épidémies simulées étaient les suivantes : la durée de l'épidémie entre 3 et 28 jours, et la variation du taux d'incidence entre les cas épidémiques et le niveau moyen de base des cas de gastro-entérite aigüe – ou taux d'incidence des cas épidémiques entre 0,5% et 6% en cohérence avec les 2 articles précédents. Afin de reproduire le niveau d'agrégation initial des données de l'Assurance Maladie et de tester l'algorithme de regroupement des communes (article 2), les cas épidémiques ont été ajoutés au bruit de fond de la gastro-entérite aigüe en les répartissant dans

une commune ou un groupement de commune en fonction de l'unité de distribution d'eau tirée au sort. Ensuite, la méthode de détection tenant compte du contour des unités de distribution d'eau (article 2) a été appliquée aux données simulées afin de rechercher des agrégats spatio-temporels pouvant correspondre aux épidémies simulées.

Sur les 2 000 épidémies simulées, 1 460 ont été détectées par la méthode (Table 1, article ci-après). Les résultats montrent que près de 9 signaux détectés sur 10 correspondent à une épidémie injectée (valeur prédictive positive = 90,5%). La valeur prédictive positive est légèrement plus faible pour les petites épidémies (inférieures à 10 cas) et pour les épidémies d'origine hydrique injectées pendant l'hiver en raison d'un bruit de fond de la gastro-entérite aigüe plus important à cette période de l'année. La sensibilité globale de la méthode est proche de 74%. Elle varie surtout en fonction de la taille de l'épidémie (inférieure à 20% pour des épidémies impliquant moins de 10 cas et supérieure à 95% pour des épidémies de 20 cas ou plus) (Table 2, article ci-après). En dehors de la taille de l'épidémie, la saison (plus sensible de 5 à 10% en dehors de la période hivernale) et le nombre de communes desservies par une unité de distribution d'eau (plus sensible de 5 à 10% pour une unité de distribution desservant 2 communes ou plus) influencent également la détection dans une moindre mesure.

L'influence du nombre de cas sur la détection dépend de la variation du taux d'incidence entre les cas épidémiques et les cas attendus de gastro-entérite aigüe (c'est-à-dire de l'ampleur de l'épidémie). Ainsi, pour les épidémies de faible ampleur, c'est-à-dire celles dont la variation du taux d'incidence est la plus faible (< 2% de variation du taux d'incidence), la capacité de détection augmente fortement lorsqu'on passe de moins de 10 cas à plus de 10 cas (jusqu'à un facteur 13 pour la classe plus de 50 cas versus la classe moins de 10 cas) (Table 3, article ci-après). Ce gain de sensibilité en fonction de la taille de l'épidémie s'estompe pour des épidémies avec une variation du taux d'incidence de 2% ou plus. Cette valeur de taux d'incidence de 2% correspond à celle d'une des épidémies décrites dans l'article 1.

Enfin, une analyse continue de la relation entre la capacité de détection et la taille de l'épidémie montre que le seuil en dessous duquel la capacité de détection est très faible (inférieure à 20%), serait de 5 cas épidémiques.

Les résultats de cette étude suggèrent une bonne capacité de détection, quelle que soit la période de l'année et le nombre de communes desservies pour les épidémies impliquant au moins 20 cas (supérieure à 95%). La capacité est très faible en dessous de 5 cas et intermédiaire pour les épidémies entre 5 et 20 cas. Pour ces épidémies, l'influence de la saison est plus importante (plus grande sensibilité en dehors de l'épidémie hivernale) ainsi que celle du nombre de communes par unité de

distribution d'eau. Enfin, la probabilité qu'une épidémie détectée corresponde à une épidémie d'origine hydrique est de 90% à partir de 10 cas.

A simulation-based study for evaluating a waterborne disease outbreak detection algorithm

Authors

Damien Mouly[1*], Sarah Goria[1], Michael Mounié[2], Pascal Beaudeau[1], Anne Gallay[1], Christian Ducrot[3], Yann Le Strat[1].

[1] Santé publique France, the French national public health agency, Saint-Maurice, France

[2] Université Paul Sabatier, Toulouse, France

[3] Institut national de la recherche agronomique, UR346 - Unité d'Épidémiologie Animale, Saint Genès Champanelle, France

*Corresponding author:

Email: damien.mouly@santepubliquefrance.fr

Abstract

Waterborne disease outbreaks (WBDO) are a public health concern in developed countries because of the high proportion of people affected when drinking water is contaminated. Because of the absence of a nationwide specific surveillance system the number of WBDO is most likely underestimated. In this context, an algorithm to detect WBDO relies on a space-time statistical method(1)(1) was previously developed.

The objective of our simulation-based study was to evaluate the performance of this algorithm for WBDO detection using data from health insurance data. A secondary objective was to estimate the factors which most influence WBDO detection.

We first simulated the daily baseline counts of acute gastrointestinal infections (AGI) by using a negative binomial distribution. A variety of 2000 simulated WBDO signals, according to a log-normal distribution, were then superimposed on the baseline data. To evaluate the performance of the WBDO detection method, the sensitivity (Se) and the positive predictive value (PPV) were both used. Multivariable Poisson regression was performed to identify the factors associated with WBDO detection and to estimate the strength of these associations.

Almost three quarters of the simulated WBDO were detected (sensitivity=73,0%). More than 9 in 10 detected signals corresponded to a WBDO (PPV=90,5%). The probability of detecting a WBDO

increased with distribution zone size (i.e., population serviced) and with the outbreak size (i.e., number of cases involved). From the results of our analysis, outbreak size had the strongest association with detection sensitivity. The model highlighted a non-linear relationship for the influence of outbreak size on the WBDO detection, the largest gain for sensitivity detection being between 5 and 10 cases. Accordingly, health insurance data constitute an adequate source for the retrospective surveillance of WBDO.

Introduction

Waterborne disease outbreaks (WBDO) are a public health concern in developed countries because of the high proportion of people affected when drinking water is contaminated (2, 3). To date, detection of these events has mainly been based on the voluntary reporting of clusters of acute gastrointestinal infections (AGI) by general practitioners to health authorities. Accordingly, because of the absence of a nationwide specific surveillance system the number of WBDO is most likely underestimated. Improving the detection of outbreaks caused by contaminated drinking water is quite a challenge for waterborne disease surveillance, as risk factors and high risk distribution zones (DZ) need to be identified more accurately. A key element in the development of such a system is that it should help decision makers formulate recommendations regarding the management of drinking water systems identified as being at risk of contamination, and to prevent recurrence of incidents. As part of a long-term plan to implement such a surveillance system nationwide, several years ago the French Institute for Public Health Surveillance (a unit within Santé Publique France) implemented a preliminary step towards this goal by using French Health Insurance administrative databases for the syndromic surveillance of medicalized AGI (4). These data have proven very relevant for describing the health impact of tap water quality and known WBDO (i.e. reported by doctors) (5, 6). Consequently, using them to automatically detect unknown WBDO would seem realistic and feasible.

In this context, an algorithm to detect unknown WBDO was previously developed and tested on real data (7) by Coly *et al.*. Their algorithm relies on a space-time statistical method developed by Kulldorff (1), and integrates the DZ into the detection process. Detected WBDO were analyzed regarding epidemiological criteria. Additional investigations were conducted for detected WBDO to investigate the existence of environmental factors (e.g. heavy rain) and technical incidents in the drinking water treatment (e.g. chlorination cessation, alarm malfunction) or in the distribution system (e.g. water pipe breaks).

One of the major challenges regarding large multiple outbreak detection systems of contaminated water is to identify the largest number of clusters corresponding to real WBDO (i.e. maximizing the sensitivity) while avoiding clusters that are not consistent with WBDO (i.e. minimizing the number of false positives). Recent reviews of simulation-based studies show that researchers are increasingly using simulations to generate realistic population health data in order to evaluate surveillance and disease control methods (8, 9). Moreover, other studies have developed and validated quantitative models for predicting the ability of commonly used surveillance algorithms to detect different types of outbreaks (10).

The main objective of our simulation-based study was to evaluate the performance of the algorithm developed by Coly *et al.* for WBDO detection using data from health administrative databases (11). A secondary objective was to estimate the factors which most influence WBDO detection.

Materials & Methods

Reference health data

We used health data from medicalized AGI cases from the French National Health Insurance Information System (SNIIRAM; Système national d'information inter régimes de l'Assurance maladie). SNIIRAM aims to evaluate beneficiaries' healthcare consumption and associated expenditures. It covers more than 98% of the French population and records all patient reimbursements for out-of-pocket medical procedures, medications and payments to professionals for consultations (12). It is possible to identify almost all medicalized AGI cases in France from SNIIRAM (4). AGI cases in SNIIRAM data were defined here as people who consulted a general practitioner for AGI and went to a pharmacy to buy medications prescribed to treat AGI. Cases were aggregated by day and municipality of residence. Because of the possibility that the SNIIRAM data included details of real (i.e. not simulated) WBDO, baseline data free of WBDO were obtained by simulation.

Study area and period

Two French administrative areas (termed "*départments"*) - Puy-de-Dôme and Isère, with 655 498 and 1 253 410 inhabitants, respectively (13) - were selected. Puy-de-Dôme was included in the previous study by Coly *et al.* for the construction of the WBDO detection algorithm. Isère is known for chronic microbiological pollution of DZ. The period of the reference dataset extended from January 1th 2010 to December 31th 2013.

Simulation study

We simulated the daily baseline counts of AGI at the municipality level (note: administrative areas in France comprise smaller municipal areas). A variety of simulated WBDO signals were superimposed on the baseline data. The simulation study was based on a methodology developed to evaluate the performance of an algorithm for outbreak detection of infectious diseases (8).

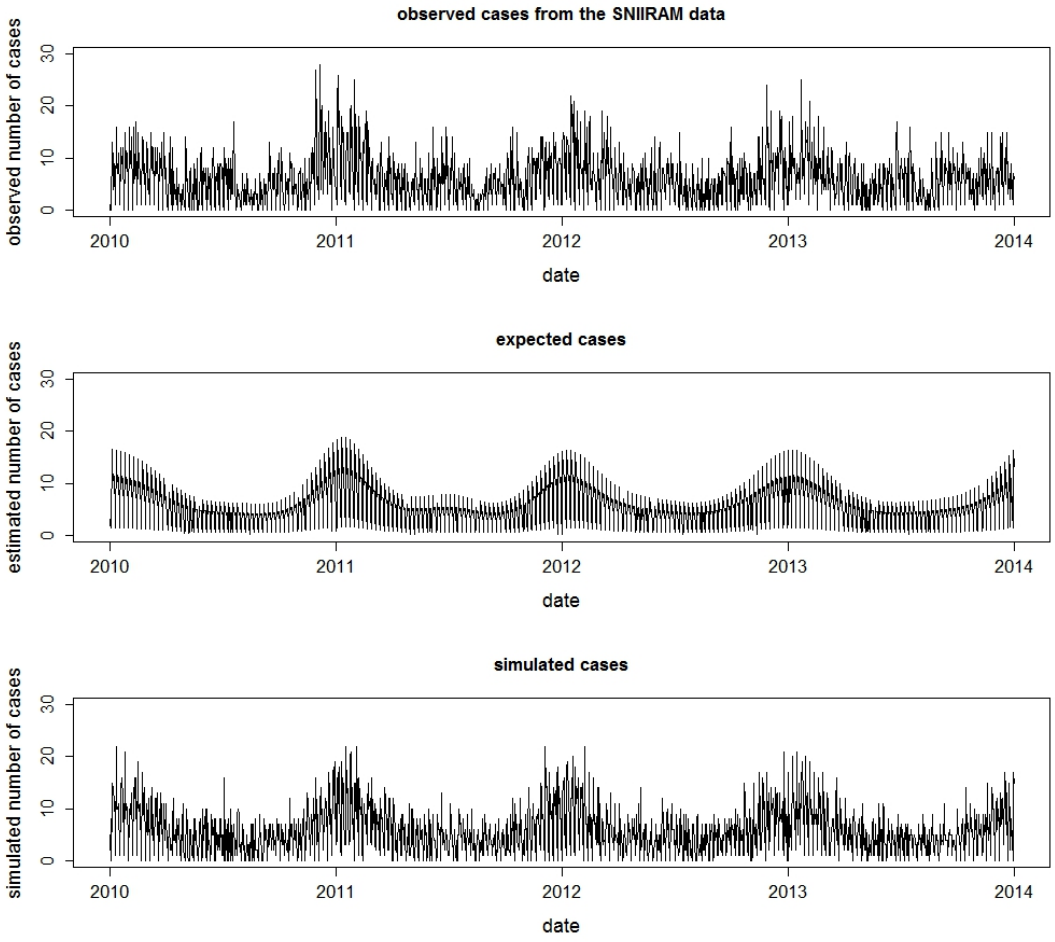
Simulation of baseline data

Baseline counts of AGI were simulated (1) at the administrative area level and (2) at the municipality level.

(1) A Poisson regression was performed to simulate the daily baseline counts of AGI at the administrative area level. A thin-plate regression spline (14) was used for modeling trend and seasonality, to take into account the seasonality of AGI cases and in particular the variability of winter outbreaks of AGI. Finally, baseline AGI data were based on observed values from the SNIIRAM data, with adjustments made for days of the week, holidays, seasonality and trend (10).

(2) To generate the baseline data at the municipality level, the AGI cases generated at the department level were distributed at the municipality level in proportion to the number of cases observed in the SNIIRAM data. Finally, in each municipality, daily baseline counts of AGI cases were generated by using a negative binomial distribution (8, 10) to introduce stochasticity (see Fig 1 for an example of the data collected from one municipality studied, chosen randomly).

Fig 1: Municipality study area with 18541 inhabitants: daily number of observed AGI cases from the SNIIRAM database between 01/01/2010 and 31/12/2013 (n=8677) *(top)*, number of estimated expected cases *(middle)* and number of simulated cases *(bottom)*.



The simulation process of waterborne disease outbreak

The spatial unit of interest for the WBDO simulation in our study was the DZ. By definition, a DZ delivers water of homogenous quality to consumers, meaning that all people serviced by the same DZ are exposed to the same risk in terms of water quality apart from situations of backflows and where contamination directly enters the network. Nevertheless, as the health outcome (i.e., AGI cases) was simulated at the municipality level, when the selected DZ serviced more than one municipality, AGI outbreak daily cases were then distributed according to the proportion of inhabitants serviced by the DZ in each municipality (15).

Several steps were implemented to simulate WBDO (Fig 2).

1/ The random selection of a DZ in the study area. DZs servicing fewer than 200 inhabitants were excluded from the simulation study to ensure statistical power of detection and because of their reduced impact on public health.

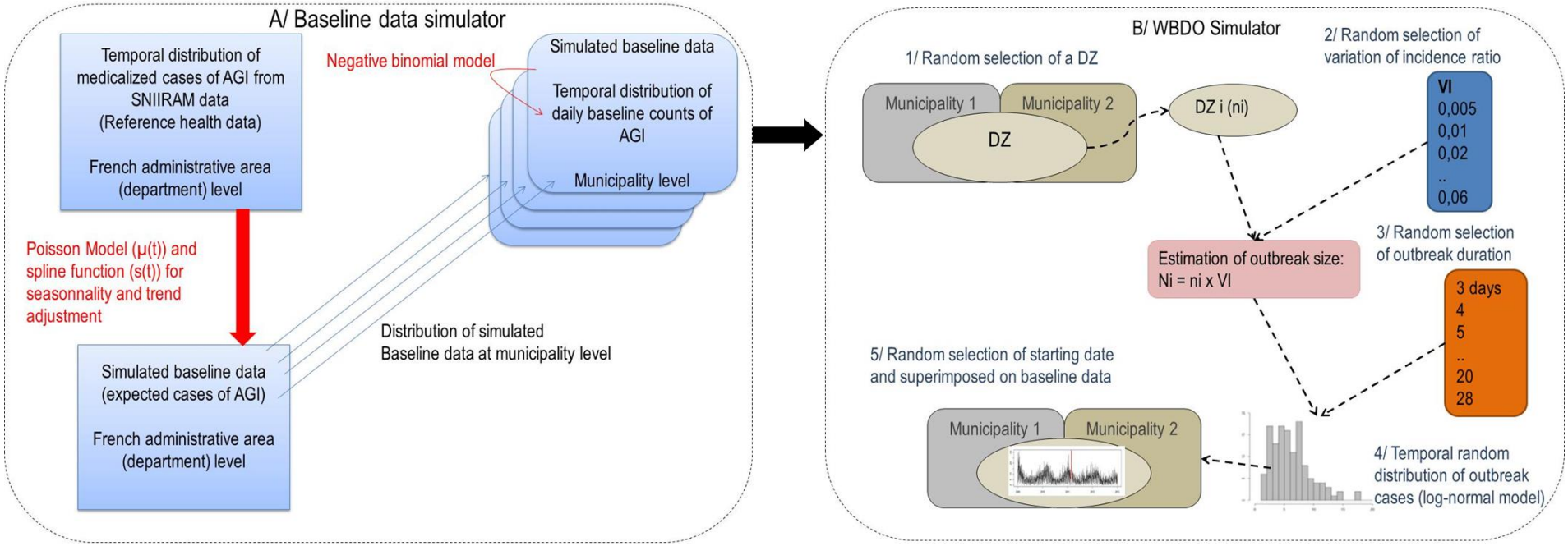
2/ The random selection of the variation of incidence ratio (VI). For each simulated outbreak, the VI was defined as the difference between the number of outbreak AGI cases and the number of expected cases of AGI (baseline data). It was randomly estimated at between 0,5% and 6% according to what we observed in previous WBDO in France (6). Comparatively with the health impact of WBDO assessed in cohort studies in which the attack rate varied between 30% and 50%, outbreak AGI cases observed in SNIIRAM data are less frequent. This difference is probably due to several factors including healthcare-seeking behavior: in French the mean consultation rate for AGI is quite low at 32% (16) and depends on age and pathogen agent (6).

3/ The random selection of the outbreak duration. Values between 3 to 28 days were randomly chosen in accordance with the observed values (17).

4/ The outbreak size (number of outbreak AGI cases) was generated by multiplying the VI by the number of inhabitants serviced by each DZ.

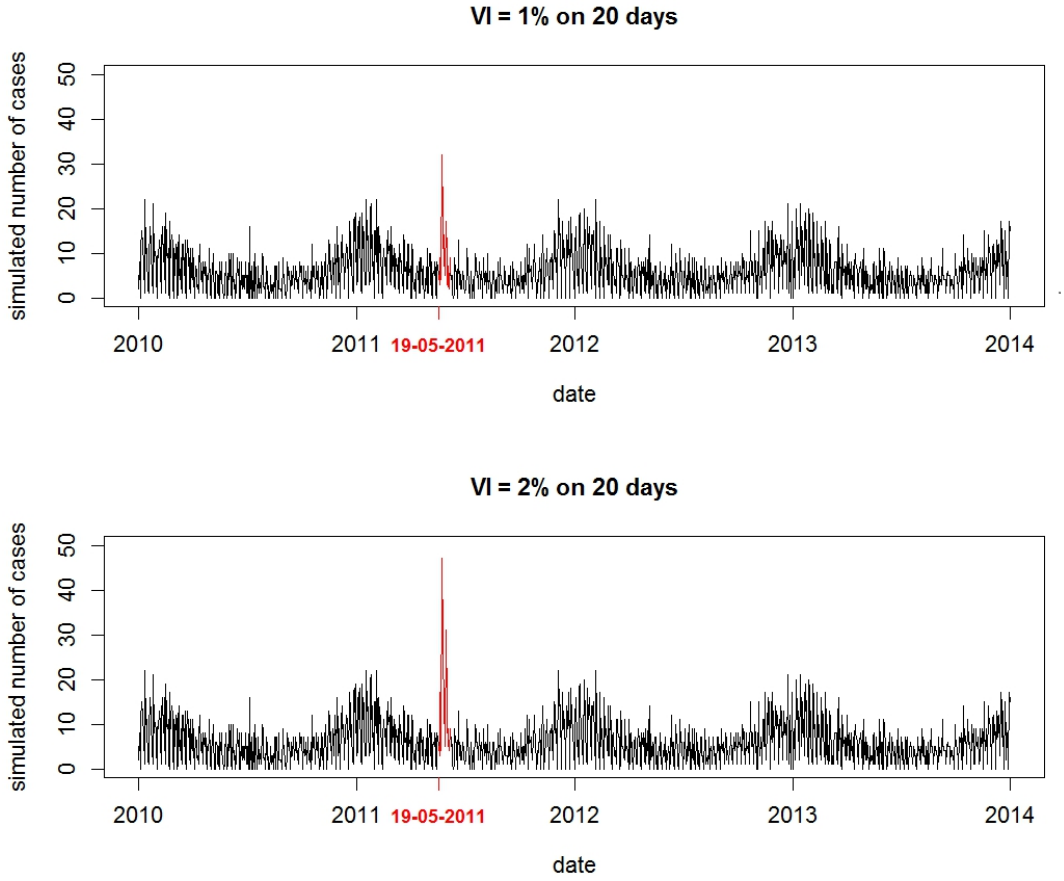
5/ Finally, outbreak cases were distributed over time according to a log-normal distribution (8, 10) (Fig 3). The parameters of the log-normal distribution used to shape the time-distribution of the outbreak AGI cases were randomly chosen between 0,33 and 0,5 for the median, and fixed at 0,5 for the standard deviation (6, 10). When the selected DZ serviced more than one municipality, AGI outbreak daily cases were then distributed according to the proportion of inhabitants serviced by the DZ in each municipality.

Fig 2: Algorithm of the overall process for simulation of baseline data and WBDO



DZ: distribution zone, WBDO: waterborne disease outbreak, VI: variation of incidence ratio; ni: population serviced by DZi; Ni: outbreak size

Fig 3: Illustration of two simulated outbreaks starting on 22/09/2011 in a municipality of 18 541 inhabitants serviced by only one DZ, with a variation of incidence ratio (VI) of 1% (top) and 2% (bottom) and with a 20-day duration.



By considering each DZ size (at least 200 people serviced) and the range of VI (0,5% to 6%), the simulated outbreak size varied between 1 and 12 cases for a DZ servicing 200 inhabitants and between 50 and 600 cases for a DZ servicing 10 000 inhabitants.

A total of 2000 simulations were run (1000 for each of the 2 French administrative areas studied). Each simulated set included the simulation of the baseline data and a WBDO.

The simulation study was performed using R v3.0.3.

Space-time detection method and implementation for WBDO detection

Space-time detection method

We used the space–time permutation scan statistic developed by Kulldorff *et al.* (1), based on overlapping cylinders, to define a scanning window. The circular base represents the geographical area of a potential outbreak. A classic approach is to iterate over a finite number of geographical grid points and gradually increase the circle radius from zero to a maximum value defined by the user. Hence, both

small and large circles are considered, all of which overlap with many other circles. The cylinder height represents the number of days, up to a maximum defined by the user. With a space-time permutation scan statistic, expected cases are calculated using only observed cases. A Poisson generalized likelihood ratio is then used as a measure of the evidence that a tested cylinder contains an outbreak. The cylinder with the maximum generalized likelihood ratio constitutes the space–time cluster of cases least likely to occur by chance and consequently it is the primary candidate for a true outbreak.

Implementation for waterborne disease outbreak detection

To detect WBDO, the space-time scan statistic developed was implemented at the municipality level by considering the DZ servicing each municipality. Municipalities were considered neighbors if they were serviced by the same DZ. An algorithm taking into account the spatial distribution of the population over municipalities serviced by a DZ was used (detail in (11)). A “municipality neighbors” table was then generated according to the DZ area. For all analyses, adjustments were made for days of the week and holidays.

This detection study was performed using SaTScan v9.3 (18) and R (v3.0.3).

Data analysis

Evaluation method

Two types of signals were distinguished among clusters detected: “true alarms”, which corresponded both to a cluster of AGI and a simulation-generated WBDO , and “false alarms”, considered by the detection method to be a significant signal (cluster of AGI) relating to statistical criteria but not generated as a WBDO by the simulation process.

To evaluate the performance of the WBDO detection method, the sensitivity (Se) and the positive predictive value (PPV) were both used. Sensitivity was estimated as the ratio between the number of detected WBDO and the number of simulated WBDO (i.e., 1000 WBDO generated per administrative area). We considered that a WBDO was detected if at least one detected day and one detected municipality corresponded to all the days and municipalities involved in the simulated WBDO (8). The PPV was defined as the ratio between the number of detected WBDO and the number of all the clusters detected. The latter corresponded to clusters associated with a statistical threshold (p-value) of 0,05, whether it concerned a true alarm (i.e., WBDO-generated) or not.

Associated factors

Multivariable Poisson regression was performed to identify the factors associated with WBDO detection (outcome) and to estimate the strength of these associations (10, 19). Five dependent variables were

considered: outbreak duration, population size of the DZ, outbreak incidence ratio (VI), outbreak size (number of AGI cases) and season ("winter" for December, January, February and March / "other seasons" for April to November). We tested for potential interactions. Incidence rate ratios (IRR) and their 95% confidence intervals (CI) were computed. The IRR is the ratio between the incidence rate in a considered group and the incidence rate in the reference group. Fractional polynomials were used to model the relationship between continuous dependent variables and WBDO detection (20, 21). All analyses were performed using Stata 12.0 (StataCorp LP, USA).

Results

Description of simulated WBDO

Simulated WBDO in both departments involved between 1 and 7392 outbreak AGI cases (median=22; mean=96,2) and most of them (90%) included 200 AGI cases or fewer (Table 1). The mean outbreak duration was 15 days (3 to 28 days). All DZ sizes were represented: 35,8% of the DZ randomly selected serviced 500 people or fewer, 37,3% between 500 and 2000 people and 27,0% more than 2000 people. For the simulated WBDO involving 200 AGI cases or fewer, outbreak size was dependent on the DZ size but independent of the season of simulation (winter/other seasons). Among all the simulated WBDO, 26,7% (n=534/2000) involved a DZ servicing more than one municipality: 162 WBDO generated for Isere with 2 to 13 municipalities serviced by the same DZ, and 372 for Puy-de-Dôme with 2 to 53 municipalities being simultaneously served.

Table 1: Description of simulated WBDO by department

		Both departments					Puy-de-Dôme					Isère				
		total	detected		undetected		total	detected		undetected		total	detected		undetected	
	N	2000	1460	%	540	%	1000	726	%	274	%	1000	734	%	266	%
DZ size (number of inhabitants served by DZ)																
	200-500	715	353	49.4%	362	50.6%	385	201	52.2%	184	47.8%	330	152	46.1%	178	53.9%
	501-1000	437	330	75.5%	107	24.5%	204	153	75.0%	51	25.0%	233	177	76.0%	56	24.0%
	1001-2000	309	264	85.4%	45	14.6%	128	107	83.6%	21	16.4%	181	157	86.7%	24	13.3%
	2001-10 000	421	396	94.1%	25	5.9%	188	171	91.0%	17	9.0%	233	225	96.6%	8	3.4%
	> 10 000	118	117	99.2%	1	0.8%	95	94	98.9%	1	1.1%	23	23	100.0%	0	0.0%
Outbreak size (number of simulated cases of AGI)																
	Min	1	5		1		2	6		2		1	5		1	
	p10	5	11		2		5	11		2		5	12		2	
	Median	22	38		6		22	35		6		23	39		6	
	Mean	96.2	128.8		8.1		122.5	165.3		8.9		69.9	92.6		7.3	
	p90	199	271		14		255	412		15		140	187		14	
	Max	7392	7392		133		5551	5551		133		7392	7392		33	
Duration (days)																
	Min	3	3		3		3	3		3		3	3		3	
	Median	16	15		17		15	14		16		16	15		18	
	Mean	15.4	15.0		16.4		15.2	14.8		16.3		15.6	15.2		16.5	
	Max	28	28		28		28	28		28		28	28		28	
DZ area (number of municipalities served)																
	1	1466	1042	71.1%	424	28.9%	628	445	70.9%	183	29.1%	838	597	71.2%	241	28.8%
	>1	534	418	78.3%	116	21.7%	372	281	75.5%	91	24.5%	162	137	84.6%	25	15.4%
Season																
	Winter	605	414	68.4%	191	31.6%	298	199	66.8%	99	33.2%	307	215	70.0%	92	30.0%
	Other seasons	1395	1046	75.0%	349	25.0%	702	527	75.1%	175	24.9%	693	519	74.9%	174	25.1%

DZ: distribution zone; *Winter: December, January, February, March

Sensitivity and positive predictive value of the detection method

Almost three quarters of the 2000 simulated WBDO were detected (sensitivity=73,0%). More than 9 in 10 detected signals corresponded to a WBDO (PPV=90,5%). The probability of detecting a WBDO increased with DZ size (i.e., population serviced) and with the outbreak size (i.e., number of cases involved) (Table 2). Moreover, WBDO in non-winter seasons (hereafter "other seasons") were better detected than WBDO simulated during the winter season. Consequently, to reach the same sensitivity value of 75%, WBDO size had to be greater in the winter season than in other seasons (at least 15 cases versus 10 cases, respectively) (Fig 4).

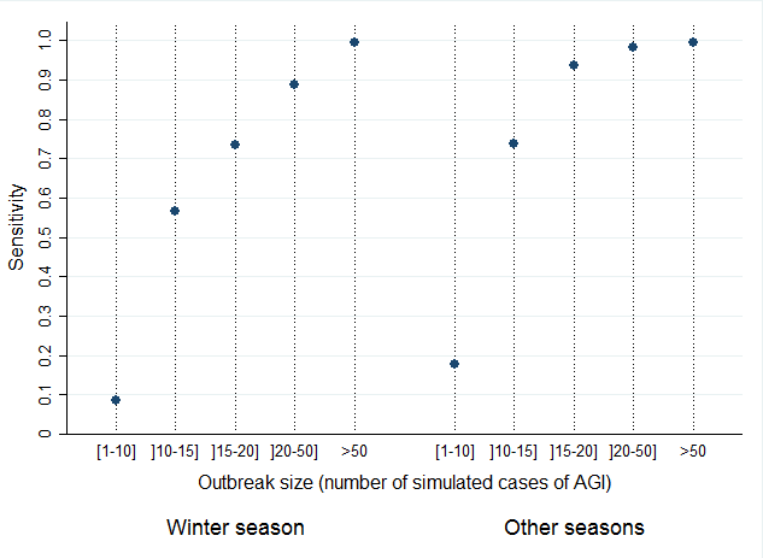
Table 2: Sensitivities and predictive positive values of the detection method according to outbreak size, distribution zone (DZ) size, season and DZ area.

	Total				Isère				Puy-de-Dôme			
	Se		PPV		Se		PPV		Se		PPV	
	%	N1	%	N2	%	N1	%	N2	%	N1	%	N2
Total	**73,0%**	**2000**	**90,5%**	**1614**	**73,4%**	**1000**	**89,0%**	**825**	**72,6%**	**1000**	**92,0%**	**789**
DZ size (number of inhabitants served by DZ)												
200-500	49,4%	715	88,0%	401	46,1%	330	82,2%	185	52,2%	385	93,1%	216
501-1000	75,5%	437	91,4%	361	76,0%	233	92,7%	191	75,0%	204	90,0%	170
1001-2000	85,4%	309	93,0%	284	86,7%	181	91,3%	172	83,6%	128	95,5%	112
2001-10 000	94,1%	421	91,5%	433	96,6%	233	89,3%	252	91,0%	188	94,5%	181
> 10 000	99,2%	118	86,7%	135	100,0%	23	92,0%	25	98,9%	95	85,5%	110
Outbreak size (number of simulated cases)												
[1-10]	15,2%	466	77,2%	92	13,8%	224	77,5%	40	16,5%	242	76,9%	52
]10-15]	68,6%	312	91,5%	234	64,7%	150	85,8%	113	72,2%	162	96,7%	121
]15-20]	86,5%	170	91,9%	160	83,3%	90	90,4%	83	90,0%	80	93,5%	77
]20-50]	95,3%	449	90,9%	471	97,9%	240	89,0%	264	92,3%	209	93,2%	207
>50	99,5%	603	91,3%	657	100,0%	296	91,1%	325	99,0%	307	91,6%	332
Season												
Winter	68,4%	605	87,7%	472	70,0%	307	84,3%	255	66,8%	298	91,7%	217
Other season	75,0%	1395	91,6%	1142	74,9%	693	91,1%	570	75,1%	702	92,1%	572
DZ area (number of municipalities served)												
1	71,1%	1466	90,2%	1155	71,2%	838	88,7%	673	70,9%	628	92,3%	482
>1	78,3%	534	91,1%	459	84,6%	162	90,1%	152	75,5%	372	91,5%	307

Se: sensitivity; PPV: positive predictive value; N1: number of WBDO simulated; N2: number of clusters detected with p-value <=0.05; DZ: distribution zone

*Winter: December, January, February, March

Fig 4: Sensitivity of detection method according to outbreak size (number of simulated AGI cases) and season (winter: December, January, February, March).



For WBDO occurring in more than one municipality (534 for both departments studied), the sensitivity of detection was higher than for WBDO associated with a single DZ servicing only one municipality (78.3% and 71,1%, respectively), while the PPV was stable (90,2% versus 91,1%, respectively). For half of the WBDO associated with a single DZ servicing several municipalities, 80% of these municipalities were included in the detected signal for Isere and 50% for Puy-de-Dôme.

The undetected WBDO involved mostly small DZ (200-500 inhabitants) and fewer outbreak cases (Table 1).

Factors associated with WBDO detection

In the multivariable Poisson regression, the outbreak size, the VI, the duration and the season of WBDO were all significantly associated with detection ($p<0,05$). The interaction of VI and the outbreak size was significant ($p<0,05$).

WBDO involving at least 10 AGI cases, with a 14-day duration or less, and occurring between April and November, had a higher probability of being detected (Table 3). The variable “outbreak size” had the strongest association with detection.

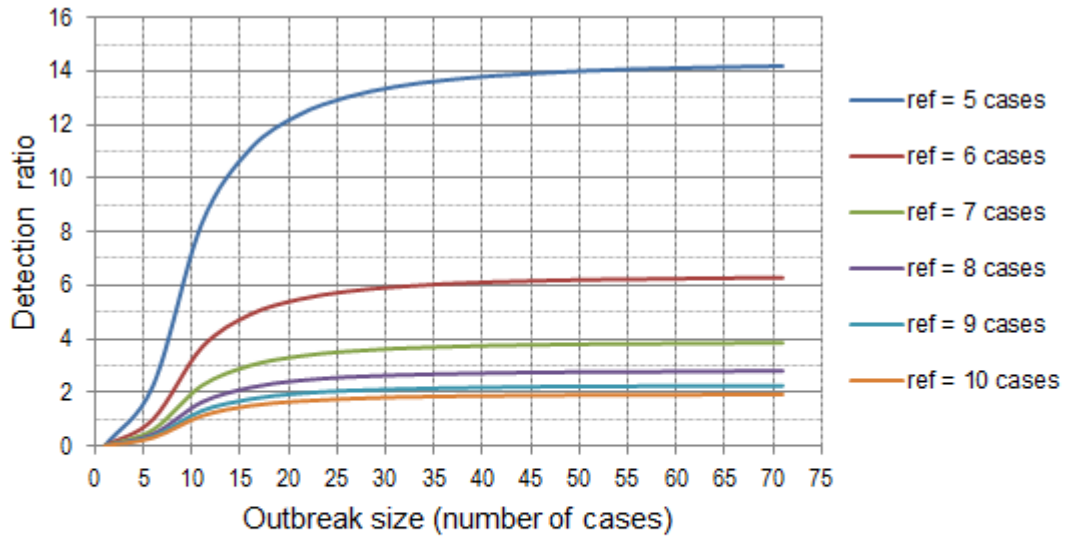
The multivariable fractional polynomials model highlighted a non-linear relationship for the influence of outbreak size on the WBDO detection ratio (Fig 5), the largest gain for sensitivity detection being between 5 and 10 cases. Compared with a 5-case outbreak size (chosen as reference), the WBDO detection ratio dramatically increases with an outbreak size between 6 and 15 AGI cases.

Table 3: Final multivariable regression model with factors significantly associated with WBDO detection, stratified by variation of incidence ratio (VI).

		VI : [0.5% - 2.0%]					VI :]2.0% - 4.0%]					VI :]4.0% - 6.0%]				
Variables		n=642	IRR	[95% CI]		p-value	n=659	IRR	[95% CI]		p-value	n=699	IRR	[95% CI]		p-value
Outbreak size (number of simulated cases)																
	[1-10]	331	Ref				129	ref			-	6	ref			
	]10-15]	79	7.70	5.03	11.74	$<10^{-3}$	130	2.01	1.55	2.61	$<10^{-3}$	103	1.59	0.71	3.58	0.26
	]15-20]	40	10.30	6.79	15.60	$<10^{-3}$	51	2.62	2.02	3.39	$<10^{-3}$	79	1.91	0.86	4.28	0.11
	]20-50]	96	12.80	8.71	18.82	$<10^{-3}$	151	2.85	2.24	3.62	$<10^{-3}$	202	1.96	0.88	4.37	0.10
	>50	96	13.70	9.29	20.07	$<10^{-3}$	198	2.92	2.30	3.70	$<10^{-3}$	309	2.03	0.91	4.53	0.08
Season																
	Winter*	193	Ref				193	ref				219	ref			
	Other seasons	449	1.37	1.20	1.56	$<10^{-3}$	466	1.11	1.03	1.19	$<10^{-2}$	480	1.05	1.01	1.10	0.01
Outbreak duration (days)																
	[3-7]	131	Ref				136	ref				133	ref			
	]7-14]	173	0.84	0.73	0.97	0.02	180	1.00	0.92	1.09	0.95	184	0.97	0.94	1.01	0.10
	]14-21]	178	0.77	0.66	0.90	$<10^{-2}$	170	0.89	0.81	0.97	0.01	178	0.94	0.90	0.99	0.01
	]21-28]	160	0.64	0.54	0.76	$<10^{-3}$	173	0.89	0.81	0.98	0.02	204	0.93	0.89	0.97	$<10^{-2}$

VI: variation of the incidence ratio; IRR: incidence rate ratio; CI: confidence Interval; WBDO: waterborne disease outbreak
*Winter: December. January. February. March.

Fig 5: WBDO detection ratio according to the outbreak size (2000 simulated WBDO).



Discussion

Simulation process

The first step of this simulation-based study was to generate the baseline incidence of the disease using the reference health data. This step employed a published method *(8)* adapted for AGI epidemiology by adding a flexible adjustment function (spline) to account for winter person-to-person outbreaks of viral AGI . An adjustment for days of the week and holidays was also made to reflect the closure of pharmacies during weekends and holidays. This ensured an acceptable representativeness of year-dependent seasonality and of the incidence of AGI (Fig 1).

Simulated WBDO were generated using a log-normal distribution model (8). The parameters used to build these epidemic signals were inspired by past observed WBDO. Accordingly, both the chosen epidemic duration ranging from 3 to 28 days and the chosen VI comprised between 0,5% and 6%, are realistic. Simulation of WBDO also considered the overlap between the DZ limit and the limits for municipalities serviced, by spreading the outbreak cases over all the municipalities serviced. The distribution of the outbreak cases between the different municipalities was set to be equal to the baseline incidence of cases in these municipalities. This choice is valid if we assume homogenous exposure to all people serviced by the affected drinking water system, a situation which is generally true when pollution comes from the water resource. However, it is not true when the contamination point of entry is located in the water system, e.g. when a contaminated backflow occurs. Indeed such backflows accounted for half of the WBDO reported in France between 1998 and 2006 (17). In these situations, the distribution of the outbreak AGI cases through the DZ can result in localized clusters of outbreak cases on a part of the DZ. In our algorithm, which successively includes –one by one - all the municipalities serviced by the same DZ in the detection process (as opposed to including all municipalities together at once), this limitation should not affect the sensitivity of detection..

Another study limitation was the choice not to perform WBDO simulations for DZ servicing fewer than 200 inhabitants. This therefore prevented us from being able to evaluate the algorithm for these DZ. However previous work from this study (unpublished data) showed that the detection sensitivity for such DZ was close to zero, except when the VI was greater than 5%.

The benefits and limitations of the detection algorithm used here in the context of an integrated detection system for WBDO is detailed elsewhere (11).

Algorithm performance for WBDO detection

Globally, the algorithm detected an estimated 73% of simulated WBDO (sensitivity). Among the detected signals, an estimated 90,5% corresponded to simulation-generated WBDO (Positive predictive value). These estimations reached, respectively, 99,2% and 86,7% for DZ servicing more than 10 000 people (Table 2). The performance (sensitivity and positive predictive value) of the algorithm mainly depended on the serviced population size, the outbreak size, their duration, the season and the intensity of the outbreak (VI) (Table 2 and 3). If we focus on the influence of serviced population size, the threshold of 500 inhabitants resulted in increased sensitivity from one in two to more than three in four detected WBDO.

Likewise, when the outbreak size exceeded 10 AGI cases, the sensitivity was four times greater compared with smaller outbreaks (10 cases or fewer). For these two parameters (i.e. the serviced population size and the outbreak size), the most significant variations in sensitivity were observed when 200 to 2000 people were serviced (from 49,4% to 94,1% respectively) and from fewer than 10 cases to over 50 (17,2% and 99,3%).

To conclude this section on the algorithm's performances, detection ability was primarily related to the serviced population size and to outbreak intensity (VI), the outbreak size being the product of both. One substantial limitation of this study is that we examined only one detection algorithm.

Factors influencing detection

In addition to the evaluation of the performance of the detection algorithm, the simulation study also allowed us to identify and quantify the three factors which most influence the performance of WBDO detection as follows: outbreak size, duration and season ("winter" or "other seasons"). The existence of a significant interaction between the outbreak size and the VI led us to consider the results according to three classifications of the incidence ratio (0,5% to 2%; 2% to 4%; 4% to 6%) (Table 3).

From the results of our analysis, outbreak size had the strongest association with detection sensitivity, especially for a VI value below 4%. Above this value, outbreak size was no longer associated with detection. For VI between 0,5% and 2%, the the outbreak size has the dominant effect (vis-à-vis duration and season) with a detection capacity 13 times greater for an outbreak of 20 AGI cases or more than for an outbreak of 10 or fewer cases (the incidence rate ratio is 7,7 when going from fewer than 10 cases to 15 cases). For a VI between 2% and 6%, the detection ratio (IRR) did not exceed 3 between the most extreme values (more than 50 cases versus fewer than 10 cases). These results suggest a strong improvement of detection ability for WBDO with more than 10 AGI cases and a VI greater than 2%. These values are consistent with the previous study's results which described the detection algorithm and its application to real health data (11). Of the 11 clusters detected in this study,

the values of the medication rate in the population (indicator close to the VI) ranged from 0,7% to 4,8%, and the cluster size from 21 to 67 AGI cases.

As mentioned above, “duration” and “season” also affected detection but much less substantially than outbreak size. WBDO with lower VI (0,5%- 2%) were primarily affected by these three factors. Accordingly, the number of detected WBDO was 1,3 times higher in non-winter season (i.e. defined “other seasons” in the text) outbreaks of AGI. Similarly, outbreaks which lasted less than 14 days were better detected than longer outbreaks.

International comparison

To our knowledge few previously published simulation studies on WBDO detection exist. Different Canadian research studies presented an agent-based simulation model for generating realistic multivariable outbreak signals (22). This model was used to simulate a WBDO caused by *Cryptosporidium*, taking into account parameters for population, water consumption and disease progression. To verify whether the presented simulation model as a whole produces credible results, in a real outbreak scenario, the authors attempted to replicate the largest documented WBDO of Cryptosporidiosis which occurred in Milwaukee in 1993. During that outbreak, over 400 000 people were estimated to have diarrhea attributable to acute *Cryptosporidium* infection (23). The simulation was repeated 1000 times using different seeds for the random number generator. The results showed that the simulated curve was slightly more positively skewed and peaked one to two days earlier than the historically observed curves. These simulated data were then used to improve early outbreak detection using a hidden Markov model (24).

Finally, the particularities of the French health database used for WBDO detection (health insurance covers almost all the French population), the nature of the data (daily medicalized AGI cases, aggregated at a municipality level) and the delay of data availability (almost 2 months), limit the possibility of international comparison. Accordingly, although health insurance data constitute an adequate source for the retrospective surveillance of WBDO, they do not allow - at least for the moment – the possibility of implementing, a prospective approach within the context of a public health alert system. On this point, health insurance's data differ from other syndromic surveillance data usually appreciated for their reactivity (25). This lack of reactivity is offset by the completeness of data covering the entire French population (overseas territories included), and the level of geographical precision (Municipality zip code). Therefore, the usefulness of these data for the prevention of WBDO epidemics focuses on identifying the affected water systems and possibly associated risk factors. Furthermore, they could also help to assess the health impact of WBDO in France. This assessment however

requires extensive further work taking into account the factors influencing healthcare-seeking behaviour for AGI and access to health services in a WBDO context (6).

Our study presents a global approach for simulating AGI baseline data using reference health data and superimposing simulated WBDO. The algorithm for WBDO detection, based on simulated WBDO, was evaluated as being able to detect almost 90% of WBDO, with few false positive alarms. We also demonstrated that it is possible to quantitatively estimate the factors which most influence WBDO detection.

Acknowledgments

The authors wish to express their appreciation and gratitude to the National Health Insurance for access to health administrative data, to Magali Corso of the French national public health agency for the preparation of case data of acute gastro-enteritis from the National Health Insurance databases, to Catherine Galey and Loïc Rambaud of the French national public health agency for their contributions to the conceptualization of the study, to the Health Ministry and Henri Davezac for water data drawn from the Sise-Eaux database and to Farida Mihoub and Jude Sweeney for her help in translation.

References

1. Kulldorff M, Heffernan R, Hartman J, Assuncao R, Mostashari F. A space-time permutation scan statistic for disease outbreak detection. PLoSMed. 2005;2(3):e59.

2. Craun GF, Brunkard JM, Yoder JS, Roberts VA, Carpenter J, Wade T, et al. Causes of outbreaks associated with drinking water in the United States from 1971 to 2006. Clinical microbiology reviews. 2010;23(3):507-28.

3. Hrudey SE, Hrudey EJ. Safe Drinking Water : Lessons from Recent Outbreaks in Affluent Nations. London: IWA publishing; 2004. 486 p.

4. Bounoure F, Beaudeau P, Mouly D, Skiba M, Lahiani-Skiba M. Syndromic surveillance of acute gastroenteritis based on drug consumption. Epidemiology and infection. 2011;139(9):1388-95.

5. Beaudeau P. Syndromic surveillance of acute gastroenteritis: an opportunity for the prevention of the infectious risk attributable to tap water. Rennes: Université de Rennes 1; 2012.

6. Mouly D, Van Cauteren D, Vincent N, Vaissiere E, Beaudeau P, Ducrot C, et al. Description of two waterborne disease outbreaks in France: a comparative study with data from cohort studies and from health administrative databases. Epidemiology and infection. 2016;144(3):591-601.

7. Coly S, Vincent N, Vaissiere E, Charras-Garridol M, Gallay A, Ducrot C, et al. Waterborne disease outbreaks detection: an integrated approach using health administrative databases. Journal of water and health. 2016;in press(accepted 12th June 2016).

8. Noufaily A, Enki DG, Farrington P, Garthwaite P, Andrews N, Charlett A. An improved algorithm for outbreak detection in multiple surveillance systems. StatMed. 2013;32(7):1206-22.

9. Buckeridge DL, Jauvin C, Okhmatovskaia A, Verma AD. Simulation Analysis Platform (SnAP): a tool for evaluation of public health surveillance and disease control strategies. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2011;2011:161-70.

10. Buckeridge DL, Okhmatovskaia A, Tu S, O'Connor M, Nyulas C, Musen MA. Predicting outbreak detection in public health surveillance: quantitative analysis to enable evidence-based method selection. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2008:76-80.

11. Coly S, Vincent N, Vaissiere E, Charras-Garrido M, Gallay A, Ducrot C, et al. Detection of waterborne disease outbreaks: an integrated approach usin health administrative databases. . Journal of water and health.under review.

12. Tuppin P, De Roquefeuil L, Weill A, Ricordeau P, Merliere Y. French national health insurance information system and the permanent beneficiaries sample. RevEpidemiolSante Publique. 2010;58(4):286-90.

13. Insee. 2015.

14. Wood SN. Generalized Additive Models: An Introduction with R: Chapman and Hall/CRC; 2006.

15. FrenchMinistryofHealth. French database on public drinking water quality.

16. Van Cauteren D, De Valk H, Vaux S, Le Strat Y, Vaillant V. Burden of acute gastroenteritis and healthcare-seeking behaviour in France: a population-based study. EpidemiolInfect. 2012;140(4):697-705.

17. Beaudeau P, de Valk H, Vaillant V, Mannschott C, Tillier C, Mouly D, et al. Lessons learned from ten investigations of waterborne gastroenteritis outbreaks, France, 1998-2006. Journal of water and health. 2008;6(4):491-503.

18. Kulldorff M. Inc. SaTScanTM v8.0: Software for the spatial and space-time scan statistics. In: Services IM, editor. 2009.

19. Barboza P, Vaillant L, Le SY, Hartley DM, Nelson NP, Mawudeku A, et al. Factors influencing performance of internet-based biosurveillance systems used in epidemic intelligence for early detection of infectious diseases outbreaks. PLoSOne. 2014;9(3):e90536.

20. Royston P, Altman D. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. Journal of the Royal Statistical Society. 1994;43(3):429-67.

21. Sauerbrei W, Royston P. Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. Journal of the Royal Statistical Society. 1999;162(1):71-94.

22. Okhmatovskaia A, Verma AD, Barbeau B, Carriere A, Pasquet R, Buckeridge DL. A simulation model of waterborne gastro-intestinal disease outbreaks: description and initial evaluation. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2010;2010:557-61.

23. MacKenzie WR, Schell WL, Blair KA, Addiss DG, Peterson DE, Hoxie NJ, et al. Massive outbreak of waterborne cryptosporidium infection in Milwaukee, Wisconsin: recurrence of illness and risk of secondary transmission. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America. 1995;21(1):57-62.

24. Morrison K, Charland K, Okhmatovskaia A, Buckeridge D. A Framework for Detecting and Classifying Outbreaks of Gastrointestinal Disease. Online J Public Health Inform. 2013;5(1).

25. Berger M, Shiau R, Weintraub JM. Review of syndromic surveillance: implications for waterborne disease detection. Journal of epidemiology and community health. 2006;60(6):543-50.

Troisième partie : discussion générale et perspectives

Nos travaux ont permis d'étudier la pertinence d'une nouvelle méthode intégrée pour améliorer la détection des épidémies de gastro-entérite aigüe d'origine hydrique. Plusieurs points de convergences sont identifiés au terme de ce travail :

- Les données de l'Assurance Maladie sont pertinentes pour la détection d'un signal épidémique de gastro-entérite aigüe médicalisés d'origine hydrique ;
- La prise en compte de l'exposition à l'eau du robinet a permis de développer une méthode intégrée utilisant les données de l'Assurance Maladie et les données de la base nationale SISE-eaux pour la détection des épidémies d'origine hydrique ;
- La méthode développée se caractérise par une bonne sensibilité et une bonne spécificité de détection ;
- La capacité de détection dépend principalement de la taille de l'épidémie, des habitudes de recours aux soins de la population impactée, de la répétition d'épidémies hydriques pour une même population ;
- Les délais de consolidation des données de l'Assurance Maladie de quelques semaines permettent uniquement une détection rétrospective ;
- La mise au point d'un outil automatisé facilite la mise en œuvre de la détection dans une perspective de surveillance ;
- La nécessité des investigations environnementales pour associer les mesures de gestion à la surveillance.

Les principaux résultats et les perspectives pour la surveillance sont discutés dans les paragraphes suivants.

1 Principaux résultats

1.1 Le profil des épidémies de gastro-entérite aigüe d'origine hydrique dans les données de l'Assurance Maladie : nature du signal à détecter

La proportion de cas de gastro-entérite aigüe, estimée à partir des données de l'Assurance Maladie[12], parmi l'ensemble de la population lors d'une épidémie d'origine hydrique est de l'ordre de 1 à 5% : 1,5 et 2,0% dans deux épidémies connues (cf. article 1) et de 0,7 à 4,8% dans les épidémies détectées en Auvergne entre 2009 et 2012 (article 2). Par ailleurs, une grande variabilité peut être observée sur la proportion de cas de gastro-entérite aigüe médicalisés entre les épidémies. Les principaux facteurs pouvant influencer la sensibilité de l'indicateur Assurance Maladie pour décrire les épidémies d'origine hydrique sont l'âge, la nature de l'agent pathogène, la connaissance du risque lié à l'eau dans la commune (existence de pollutions chroniques pouvant entraîner des constitutions de stocks de médicaments dans l'armoire familiale) et l'accès aux structures de soins (médecins, pharmacies). Concernant l'effet de l'âge, la classe des enfants de moins de 15 ans est intéressante car la proportion de cas de gastro-entérite aigüe médicalisés y est plus importante que chez les adultes. Ainsi, une plus grande proportion de cas de gastro-entérite aigüe médicalisés chez les enfants, peut être un argument en faveur d'une épidémie d'origine hydrique.

La durée des signaux détectés en Auvergne entre 2009 et 2012, correspondant à des épidémies dont l'origine hydrique a été confortée, est de l'ordre de 2 semaines (moyenne 16 jours pour les 11 signaux de l'article 2) et s'étale de 7 à 35 jours. Cette valeur correspond à la durée du signal pour laquelle la statistique de test était la plus significative. Il n'est pas possible de l'interpréter comme étant la durée d'une épidémie. A titre d'exemples, les deux épidémies de 2010 et 2012 détaillées dans l'article 1, avaient des durées d'environ 3 semaines chacune d'après les études de cohortes alors qu'elles étaient associées à des signaux d'une durée respective de 7 et 28 jours par la méthode de détection. Ces deux exemples montrent que la méthode de détection peut se comporter différemment en fonction du signal présent dans les données (amplitude, variation de l'incidence par rapport au bruit de fond de la maladie) et que la durée des signaux détectés peut être augmentée ou réduite par rapport à la réalité. Elle pourra ainsi détecter un signal sans prendre en compte les cas secondaires tardifs si l'épidémie est soudaine et intense (cas de l'épidémie de 2010) ; avec potentiellement une durée du signal inférieure à la durée

[12] C'est-à-dire les malades allant consulter un médecin et acheter des médicaments remboursés à la pharmacie dans les 24 heures après la consultation

réelle de l'épidémie. A l'inverse, pour des épidémies d'intensité plus faible comme celle de 2012, le signal détecté pourra avoir une durée supérieure à la durée de l'épidémie réelle.

Dans l'étude de simulation, la durée du signal des épidémies détectées est dans la plupart des cas inférieure à la durée des épidémies simulées (Figure 15).

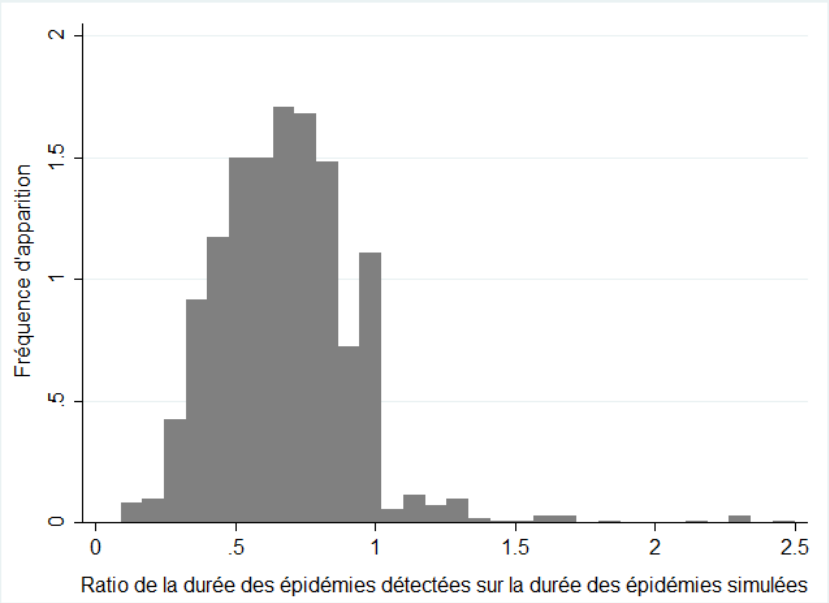


Figure 15 : Ratio de la durée des épidémies détectées sur la durée des épidémies simulées pour 1460 épidémies simulées, départements Auvergne et Isère, 2010-2014

1.2 Les performances et les limites de la méthode de détection des agrégats de gastro-entérite aigüe ayant comme point commun la même exposition à l'eau du robinet

La méthode de détection spatio-temporelle de Kulldorff appliquée rétrospectivement aux cas de gastro-entérite aigüe de l'Assurance Maladie, en tenant compte du contour des réseaux de distribution d'eau permet d'améliorer la détection des épidémies d'origine hydrique.

En appliquant cette méthode sur des données simulées, on estime que plus de 95% des épidémies d'origine hydrique impliquant plus de 20 cas sont détectées, quelle que soit la période de l'année et quel que soit le nombre de communes desservies par une même unité de distribution d'eau.

Les épidémies inférieures à 5 cas sont peu ou pas détectées (<20%).

Entre 5 et 20 cas, la méthode a une capacité de détection supérieure de 10% hors période hivernale et/ou pour les unités de distribution desservant plusieurs communes.

Enfin, la sensibilité de la méthode de détection est d'environ 50% pour des unités de distribution d'eau qui desservent entre 200 et 500 personnes et de plus de 95% pour les unités de distribution d'eau

desservant 2 000 personnes ou plus. Entre ces deux classes de population, la sensibilité varie de 75% (entre 500 et 1 000 personnes desservies) à 85% (entre 1 000 et 2 000 personnes desservies).

D'après les tests effectués sur des données simulées, 90% des signaux détectés correspondent à un agrégat de cas de gastro-entérite aigüe avec un profil épidémique compatible avec une épidémie d'origine hydrique et une emprise géographique compatible avec un réseau de distribution d'eau. Pour ces signaux, la réalisation d'une enquête environnementale permettrait d'identifier les causes et les circonstances de la pollution (cf. partie 1.3 ci-dessous). Cette valeur prédictive positive est associée dans l'étude de simulation (article 3) à des épidémies d'une durée de 3 jours ou plus et impliquant au moins 10 cas épidémiques. Dans l'étude sur des données réelles (article 2), on retrouve la même valeur prédictive positive (10 signaux/11 ont été confortés par des indicateurs environnementaux) pour les signaux d'une durée supérieure à 6 jours, impliquant au moins 10 cas épidémiques, et avec un rapport entre le nombre de cas observé et le nombre de cas attendu supérieur à 3.

Sur la base de ces études, un outil de détection automatisé a été élaboré pour faciliter l'implémentation de la méthode (Annexe).

1.3 Les critères de sélection des signaux détectés sur la base du profil des épidémies d'origine hydrique

Bien que la gastro-entérite aigüe médicalisée (Assurance Maladie) soit un indicateur sanitaire pertinent pour surveiller les épidémies d'infections liées à l'eau du robinet car elle représente le symptôme le plus fréquemment identifié dans ces épidémies et qu'elle bénéficie d'un niveau d'agrégation fin (au jour et à la commune) permettant la détection de cas groupés localisés, elle n'est pas spécifique de ce risque et d'autres modes de contamination sont possibles (principalement les épidémies hivernales et les intoxications alimentaires principalement).

La méthode de détection automatisée des épidémies de gastro-entérite aigüe d'origine hydrique développée dans le cadre de nos travaux de recherche permet de contrôler les épidémies hivernales et prend en compte l'exposition à l'eau du robinet. En pratique, la méthode permet de remplir les caractéristiques 2 à 4 proposées par Poullis (Poullis 2005) (Figure 10, page 40) : respectivement apparition soudaine de malades, augmentation rapide de cas avec des symptômes similaires et cas localisés dans une zone desservie par un même réseau d'eau.

Concernant le contrôle des intoxications alimentaires, la typologie des épidémies associées permet d'identifier des critères d'exclusion potentiels comme l'impact, la durée et l'emprise géographique (Tableau 10). En effet, les toxi-infections alimentaires collectives (Tiac), dont l'emprise géographique est

compatible avec celle d'un réseau de distribution d'eau potable, représentent le principal risque de confusion possible avec des épidémies hydriques. Le nombre médian de malades par foyer de Tiac qui se situait entre 4 et 5 cas en 2006-2008 (Delmas 2010), montre que la probabilité de confusion est nulle pour la grande majorité des Tiac mais existe pour les Tiac localisées de grande taille.

Tableau 10 : Typologie des épidémies de gastro-entérite aigüe survenant en population générale en France

Origine	Epidémie hivernale	Epidémie d'origine hydrique	Intoxication d'origine alimentaire
Mode de transmission	Inter-humaine	Eau du robinet	Aliments
Ordre de grandeur de la durée	12 à 16 semaines	1 semaine à 1 mois	1 jour à 1 semaine
Ordre de grandeur du taux d'attaque	-	20 à 50%	20 à 50%
Ordre de grandeur de l'impact sanitaire (nombre de cas)	Entre 5.10^6 et 10^7 cas/an	10^1 à 10^3 cas/épidémie	2 à 10 cas/intoxication
Type d'agent pathogène le plus couramment rencontré	Virus entériques	Virus, bactéries, parasites	Bactéries, virus, toxines
Ordre de grandeur de l'emprise géographique	Locale, nationale, internationale	Localisée sur la zone desservie par le réseau d'eau contaminé	Locale, nationale ou internationale

Ainsi, la prise en compte de l'épidémie hivernale et de l'exposition à l'eau du robinet augmente significativement la capacité de la méthode de détection des épidémies de gastro-entérite aigüe hydriques. Elle a montré sa pertinence pour améliorer l'identification des agrégats de cas de gastro-entérite aigüe qui partagent en commun une même unité de distribution d'eau. Pour autant, l'appartenance à un même réseau d'eau n'est pas le seul critère à prendre en compte pour conforter l'existence d'une épidémie d'origine hydrique.

Des critères de sélection supplémentaires, fondés sur la durée du signal et sur le nombre de cas, doivent être appliqués pour améliorer la probabilité que les agrégats détectés (partageant la même unité de distribution d'eau), correspondent effectivement à une épidémie hydrique :

- Une durée du signal de plus de 3 jours : d'après les résultats de l'étude de simulation dans laquelle les épidémies simulées avaient une durée de 3 jours ou plus. Par ailleurs, les Tiac ont dans la plupart des cas une durée inférieure à 3 jours ;

- Un nombre de cas en excès (valeur observée – valeur attendue) supérieur à 10 : seuil définit d'après l'étude de simulation. Par ailleurs, la médiane du nombre de cas dans les Tiac se situe autour de 5 cas.

Un critère basé sur le rapport (valeur observée) / (valeur attendue) pourrait être discuté si on veut se concentrer uniquement sur le risque épidémique (rapport au-dessus de 2 par exemple).

Enfin, d'autres critères caractéristiques des épidémies hydriques, comme par exemple la répartition des cas par classe d'âge (plus élevée chez les enfants), peuvent également être pris en compte.

1.4 Les investigations complémentaires des signaux détectés pour conforter l'origine hydrique

Pour les signaux retenus sur des critères statistiques et épidémiologiques, des investigations complémentaires doivent être mise en place pour conforter l'origine hydrique de l'épidémie : i) en vérifiant la cohérence géographique entre le contour des cas groupés de gastro-entérite aigüe (contour administratif des communes hébergeant les cas) et le contour de la ou des unités de distribution d'eau ciblée(s) ; ii) en cherchant des indicateurs environnementaux qui pourraient objectiver une contamination microbiologique de l'unité de distribution d'eau ciblée, les jours précédant le signal ; et iii) en cherchant une répétition dans le temps d'épisodes de cas groupés de gastro-entérite aigüe impliquant un même réseau de distribution d'eau. Ce dernier cas peut s'observer dans une situation de pollution chronique d'un réseau d'eau uniquement.

Les indicateurs environnementaux pouvant témoigner de l'existence d'une pollution microbiologique du réseau d'eau sont de différentes natures :

- Signaux microbiologiques : résultats de non-conformité des indicateurs de contamination fécale issus du contrôle sanitaire ou de l'autocontrôle par l'exploitant (résultats contemporains de l'épidémie ou à d'autres moments de façon répétée témoignant d'une pollution chronique du système de distribution d'eau) ;
- Données de précipitation : localisation et intensité, comparaison de l'intensité par rapport à la saison ;
- Incident au niveau de la ressource (inondation de captage, forte pollution de la ressource par mauvais fonctionnement d'une station d'épuration des eaux usées en amont de la prise d'eau par exemple) ;
- Incident au niveau de la station de traitement (panne de désinfection, incident de filtration) ;
- Incident au niveau du réseau de distribution (rupture de canalisation, travaux sur le réseau) ;

- Plaintes d'abonnés les jours précédant le signal.

Ainsi, à l'issue de ces étapes de détection, d'analyse et d'investigations environnementales, 4 des 5 caractéristiques retenues par Poullis (Poullis 2005) pour définir la détection d'une épidémies d'origine hydrique pourront être satisfaites (Figure 10, page 40) : la présence d'indicateurs de contamination de l'eau (caractéristique 1), associés à l'apparition soudaine et étendue de malades (caractéristique 2) ou à une augmentation rapide de cas (caractéristique 3) et/ou à la présence de cas de gastro-entérite aigüe localisés dans une zone desservie par un même réseau d'eau (et peu de cas dans les zones desservies par des réseaux d'eau adjacents) (caractéristique 4). La dernière caractéristique qui concerne la cohérence entre la localisation des cas et le tracé du réseau d'eau (caractéristique 5) n'est pas possible avec les données de l'Assurance Maladie en l'absence de l'adresse exacte des cas.

1.5 Les conditions d'application de la méthode de détection

1.5.1 Une détection rétrospective

Actuellement, les délais nécessaires pour disposer de données consolidées dans les bases de l'Assurance Maladie (2 mois environ) ne permettent pas d'envisager une détection en temps réel mais sont adaptés à la recherche d'épidémies de façon rétrospective. Sur ce point, les données de l'Assurance Maladie se distinguent des autres données de surveillance syndromique habituellement appréciées pour leur réactivité (Berger 2006). Ce manque de réactivité est compensé par l'exhaustivité des données, couvrant l'ensemble de la population française (territoires d'outre-mer inclus), et le niveau de précision géographique très fin (au code Insee). De ce fait, l'utilité de ces données pour la prévention des épidémies d'origine hydrique porte essentiellement sur l'identification des réseaux d'eau impactés et le cas échéant des facteurs de risque associés. Par ailleurs, elles pourraient également permettre d'estimer l'impact sanitaire du risque épidémique lié à l'eau du robinet. Cette estimation nécessiterait néanmoins des travaux ultérieurs approfondis prenant en compte les facteurs influençant le recours aux soins de la population pour la gastro-entérite aigüe dans un contexte d'épidémie d'origine hydrique (cf. 4. L'estimation de l'impact sanitaire : une question qui reste ouverte, p 147).

1.5.2 Influence de la qualité des données de la base SISE-eaux sur la méthode de détection

La méthode de détection des épidémies d'origine hydrique à partir des données de l'Assurance Maladie nécessite également d'utiliser des données de population dans la base de données SISE-eaux dont il convient de vérifier la qualité lors de l'implantation de la méthode. En effet, la méthode de détection exploite la variable « population quartier » dans SISE-eaux pour effectuer les associations de

communes desservies par une même unité de distribution. Une valeur nulle ou une donnée manquante sur cette variable entraîne l'exclusion des cas de gastro-entérite aigüe associés à cette commune lors de la recherche d'épidémies d'origine hydrique. Cette variable est remplie de façon hétérogène selon les départements. Par exemple, l'analyse pour les 4 départements de la région Auvergne montre qu'aucune valeur nulle ou donnée manquante n'est identifiée. Pour la région Midi-Pyrénées, la Haute-Garonne ne présente aucune valeur nulle ou donnée manquante alors que pour 82 communes des Hautes-Pyrénées, une valeur nulle ou manquante entraîne l'exclusion de près de 16% des cas de gastro-entérite aigüe médicalisés (Tableau 11).

Un projet en cours d'une base géographique nationale, a pour ambition d'améliorer l'homogénéité et la qualité des données sur l'eau potable, dont les données concernant le contour des unités de distribution dans SISE-eaux.

Tableau 11 : Qualité des données SISE-eaux et impact sur la détection automatisée des épidémies de gastro-entérite aigüe hydrique – Ariège, Haute-Garonne, Lot, Hautes-Pyrénées, 2012-2014.

Départements	Ariège		Haute-Garonne		Lot		Hautes-Pyrénées		4 départements	
Population exclue de l'analyse*										
Nb communes exclues**	8		0		4		82		94	
Pop exclue (nombre et %)	2796	1,9%	0	0,0%	988	0,6%	36068	15,7%	39852	2,2%
Cas de GEA exclus de l'analyse***										
GEA 2012 exclus (nombre et %)	283	1,8%	0	0,0%	53	0,4%	3650	15,7%	3986	1,9%
GEA 2013 exclus (nombre et %)	320	2,0%	0	0,0%	42	0,3%	3381	15,7%	3743	2,0%
GEA 2014 exclus (nombre et %)	337	2,3%	0	0,0%	36	0,3%	2980	15,5%	3353	1,9%

* source : SISE-eaux, Ministère de la Santé

** si la variable SISE-eaux "pop quartier" = 0 ou manquant

*** source : InVS-DSE à partir de données de l'Assurance Maladie, SNIIRAM

2 Perspectives pour la surveillance des épidémies de gastro-entérite aigüe d'origine hydrique

2.1 Objectifs et périmètre de la surveillance

Les objectifs de la surveillance des épidémies de gastro-entérite aigüe d'origine hydrique à partir des données de l'Assurance Maladie sont (i) de détecter *a posteriori* à l'échelle départementale les agrégats de cas de gastro-entérite aigüe partageant un même réseau d'eau potable ; (ii) de mener des investigations sur les unités de distribution d'eau qui desservent les agrégats de cas identifiés pour rechercher la présence de facteurs environnementaux pouvant être à l'origine d'une contamination microbiologique du réseau d'eau (confirmation de l'origine hydrique) ; et (iii) *in fine* de renseigner une base de données nationale avec des informations épidémiologiques et environnementales pour chaque

agrégats de cas de gastro-entérite aigüe identifié. Des analyses ultérieures pourront être réalisées à partir de cette base de données dans un objectif de description des facteurs de risque, de bilan épidémiologique, de rétro-information et de prévention.

La mise en œuvre de la surveillance nécessite l'implication d'un ensemble de partenaires ayant chacun leur domaine de compétence : Santé publique France pour la détection des épidémies et l'évaluation des mesures de prévention, les autorités sanitaires (Agences régionales de santé) et les exploitants pour les enquêtes environnementales et la mise en œuvre de mesures correctives et préventives.

2.2 Mise en œuvre opérationnelle de la surveillance

2.2.1 Présentation de l'outil « EpiGEH » pour détecter les épidémies d'origine hydrique

La surveillance des épidémies de gastro-entérite aigüe d'origine hydrique nécessite des compétences en épidémiologie, en gestion de bases de données et en statistique. Elle relève des missions de Santé publique France qui a notamment en charge la surveillance de l'état de santé de la population et de ses déterminants.

D'après les résultats de l'évaluation de la méthode de détection développée dans les travaux de cette thèse qui ont montré qu'elle avait à la fois une bonne sensibilité et une bonne spécificité (cf article 3, p. 103) ; une application nommée « EpiGEH », a été développée dans le cadre de cette thèse pour faciliter la mise en œuvre de la détection des Epidémies de Gastro-Entérite aigüe d'origine Hydrique. Testée dans les régions Auvergne et Midi-Pyrénées, cette application est adaptable à l'ensemble des régions de France, et pourrait, à terme, être utile au déploiement de la détection à l'échelle nationale.

Les traitements informatiques ont été automatisés de l'intégration des données à la production des résultats à l'aide des logiciels R (v3.2.2.) et SatScan (9.3.1).

L'utilisation de l'outil a été facilitée grâce à l'adoption d'une interface développée à l'aide du package Shiny pour le logiciel R ne nécessitant aucune programmation supplémentaire de la part de l'utilisateur (interface de paramétrage et de consultation nécessitant un navigateur Internet) (Figure 16).

Figure 16 : Interface de l'application EpiGEH

2.2.1.1 Données pré-requises

L'utilisation de l'outil requiert pour chaque analyse la récupération de données spécifiques concernant l'indicateur sanitaire, l'indicateur d'exposition, les paramètres et les critères d'analyse. La récupération et la mise en forme de ces données constitue une étape préalable indispensable à la réalisation des analyses.

2.2.1.1.1 Indicateur sanitaire

Il s'agit du détail, sur une période donnée, des cas de gastro-entérite aigüe médicalisés au niveau communal et précisant le nombre de cas résidents attachés à chaque commune pour chaque jour de la période d'étude (source : données Santé publique France-DSE à partir des données de l'Assurance Maladie, Sniir-AM).

2.2.1.1.2 Indicateur d'exposition

Il s'agit des caractéristiques des unités de distribution (UDI) disponibles dans la base nationale SISE-eaux.

Ces données détaillent, pour chaque commune de la zone d'étude les unités de distribution qui les alimentent, ce détail allant jusqu'à l'échelle la plus fine disponible c'est à dire au niveau des quartiers ou

hameaux desservis. Elles servent à définir des couples UDI-communes, notamment en termes de poids de population représentée, grâce à l'algorithme développé dans l'article 2 (p74).

2.2.1.2 Options pour l'analyse

2.2.1.2.1 La prise en compte, dans la détection, d'épidémies liées à une pollution diffuse ou localisée des réseaux d'eau : les regroupements figés ou flexibles

- Test sur des regroupements figés de communes desservies par une même UDI

La méthode d'origine telle que publiée dans l'article 2 (p. 74) effectue l'analyse spatio-temporelle de Kulldorff à partir des regroupements de communes définis par l'algorithme selon l'adéquation UDI-communes en figeant le contour de ces regroupements.

L'unité écologique d'analyse pour la détection spatio-temporelle est alors le regroupement figé de communes définis par l'algorithme. La zone témoin est définie comme l'ensemble des autres regroupements.

Cette option est particulièrement bien adaptée en cas de pollution d'eau touchant l'ensemble des habitations desservies par une UDI (cas d'une pollution venant de la ressource par exemple). Ses capacités de détection peuvent en revanche être diminuées en cas de pollution touchant une partie du réseau d'eau (cas d'un retour d'eau usée dans le réseau d'eau potable par exemple).

- Test sur des regroupements flexibles de communes desservies par une même UDI

Afin d'optimiser la méthode de détection, en particulier pour les pollutions d'eau touchant une partie du réseau d'eau, la « matrice de voisinage » ou « table des voisins » disponible dans le logiciel Satscan a été également exploitée. Elle permet de cumuler progressivement dans l'analyse l'ensemble des communes d'un regroupement défini par l'algorithme selon l'adéquation UDI-communes. Ainsi, pour chaque regroupement de communes partageant une même unité de distribution d'eau, les communes « liées » ont été inscrites dans la matrice de voisinage. Cette option d'analyse a été utilisée dans l'étude de simulation pour évaluer les capacités de détection de la méthode développée (article 3, p. 103).

L'unité écologique d'analyse pour la détection spatio-temporelle est alors la commune au sein de chaque regroupement. La zone témoin étant les communes de la zone d'étude, situées en dehors du regroupement de communes desservis par une même UDI faisant l'objet du test.

Cette variante présente l'avantage pour la détection d'intégrer progressivement les communes « liées » et de tester toutes les associations possibles de communes qui partagent un même réseau de

distribution d'eau. L'agrégat le plus probable identifié pourra ainsi correspondre à tout ou partie des communes desservies par une même unité de distribution d'eau. Il est alors possible, grâce à cette variante, d'améliorer les capacités de détection des épidémies d'origine hydrique localisées sur une partie du réseau d'eau comme c'est le cas lorsque le point d'entrée de la pollution est situé sur le réseau d'eau (retours d'eaux usées d'une station d'épuration par exemple) et non au niveau de la ressource.

Les deux options (regroupements figés ou flexibles de communes partageant une même unité de distribution) sont implémentées dans l'outil EpiGEH.

2.2.1.2.2 La prise en compte, dans la détection, d'épidémies répétées sur un même réseau d'eau : les boucles d'analyse

La méthode de permutation spatio-temporelle détecte et classe entre eux les signaux en fonction de leur significativité statistique (Kulldorff *et al.*, 2005). La survenue de plusieurs signaux (épidémies) sur un même réseau d'eau peut alors se traduire par l'identification exclusive du signal le plus important, sa significativité étant par ailleurs diminuée en raison de la fréquence du phénomène. Cette méthode n'est donc pas complètement adaptée à l'identification d'épidémies répétées, sur un même réseau d'eau, à plusieurs périodes de l'année, ce qui peut se produire en cas de pollutions chroniques par exemple.

Pour pallier ce phénomène de sélection de l'agrégat le plus significatif, la possibilité de répéter les analyses plusieurs fois a été intégrée (principe de création de « boucles ») en supprimant à chaque boucle l'agrégat le plus significatif et en remplaçant les effectifs de cas de gastro-entérite aigüe par la valeur médiane de la série. La détection peut ainsi être reproduite pour identifier plusieurs signaux épidémiques pour un même réseau d'eau à des périodes différentes, jusqu'à ce qu'aucun agrégat ne ressorte. Ce processus qui réduit progressivement la taille de l'échantillon et introduit des tests multiples pourrait avoir un impact sur les résultats dès lors que le nombre de cas soustraits représenterait une proportion non négligeable de l'échantillon et que le nombre de tests serait important. En pratique, les agrégats concernés regroupent un faible nombre de cas (quelques dizaines pour les plus importants) et le nombre de tests est limité.

2.2.1.3 Rendus des résultats

L'outil fournit différents supports synthétiques de résultats :

- Un tableau des caractéristiques des agrégats potentiels identifiés comprenant différentes variables (Tableau 12) ;
- Un rapport sous forme de texte résumant les précédents résultats (Tableau 13);

- Une cartographie représentant le nombre de signaux par commune et indiquant la (les) commune(s) concernée(s) par le plus grand nombre de signaux sur la période d'étude (Figure 17) ;
- Une cartographie du nombre total de cas de gastro-entérite aigüe en excès estimé par l'analyse (nombre maximal de cas estimés au niveau communal sur la période d'étude) (Figure 18).

Pour chaque agrégat potentiel identifié, l'outil génère également :

- Une carte de localisation de la zone concernée ;
- Des figures représentant les séries chronologiques centrées sur la période de survenue des cas de gastro-entérite aigüe (Tableau 14 et Tableau 15).

Tableau 12 : Description des périodes épidémiques et zones géographiques identifiables par la méthode de détection, résidents du département, tous âges, par ordre chronologique – Pyrénées – orientales, année 2014.

Référence de l'analyse		Zones géographiques concernée(s)				Caractéristiques du signal						
Boucle d'analyse	Numéro de signal	Caractéristiques de(des) UDI identifiée(s)			Commune(s) concernée(s)*	Durée (jours)	Début	Fin	Nb de cas observés	Nb de cas en excès	Rapport de risque**	p-value
		Numéro	Identifiant	Nom(s)								
1	10	2	66000026	VILLELONGUE DE LA SALANQUE	VILLELONGUE DE LA SALANQUE (66224)	8	20/02/2014	27/02/2014	28	19	3,02	1,30E-02
1	1	138	66000747	CONFLENT 1	EUS (66074), PRADES (66149)	6	28/03/2014	02/04/2014	48	38	4,85	4,00E-15
1	2	180	66002179	CONFLENT 3	PRADES (66149)	6	28/03/2014	02/04/2014	46	36	4,81	3,02E-14
3	1	49	66000288	LATOUR DE FRANCE	LATOUR DE FRANCE (66096)	7	29/03/2014	04/04/2014	15	13	8,24	8,33E-06
30	1	20	66000140	CABESTANY	CABESTANY (66028)	31	14/04/2014	14/05/2014	39	27	3,16	1,60E-05
6	1	7	66000065, 66000067	TORREILLES VILLAGE, TORREILLES PLAGE	TORREILLES (66212)	7	17/06/2014	23/06/2014	22	17	4,06	8,01E-04
31	1	20	66000140	CABESTANY	CABESTANY (66028)	35	01/07/2014	04/08/2014	36	28	4,29	5,96E-09
9	1	173	66000972	BANYULS VAL AUGER	BANYULS SUR MER (66016)	4	05/11/2014	08/11/2014	16	13	4,82	1,00E-02
11	1	79	66000443	RIA SIRACH	RIA SIRACH (66161)	3	17/12/2014	19/12/2014	12	10	5,73	3,50E-02
2	2	138	66000747	CONFLENT 1	EUS (66074), PRADES (66149)	7	17/12/2014	23/12/2014	56	41	3,70	8,34E-13
2	1	180	66002179	CONFLENT 3	PRADES (66149)	6	18/12/2014	23/12/2014	52	39	4,05	2,13E-13

* Communes desservies par les UDI détectées

** Rapport de risque estimé de la façon suivante : valeur observée / valeur attendue

Tableau 13 : Illustration d'un rapport sous forme de texte décrivant résumant les résultats des analyses - Ariège, Haute-Garonne, Lot et Hautes-Pyrénées – année 2014

N° signal	Interprétation commentée des résultats
1	Le signal 1 concerne le regroupement 352. Il touche la commune de SAINT ORENS DE GAMEVILLE (31506). L UDI en cause est CUTM ST ORENS MTGNE NOIRE P DAV (n°31003025). Le signal relevé couvre une période de 3 jours, allant du mercredi 10 septembre au vendredi 12 septembre 2014. Sur cette période, 47 cas de GEA sont observés, contre 9.65 cas attendus. Le rapport de risque est donc de 4.87 et l'excès de cas mesuré est de 37. La statistique de test associée à ce signal vaut 37.07224, la p-value associée valant 2.18e-13.
2	Le signal 2 concerne le regroupement 384. Il touche la commune de GOURDON (46127). L UDI en cause est GOURDON (n°46000421). Le signal relevé couvre une période de 12 jours, allant du lundi 24 novembre au vendredi 05 décembre 2014. Sur cette période, 46 cas de GEA sont observés, contre 14.86 cas attendus. Le rapport de risque est donc de 3.1 et l' excès de cas mesuré est de 31. La statistique de test associée à ce signal vaut 20.850929, la p-value associée valant 8.22e-06.
3	Le signal 3 concerne le regroupement 215. Il touche la commune d'AUTERIVE (31033). L UDI en cause est AUTERIVE (n°31000042). Le signal relevé couvre une période de 26 jours, allant du dimanche 19 janvier au jeudi 13 février 2014. Sur cette période, 227 cas de GEA sont observés, contre 143.61 cas attendus. Le rapport de risque est donc de 1.58 et l'excès de cas mesuré est de 83. La statistique de test associée à ce signal vaut 20.562545, la p-value associée valant 1.12e-05.
4	Le signal 4 concerne le regroupement 214. Il touche les communes de MONESTROL (31002), SEYRE (31024), BEAUTEVILLE (31054), MONTCLAR LAURAGAIS (31099), MAUVAISIN (31100), LABRUYERE DORSA (31145), AIGNES (31210), CAIGNAC (31220), GIBEL (31233), LAGARDE (31256), VIEILLEVIGNE (31262), AURAGNE (31332), MONTGEARD (31354), RENNEVILLE (31368), MONTESQUIEU LAURAGAIS (31374), GREPIAC (31380), SAINT LEON (31396), GARDOUCH (31450), CALMONT (31495), CINTEGABELLE (31546) et NAILLOUX (31576). L UDI en cause est HERS ARIEGE (n°31000039). Le signal relevé couvre une période de 36 jours, allant du mercredi 15 janvier au mercredi 19 février 2014. Sur cette période, 476 cas de GEA sont observés, contre 355.86 cas attendus. Le rapport de risque est donc de 1.34 et l'excès de cas mesuré est de 120. La statistique de test associée à ce signal vaut 18.362048, la p-value associée valant 0.00012.
…	…

GEA : gastro-entérite aigüe

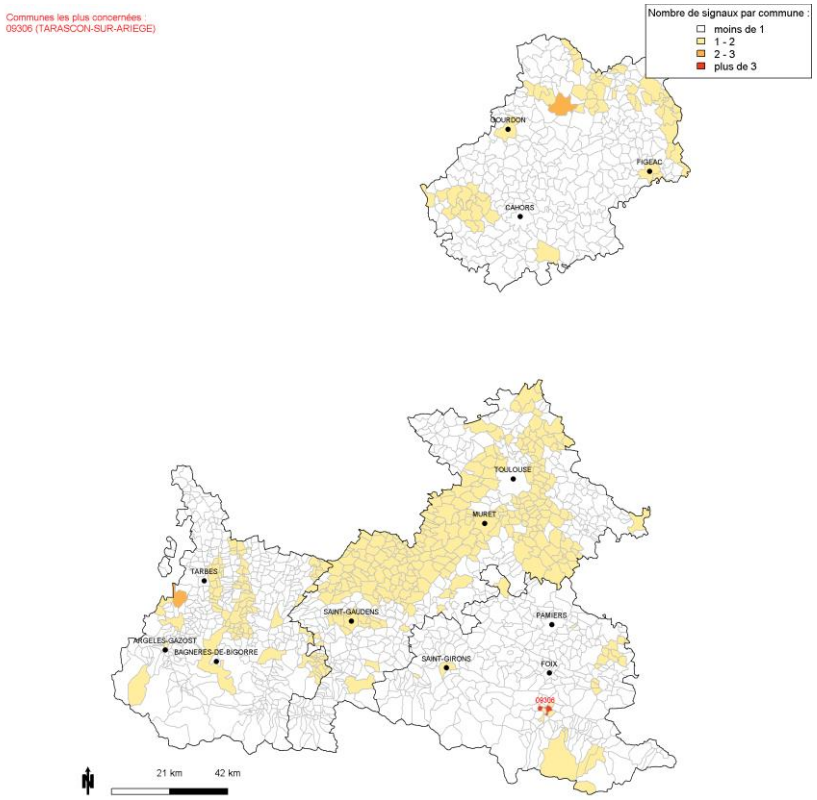


Figure 17 : Nombre de signaux par commune - Ariège, Haute-Garonne, Lot et Hautes-Pyrénées – année 2014

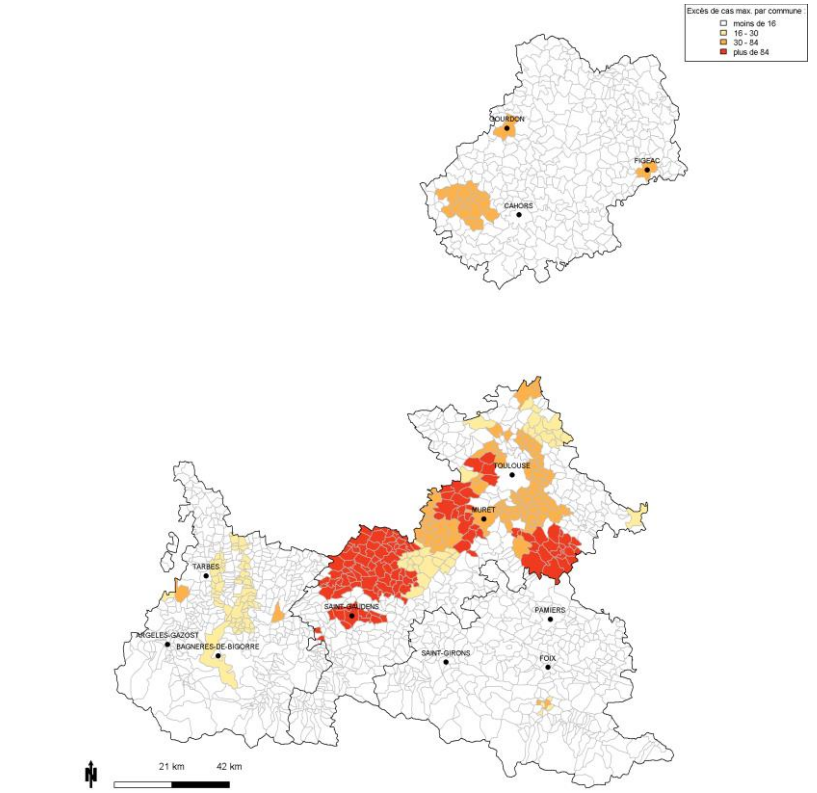


Figure 18 : Nombre de cas en excès estimés par l'analyse - Ariège, Haute-Garonne, Lot et Hautes-Pyrénées – année 2014

Tableau 14 : Localisation des cas groupés de cas de gastro-entérite aigüe (GEA) et distribution quotidienne des cas résidents, tous âges, par ordre chronologique sur le premier semestre de 2014 – Pyrénées – orientales.

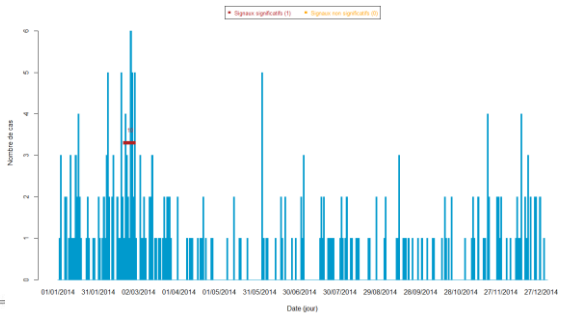
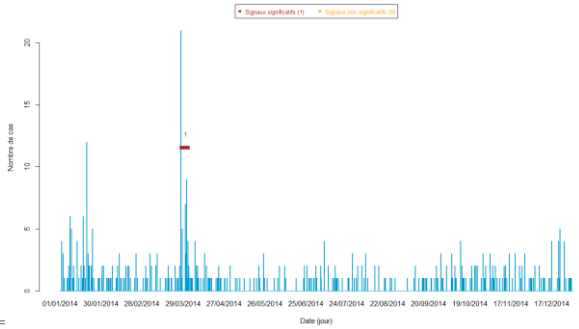
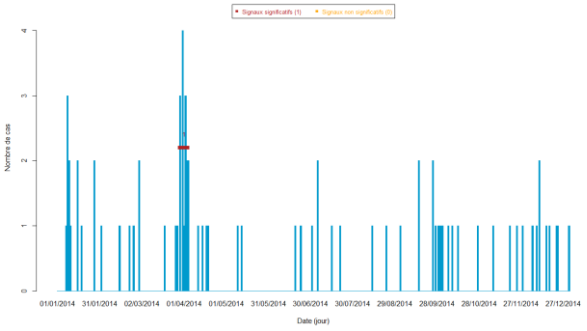
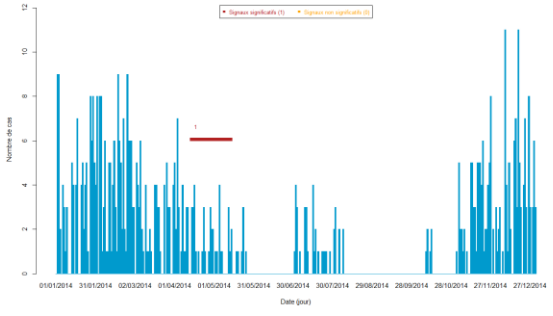
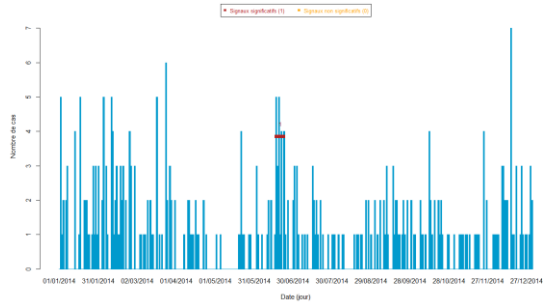
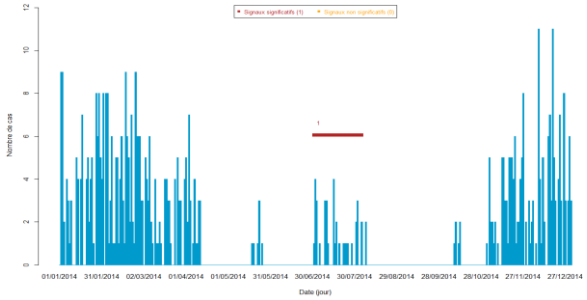
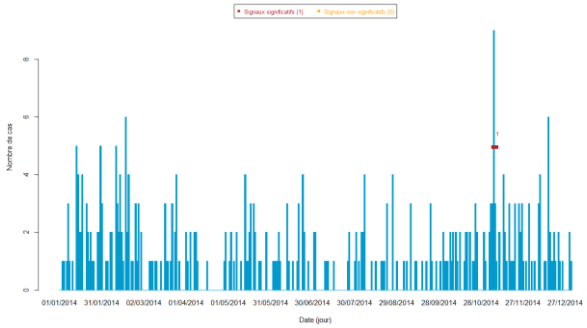
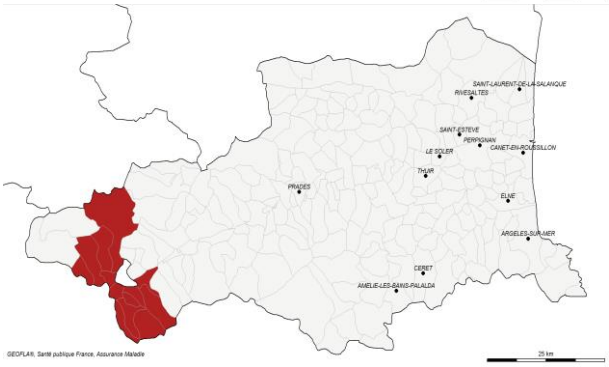
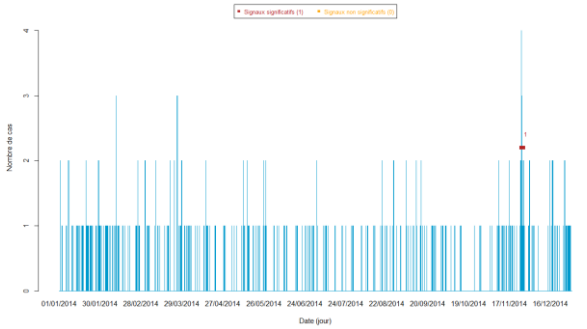
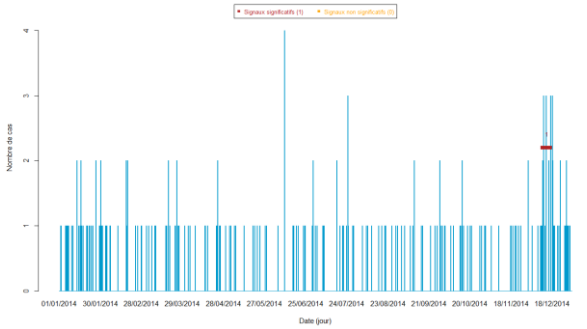
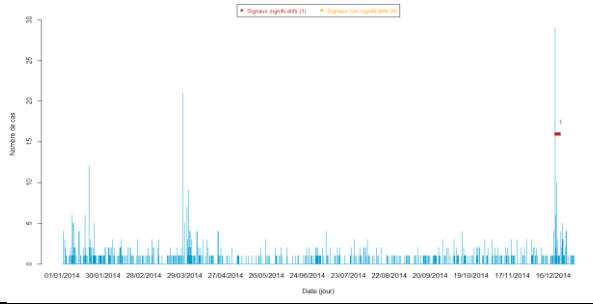
Périodes épidémiques	Commune(s) potentiellement impactée(s)	Distribution temporelle des cas de GEA utilisée pour la détection
20 au 27 février		
28 mars au 2 avril		
29 mars au 4 avril		
14 avril au 14 mai		
17 au 23 juin		

Tableau 15 - Localisation des cas groupés de cas de gastro-entérite aigüe (GEA) et distribution quotidienne des cas résidents, tous âges, par ordre chronologique sur le deuxième semestre de 2014 – Pyrénées – orientales

Périodes épidémiques	Commune(s) potentiellement impactée(s)	Distribution temporelle des cas de GEA utilisée pour la détection
1er juillet au 4 août		
5 au 8 novembre		
25 au 27 novembre		
10 au 16 décembre		
17 au 23 décembre		

2.2.2 Mise en place de l'enquête environnementale : le rôle de l'autorité sanitaire et des exploitants

La détection d'épidémies de gastro-entérite aigüe dont l'origine hydrique est suspectée, permet d'établir une liste d'unités de distribution d'eau pour lesquelles des enquêtes environnementales, ciblées sur la période des épidémies détectées, sont nécessaires. La mise à disposition de la liste des unités de distribution d'eau à investiguer et des épidémies d'origine hydrique suspectées est portée au niveau local par Santé publique France et constitue le point de départ des enquêtes environnementales.

Ces enquêtes viseront à conforter l'origine hydrique des épidémies détectées en identifiant des indicateurs directs ou indirects de contaminations microbiologiques des réseaux d'eau. Elles viseront également à comprendre les circonstances dans lesquelles ces contaminations sont survenues.

Plusieurs sources d'information pourront être utiles à ces enquêtes : données du contrôle sanitaire de l''eau destinée à la consommation humaine (base SISE-eaux), données d'autocontrôle (SISE-eaux et exploitant), données de pluviométrie (météo France), plans des réseaux d'eau, cahier des plaintes (exploitant), liste des travaux effectués (exploitant), liste des incidents d'exploitation (exploitant), carte des stations d'épuration, visite de terrain, etc..

La réalisation de ces enquêtes environnementales nécessite une implication et une interaction forte au niveau local de l'autorité sanitaire locale en charge du contrôle sanitaire des eaux destinées à la consommation humaine (Agences régionales de santé) et de l'exploitant. Elle doit se conclure par la transmission à Santé publique France, pour chaque unité de distribution d'eau et épidémie investiguées, d'un ensemble d'informations homogènes et standardisées.

2.2.3 La phase pilote : une étape nécessaire avant le déploiement de la surveillance à l'échelle nationale

Le déploiement de la surveillance à l'échelle nationale nécessite au préalable de confronter la méthode développée pour la détection d'épidémies de gastro-entérite aigüe et l'approche proposée pour la réalisation d'enquêtes environnementales, à la réalité du terrain dans des zones pilotes. Une étude de faisabilité, pilotée au niveau national, a été initiée en ce sens dans plusieurs régions de France, incluant l'Auvergne et l'Occitanie. Cette étude permettra notamment de tester la méthode de détection sur des données d'autres régions n'ayant pas été utilisées pour sa construction ; de comparer les performances de la méthode de détection, obtenues à partir une l'étude de simulation (article 3, p. 103), aux données

de terrain ; d'établir un protocole pour l'enquête environnementale ; de définir les modalités de rétro-information vers l'autorité sanitaire et les exploitants.

3 Synthèse : représentation schématique du dispositif de surveillance des épidémies de gastro-entérite aigüe d'origine hydrique

En se basant sur les travaux antérieurs puis sur nos travaux, il est possible de proposer le schéma d'un dispositif de surveillance des épidémies de gastro-entérite aigüe depuis la récupération des données nécessaires à la mise en place des actions de prévention (Figure 19). Pour atteindre sa finalité de réduction du risque infectieux lié à l'eau du robinet, ce dispositif nécessite une sensibilisation et une mobilisation de l'ensemble des acteurs concernés (épidémiologistes, autorités sanitaires, exploitants). En effet, bien que les données requises (Assurance Maladie et SISE-eaux) soient directement accessibles, et que le processus de détection soit automatisé, la réalisation des enquêtes environnementales pour conforter l'origine hydrique et identifier les causes de la pollution va mobiliser des moyens humains. Néanmoins, on peut postuler que plusieurs éléments discutés précédemment interviendront pour en faire un outil opérationnel : peu de faux positifs, un faible nombre d'alertes gérables par les moyens des ARS et un nombre suffisant de données de terrain pour étayer les suspicions.

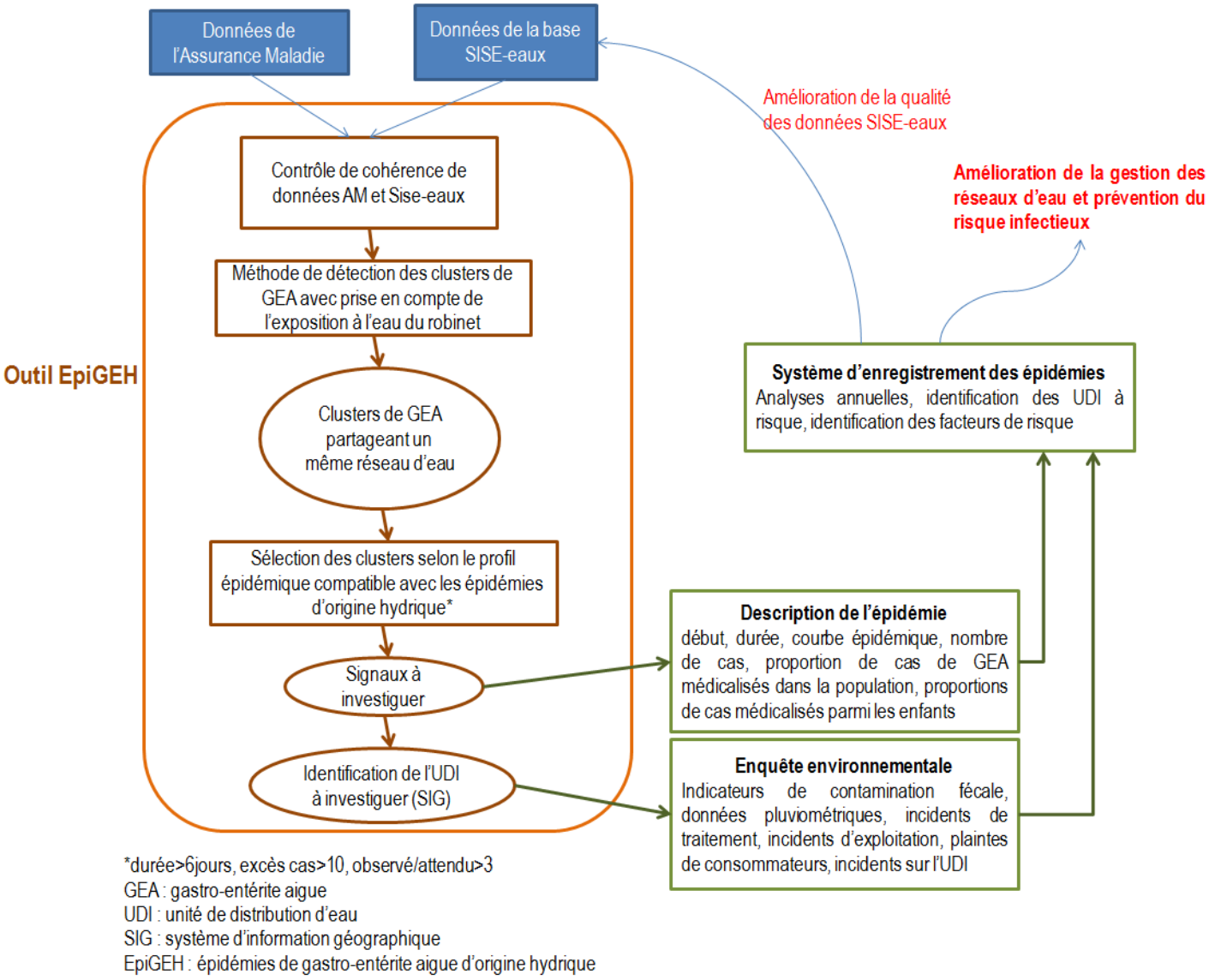


Figure 19 : Schéma du dispositif de surveillance des épidémies de gastro-entérite aigüe d'origine hydrique à partir des données de l'Assurance Maladie et des données de la base SISE-eaux.

4 L'estimation de l'impact sanitaire : une question qui reste ouverte

L'analyse des profils des épidémies de gastro-entérite aigüe d'origine hydrique montre qu'il y a des différences possibles entre les caractéristiques du signal détecté (variation du taux d'incidence et durée du signal) et les caractéristiques de l'épidémie réelle (taux d'attaque, durée de l'épidémie). Ces différences peuvent s'expliquer d'une part par la sensibilité de l'indicateur Assurance Maladie qui varie en fonction des caractéristiques des épidémies et du contexte local : nature de l'agent pathogène, les habitudes de consommation de l'eau de la population lorsque le risque de pollution est connu dans une commune et les habitudes de consommation de soins en fonction des facilités d'accès. D'autre part par la méthode de détection elle-même qui ne cherche pas à détecter la totalité d'une épidémie d'origine hydrique mais le signal le plus significatif sur le plan statistique et dont l'emprise géographique est compatible avec le contour d'une unité de distribution d'eau.

Ainsi, une fois que l'origine hydrique d'un signal a pu être appuyée par des indicateurs environnementaux (cf. partie « Les critères de sélection des signaux détectés », page 129), il est nécessaire de représenter la série temporelle du signal détecté à partir des données brutes pour améliorer l'estimation de l'impact de l'épidémie. Cette représentation permettra de définir une durée de l'épidémie et la proportion de cas de gastro-entérite aigüe médicalisés dans la population touchée. Il ne sera en revanche pas possible, sans mener des investigations supplémentaires dans la population touchée, d'estimer avec précision le taux de redressement à opérer sur les données de l'Assurance Maladie pour estimer le nombre de cas réel. Des hypothèses sur le taux de consultations médicales suivies d'un achat à la pharmacie donneront un intervalle de valeurs possibles.

Si une estimation précise de l'impact sanitaire en nombre de cas paraît difficile à envisager pour les raisons expliquées ci-dessus, la connaissance des performances de la méthode de détection devrait permettre d'exprimer l'impact sanitaire « agrégé » des épidémies de gastro-entérite aigüe d'origine hydrique en nombre d'épidémies.

Au regard des objectifs de la surveillance, la détection des épidémies de gastro-entérite aigüe d'origine hydrique, l'identification des unités de distribution d'eau à risque et des circonstances de la contamination, paraissent suffisants pour permettre des actions de santé publique de réduction des risques et de prévention de ces évènements.

5 Comparaison internationale : un exercice limité

Concernant le risque épidémique de gastro-entérite aigüe d'origine hydrique, la France ne se distingue pas des pays développés quant aux circonstances de survenues (cf. « Les épidémies liées à l'eau du robinet : synthèse bibliographique », page 19). En revanche, l'existence concomitante de deux bases de données nationale exploitables, contenant pour l'Assurance Maladie les informations sur les prestations médicales remboursées et pour SISE-eaux les informations sur la distribution de l'eau, est spécifique à la France. Cette spécificité limite les possibilités de comparaison des résultats obtenus dans le cadre nos travaux. Les sources de données de surveillance syndromique présentes dans d'autres pays sont principalement les ventes de médicaments au comptoir des pharmacies, les lignes d'appels téléphoniques d'urgence et l'absentéisme scolaire ou professionnel (Berger 2006).

Plusieurs pays ont testé et évalué l'utilisation des données de surveillance syndromiques pour décrire les épidémies d'origine hydrique de façon rétrospective sur des évènements connus (Proctor 1998; Edge 2004; Berger 2006; Smith 2010; Andersson 2014).

Ces analyses ont montré que les données de ventes de médicaments en pharmacie ou les données des appels téléphoniques pouvaient fournir des informations permettant de détecter le début d'une épidémie de façon plus précoce que les systèmes de surveillance traditionnels (Proctor 1998; Smith 2010).

Bien que plus réactives que les systèmes de surveillance traditionnels, les principales limites identifiées pour les données de surveillance syndromiques sont leur manque de spécificité vis-à-vis de la maladie (pour les données de ventes de médicaments ou absentéisme scolaire) et l'absence d'information géographique suffisamment précise pour permettre une localisation des cas cohérente avec le contour des réseaux d'eau potable (Figure 20). Enfin, la sensibilité des données de surveillance syndromique sera variable en fonction de la source de données utilisée, des habitudes de consommation des soins des populations et des facilité d'accès aux structures de soins.

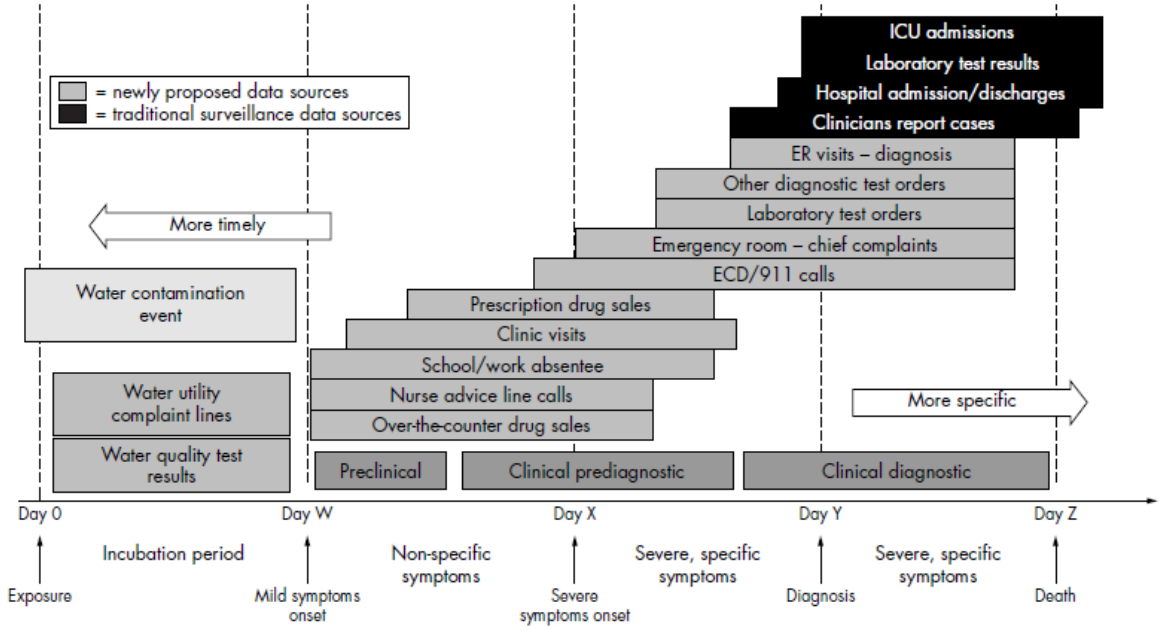


Figure 20 : Sources de données de surveillance syndromique et traditionnelles (source : (Berger 2006))

Compte tenu de l'hétérogénéité des sources de données utilisées entre les pays et des possibilités d'accès aux soins, certains auteurs recommandent d'étudier et d'évaluer davantage les bénéfices de la surveillance syndromique pour surveiller les épidémies hydriques, au niveau de chaque pays (Risebro & Hunter 2007).

Conclusion générale

Les épidémies de gastro-entérite aigüe liées à l'eau du robinet demeurent un enjeu de santé publique au 21ème siècle dans les pays développés. Depuis des décennies, les pays ont élaboré des systèmes de surveillance pour détecter, étudier et prévenir ces évènements. La grande majorité des dispositifs reposent sur la déclaration des épisodes de cas groupés de maladies (le plus souvent de type gastro-entérite aigüe) par les professionnels de santé ou les responsables d'établissements recevant du publique, et la déclaration d'un nombre anormalement élevé d'agents pathogènes responsables de symptômes digestifs identifiés par les laboratoires. Bien que ces dispositifs enregistrent un ensemble d'informations qui permettent d'étudier les principaux facteurs de risque des épidémies hydrique ainsi que les tendances épidémiologiques, le système de collecte, de type déclaratif, entraîne une hétérogénéité dans la qualité des données et une sous déclaration manifeste. Ces dernières années, l'apport de la surveillance syndromique a été considéré pour évaluer sa capacité à contribuer à l'étude et à la surveillance des épidémies d'origine hydrique.

Nos travaux ont permis de proposer une méthode pour améliorer la détection des épidémies de gastro-entérite aigüe d'origine hydrique. Cette méthode peut être mise en œuvre à un niveau départemental, régional ou national grâce au développement d'un outil automatisé. D'après les résultats d'une étude de simulation, cette méthode pourrait détecter près de 90% des épidémies d'origine hydrique en France supérieures à 10 cas de gastro-entérite aigüe médicalisés. Les signaux détectés bénéficieraient d'une valeur prédictive positive d'environ 90% vis-à-vis de l'origine hydrique. Les principales limites résident dans les délais de mise à disposition des données de l'Assurance Maladie qui imposent une détection rétrospective (avec un retard de quelques semaines), et dans la variabilité des comportements humains vis-à-vis du recours aux soins qui peuvent influencer les capacités de détection, dès lors qu'elle repose sur l'indicateur de gastro-entérite aigüe médicalisé.

Les investigations environnementales à mener autour des cas groupés de gastro-entérite aigüe partageant une même unité de distribution d'eau doivent répondre à un double objectif : identifier des indicateurs environnementaux pour conforter l'origine hydrique et identifier les circonstances d'introduction de la pollution dans le réseau d'eau dans un objectif de prévention. Il paraît enfin envisageable de classer les épidémies dont l'origine est suspectée en fonction de l'interprétation des données sanitaires et des données environnementales, à l'instar des critères proposés par les américains ou les gallois (cf. partie « Critères de classification », page 20).

En termes de santé publique, bien que le risque épidémique ne représente probablement pas la part la plus importante du risque infectieux lié à l'eau du robinet, l'amélioration de la détection des épidémies

d'origine hydrique peut entraîner une réduction du risque infectieux dans son ensemble. Cette détection contribue ainsi à améliorer la connaissance de l'impact sanitaire de ces évènements (objectif de surveillance) et inciter au renforcement des dispositifs de sécurisation et de maîtrise des systèmes de distribution d'eau vis-à-vis du risque infectieux dans son ensemble (objectif de prévention). La mise en place d'un dispositif de surveillance, incluant la méthode de détection développée, paraît envisageable à l'issu de ce travail.

Références bibliographiques

Addiss, D. G., Pond, R. S., Remshak, M., Juranek, D. D., Stokes, S. & Davis, J. P. 1996 Reduction of risk of watery diarrhea with point-of-use water filters during a massive outbreak of waterborne Cryptosporidium infection in Milwaukee, Wisconsin, 1993. *Am J Trop Med Hyg.* **54**, 549-553.

Ali, S. H. 2004 A socio-ecological autopsy of the E. coli O157:H7 outbreak in Walkerton, Ontario, Canada. *Soc Sci Med.* **58**, 2601-2612.

Andersson, T., Bjelkmar, P., Hulth, A., Lindh, J., Stenmark, S. & Widerstrom, M. 2014 Syndromic surveillance for local outbreak detection and awareness: evaluating outbreak signals of acute gastroenteritis in telephone triage, web-based queries and over-the-counter pharmacy sales. *Epidemiol Infect.* **142**, 303-313.

Assuncao, R. & T., C. 2009 Surveillance to detect emerging space-time clusters. *Computational Statistics & Data Analysis.* **53**, 2817-2830.

Auld, H., MacIver, D. & Klaassen, J. 2004 Heavy rainfall and waterborne disease outbreaks: the Walkerton example. *J Toxicol Environ Health A.* **67**, 1879-1887.

Balter, S., Weiss, D., Hanson, H., Reddy, V., Das, D. & Heffernan, R. 2005 Three years of emergency department gastrointestinal syndromic surveillance in New York City: what have we found? *MMWR Suppl.* **54**, 175-180.

Barboza, P., Vaillant, L., Le, S. Y., Hartley, D. M., Nelson, N. P., Mawudeku, A., Madoff, L. C., Linge, J. P., Collier, N., Brownstein, J. S. & Astagneau, P. 2014 Factors influencing performance of internet-based biosurveillance systems used in epidemic intelligence for early detection of infectious diseases outbreaks. *PLoS.One.* **9**, e90536.

Beaudeau, P., de Valk, H., Vaillant, V., Mannschott, C., Tillier, C., Mouly, D. & Ledrans, M. 2008 Lessons learned from ten investigations of waterborne gastroenteritis outbreaks, France, 1998-2006. *J Water Health.* **6**, 491-503.

Beaudeau,P. Surveillance syndromique des gastroentérites aigües : une opportunité pour la prévention du risque infectieux attribuable à l'ingestion d'eau du robinet. [Syndromic surveillance of acute gastroenteritis:an opportunity for the prevention of the infectious risk attributable to tap water] Thèse de doctorat. Université de Rennes 1,2012a.

Beaudeau, P., Le Tertre, A., Zeghnoun, A., Zanobetti, A. & Schwartz, J. 2012b A time series study of drug sales and turbidity of tap water in Le Havre, France. *J Water Health.* **10**, 221-235.

Beaudeau, P., Zeghnoun, A., Ledrans, M. & L., V. J. 2003 Consommation d'eau du robinet pour la boisson en France métropolitaine : résultats tirés de l'enquête alimentaire INCA1. *Environnement, Risques et Santé.* **2**, 147-158.

Beer, K. D., Gargano, J. W., Roberts, V. A., Hill, V. R., Garrison, L. E., Kutty, P. K., Hilborn, E. D., Wade, T. J., Fullerton, K. E. & Yoder, J. S. 2015 Surveillance for Waterborne Disease Outbreaks

Associated with Drinking Water - United States, 2011-2012. *MMWR Morb Mortal Wkly Rep.* **64**, 842-848.

Berger, M., Shiau, R. & Weintraub, J. M. 2006 Review of syndromic surveillance: implications for waterborne disease detection. *J Epidemiol Community Health.* **60**, 543-550.

Bounoure, F., Beaudeau, P., Mouly, D., Skiba, M. & Lahiani-Skiba, M. 2011 Syndromic surveillance of acute gastroenteritis based on drug consumption. *Epidemiol Infect.* **139**, 1388-1395.

Brown, R. S. & Hussain, M. 2003 The Walkerton tragedy --issues for water quality monitoring. *Analyst.* **128**, 320-322.

Buckeridge, D. L. 2007 Outbreak detection through automated surveillance: a review of the determinants of detection. *J.Biomed.Inform.* **40**, 370-379.

Buckeridge, D. L., Jauvin, C., Okhmatovskaia, A. & Verma, A. D. 2011 Simulation Analysis Platform (SnAP): a tool for evaluation of public health surveillance and disease control strategies. *AMIA Annu Symp Proc.* **2011**, 161-170.

Buckeridge, D. L., Okhmatovskaia, A., Tu, S., O'Connor, M., Nyulas, C. & Musen, M. A. 2008 Predicting outbreak detection in public health surveillance: quantitative analysis to enable evidence-based method selection. *AMIA Annu Symp Proc.* 76-80.

Canada, P. H. A. o. 2000 Waterborne outbreak of gastroenteritis associated with a contaminated municipal water supply, Walkerton, Ontario, May-June 2000. *Can Commun Dis Rep.* **26**, 170-173.

Caserio-Schonemann, C. & Meynard, J. B. 2015 Ten years experience of syndromic surveillance for civil and military public health, France, 2004-2014. *Euro Surveill.* **20**, 35-38.

Centers for Disease, C. 2001 Updated guidelines for evaluating public health surveillance systems: recommendations from the guidelines working group. *MMWR CDC Surveill Summ.* **50**, 1-35.

Cicirello, H. G., Kehl, K. S., Addiss, D. G., Chusid, M. J., Glass, R. I., Davis, J. P. & Havens, P. L. 1997 Cryptosporidiosis in children during a massive waterborne outbreak in Milwaukee, Wisconsin: clinical, laboratory and epidemiologic findings. *Epidemiol Infect.* **119**, 53-60.

Clark AB & Lawson AB (2006). Surveillance of individual level disease maps. *Stat Methods Med Res* **15**, 353-362

Clark, C. G., Bryden, L., Cuff, W. R., Johnson, P. L., Jamieson, F., Ciebin, B. & Wang, G. 2005 Use of the oxford multilocus sequence typing protocol and sequencing of the flagellin short variable region to characterize isolates from a large outbreak of waterborne Campylobacter sp. strains in Walkerton, Ontario, Canada. *J Clin Microbiol.* **43**, 2080-2091.

Clark, C. G., Price, L., Ahmed, R., Woodward, D. L., Melito, P. L., Rodgers, F. G., Jamieson, F., Ciebin, B., Li, A. & Ellis, A. 2003 Characterization of waterborne outbreak-associated Campylobacter jejuni, Walkerton, Ontario. *Emerg Infect Dis.* **9**, 1232-1241.

Colford, J. M., Jr., Roy, S., Beach, M. J., Hightower, A., Shaw, S. E. & Wade, T. J. 2006 A review of household drinking water intervention trials and an approach to the estimation of endemic waterborne gastroenteritis in the United States. *J Water Health.* **4 Suppl 2**, 71-88.

Coly, S., Vincent, N., Vaissiere, E., Charras-Garrido, M., Gallay, A., Ducrot, C. & Mouly, D. Detection of waterborne disease outbreaks: an integrated approach using health administrative databases. *J Water Health.* (accepté le 12 juin 2016, en cours d'édition)

Cooper, D. L., Smith, G. E., Hollyoak, V. A., Joseph, C. A., Johnson, L. & Chaloner, R. 2002 Use of NHS Direct calls for surveillance of influenza--a second year's experience. *Commun Dis Public Health.* **5**, 127-131.

Cordell, R. L., Thor, P. M., Addiss, D. G., Theurer, J., Lichterman, R., Ziliak, S. R., Juranek, D. D. & Davis, J. P. 1997 Impact of a massive waterborne cryptosporidiosis outbreak on child care facilities in metropolitan Milwaukee, Wisconsin. *Pediatr Infect Dis J.* **16**, 639-644.

Corso, P. S., Kramer, M. H., Blair, K. A., Addiss, D. G., Davis, J. P. & Haddix, A. C. 2003 Cost of illness in the 1993 waterborne Cryptosporidium outbreak, Milwaukee, Wisconsin. *Emerg Infect Dis.* **9**, 426-431.

Craun, G. F. 2012 The importance of waterborne disease outbreak surveillance in the United States. *Ann Ist Super Sanita.* **48**, 447-459.

Craun, G. F., Brunkard, J. M., Yoder, J. S., Roberts, V. A., Carpenter, J., Wade, T., Calderon, R. L., Roberts, J. M., Beach, M. J. & Roy, S. L. 2010 Causes of outbreaks associated with drinking water in the United States from 1971 to 2006. *Clin Microbiol Rev.* **23**, 507-528.

Craun, G. F. & McCabe, L. J. 1973 Review of the causes of waterborne-disease outbreaks. *J. Am. Water Works Assoc.* **65**, 74-84.

Das, D., Metzger, K., Heffernan, R., Balter, S., Weiss, D., Mostashari, F., New York City Department of, H. & Mental, H. 2005 Monitoring over-the-counter medication sales for early detection of disease outbreaks-New York City. *MMWR Morb Mortal Wkly Rep.* **54 Suppl**, 41-46.

Delmas, G., Jourdan Da Silva, N., Pihier, N., Weil, F. X., Vaillant, V. & De Valk, H. 2010 Les toxi-infections alimentaires collectives en France entre 2006 et 2008. *Bull Epidemio Hebdo.* **31-32**, 344-348.

Dematteï C. Detection d'agregats temporels et spatiaux. 1-170. 2006. Thèse de doctorat. Université de Montpellier 1.

Edge, V. L., Pollari, F., Lim, G., Aramini, J., Sockett, P., Martin, S. W., Wilson, J. & Ellis, A. 2004 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.* **95**, 446-450.

Edge, V. L., Pollari, F., Ng, L. K., Michel, P., McEwen, S. A., Wilson, J. B., Jerrett, M., Sockett, P. N. & Martin, S. W. 2006 Syndromic Surveillance of Norovirus using Over-the-counter Sales of Medications Related to Gastrointestinal Illness. *Can.J.Infect.Dis.Med.Microbiol.* **17**, 235-241.

Eisenberg, J. N., Lei, X., Hubbard, A. H., Brookhart, M. A. & Colford, J. M., Jr. 2005 The role of disease transmission and conferred immunity in outbreaks: analysis of the 1993 Cryptosporidium outbreak in Milwaukee, Wisconsin. *Am J Epidemiol.* **161**, 62-72.

Eisenberg, J. N., Seto, E. Y., Colford, J. M., Jr., Olivieri, A. & Spear, R. C. 1998 An analysis of the Milwaukee cryptosporidiosis outbreak based on a dynamic model of the infection process. *Epidemiology.* **9**, 255-263.

Elderer, F., Myers, E. & Mantel, M. 1964 A statistical problem in space and time: Do leukemia cases come in clusters? *Biometrics.* **20**, 626-638.

Eliassen, R. & Cummings, R. H. 1948 Analysis of waterborne outbreaks, 1938-45. *J. Am. Water Works Assoc.* **40**, 509-5028.

Furtado, C., Adak, G. K., Stuart, J. M., Wall, P. G., Evans, H. S. & Casemore, D. P. 1998 Outbreaks of waterborne infectious intestinal disease in England and Wales, 1992-5. *Epidemiol Infect.* **121**, 109-119.

Gaudart, J., Giorgi, R., Poudiougou, B., Toure, O., Ranque, S., Doumbo, O. & Demongeot, J. 2007 [Spatial cluster detection without point source specification: the use of five methods and comparison of their results]. *Rev Epidemiol Sante Publique.* **55**, 297-306.

German, R. R., Lee, L. M., Horan, J. M., Milstein, R. L., Pertowski, C. A. & Waller, M. N. 2001 Updated guidelines for evaluating public health surveillance systems: recommendations from the Guidelines Working Group. *MMWR Recomm.Rep.* **50**, 1-35.

Glouberman, S. 2001 Walkerton water and complex adaptive systems. *Hosp Q.* **4**, 28-31.

Gorman, A. E. & Wolman, A. 1939 Waterborne outbreaks in the United States and Canada and their significance. *J. Am. Water Works Assoc.* **31**, 225-275.

Gossner, C. M., de Jong, B., Hoebe, C. J., Coulombier, D., European, F. & Waterborne Diseases Study, G. 2015 Event-based surveillance of food- and waterborne diseases in Europe: urgent inquiries (outbreak alerts) during 2008 to 2013. *Euro Surveill.* **20**, 19-28.

Gupta, M. & Haas, C. N. 2004 The Milwaukee Cryptosporidium outbreak: assessment of incubation time and daily attack rate. *J Water Health.* **2**, 59-69.

Guzman-Herrador, B., Carlander, A., Ethelberg, S., Freiesleben de Blasio, B., Kuusi, M., Lund, V., Lofdahl, M., MacDonald, E., Nichols, G., Schonning, C., Sudre, B., Tronnberg, L., Vold, L., Semenza, J. C. & Nygard, K. 2015 Waterborne outbreaks in the Nordic countries, 1998 to 2012. *Euro Surveill.* **20**,

Halonen, J. I., Kivimaki, M., Oksanen, T., Virtanen, P., Virtanen, M. J., Pentti, J. & Vahtera, J. 2012 Waterborne outbreak of gastroenteritis: effects on sick leaves and cost of lost workdays. *PLoS One.* **7**, e33307.

Heffernan, R., Mostashari, F., Das, D., Besculides, M., Rodriguez, C., Greenko, J., Steiner-Sichel, L., Balter, S., Karpati, A., Thomas, P., Phillips, M., Ackelsberg, J., Lee, E., Leng, J., Hartman, J., Metzger, K., Rosselli, R. & Weiss, D. 2004 New York City syndromic surveillance systems. *MMWR Morb Mortal Wkly Rep.* **53 Suppl**, 23-27.

Henning, K. J. 2004 What is syndromic surveillance? *MMWR Suppl.* **53**, 5-11.

Holme, R. 2003 Drinking water contamination in Walkerton, Ontario: positive resolutions from a tragic event. *Water Sci Technol.* **47**, 1-6.

Hoxie, N. J., Davis, J. P., Vergeront, J. M., Nashold, R. D. & Blair, K. A. 1997 Cryptosporidiosis-associated mortality following a massive waterborne outbreak in Milwaukee, Wisconsin. *Am J Public Health.* **87**, 2032-2035.

Hrudey, S. E. & Hrudey, E. J. 2002 Walkerton and North Battleford--key lessons for public health professionals. *Can J Public Health.* **93**, 332-333.

Hrudey, S. E., Payment, P., Huck, P. M., Gillham, R. W. & Hrudey, E. J. 2003 A fatal waterborne disease epidemic in Walkerton, Ontario: comparison with other waterborne outbreaks in the developed world. *Water Sci Technol.* **47**, 7-14.

Journal, C. M. A. 2000 Leadership and fecal coliforms: Walkerton 2000. *CMAJ.* **163**, 1417, 1419.

Kleinman, K. P., Abrams, A. M., Kulldorff, M. & Platt, R. 2005 A model-adjusted space-time scan statistic with an application to syndromic surveillance. *Epidemiol Infect.* **133**, 409-419.

Kleinman KP, Lazarus R, & Platt R (2004). A generalized linear mixed models approach for detecting incident clusters of disease in small areas, with an application to biological terrorism. *Am J Epidemiol* **159**, 217-224.

Koutsotoli, A. D., Papassava, M. E., Maipa, V. E. & Alamanos, Y. P. 2006 Comparing Shigella waterborne outbreaks in four different areas in Greece: common features and differences. *Epidemiol Infect.* **134**, 157-162.

Krewski, D., Balbus, J., Butler-Jones, D., Haas, C., Isaac-Renton, J., Roberts, K. J. & Sinclair, M. 2002 Managing health risks from drinking water--a report to the Walkerton inquiry. *J Toxicol Environ Health A.* **65**, 1635-1823.

Kulldorff, M. 2001 Prospective time periodic geographical disease surveillance using a scan statistic. *Journal of the Royal Statistics Society.* **A 164**, 61-72.

Kulldorff, M. 2010 SaTScan User Guide for version 9.0. 1-109.

Kulldorff, M., Heffernan, R., Hartman, J., Assuncao, R. & Mostashari, F. 2005 A space-time permutation scan statistic for disease outbreak detection. *PLoS.Med.* **2**, e59.

Kulldorff, M., Mostashari, F., Duczmal, L., Katherine Yih, W., Kleinman, K. & Platt, R. 2007 Multivariate scan statistics for disease surveillance. *Stat Med.* **26**, 1824-1833.

Kulldorff, M. & Nagarwalla, N. 1995 Spatial disease clusters: detection and inference. *Stat Med.* **14**, 799-810.

Laine, J., Huovinen, E., Virtanen, M. J., Snellman, M., Lumio, J., Ruutu, P., Kujansuu, E., Vuento, R., Pitkanen, T., Miettinen, I., Herrala, J., Lepisto, O., Antonen, J., Helenius, J., Hanninen, M. L., Maunula, L., Mustonen, J., Kuusi, M. & Pirkanmaa Waterborne Outbreak Study, G. 2011 An extensive

gastroenteritis outbreak after drinking-water contamination by sewage effluent, Finland. *Epidemiol Infect.* **139**, 1105-1113.

Laursen, E., Mygind, O., Rasmussen, B. & Ronne, T. 1994 Gastroenteritis: a waterborne outbreak affecting 1600 people in a small Danish town. *J Epidemiol Community Health.* **48**, 453-458.

Lawson AB, Clark A, & Rodeiro CLV (2003). Developments in general and syndormic surveillance for small area health data. *J of App Stat* **31**, 951-966.

Le Strat, Y. 2015 Détection statistique d'événements inhabituels à partir d'un système de surveillance. *Bulletin épidémiologique - Santé animale et alimentation.* **68**, 50-51.

Mac Kenzie, W. R., Hoxie, N. J., Proctor, M. E., Gradus, M. S., Blair, K. A., Peterson, D. E., Kazmierczak, J. J., Addiss, D. G., Fox, K. R., Rose, J. B. & et al. 1994 A massive outbreak in Milwaukee of cryptosporidium infection transmitted through the public water supply. *N Engl J Med.* **331**, 161-167.

Mackay, B. 2002 Walkerton, 2 years later: "Memory fades very quickly". *CMAJ.* **166**, 1326.

MacKenzie, W. R., Schell, W. L., Blair, K. A., Addiss, D. G., Peterson, D. E., Hoxie, N. J., Kazmierczak, J. J. & Davis, J. P. 1995 Massive outbreak of waterborne cryptosporidium infection in Milwaukee, Wisconsin: recurrence of illness and risk of secondary transmission. *Clin Infect Dis.* **21**, 57-62.

Majowicz, S. E., Edge, V. L., Fazil, A., McNab, W. B., Dore, K. A., Sockett, P. N., Flint, J. A., Middleton, D., McEwen, S. A. & Wilson, J. B. 2005 Estimating the under-reporting rate for infectious gastrointestinal illness in Ontario. *Can J Public Health.* **96**, 178-181.

Manthey, M. W., Ross, A. B. & Soergel, K. H. 1997 Cryptosporidiosis and inflammatory bowel disease. Experience from the Milwaukee outbreak. *Dig Dis Sci.* **42**, 1580-1586.

Marshall, J. K., Thabane, M., Garg, A. X., Clark, W., Meddings, J., Collins, S. M. & Investigators, W. E. L. 2004 Intestinal permeability in patients with irritable bowel syndrome after a waterborne outbreak of acute gastroenteritis in Walkerton, Ontario. *Aliment Pharmacol Ther.* **20**, 1317-1322.

Matsell, D. G. & White, C. T. 2009 An outbreak of diarrhea-associated childhood hemolytic uremic syndrome: the Walkerton epidemic. *Kidney Int Suppl.* S35-37.

McDonald, A. C., Mac Kenzie, W. R., Addiss, D. G., Gradus, M. S., Linke, G., Zembrowski, E., Hurd, M. R., Arrowood, M. J., Lammie, P. J. & Priest, J. W. 2001 Cryptosporidium parvum-specific antibody responses among children residing in Milwaukee during the 1993 waterborne outbreak. *J Infect Dis.* **183**, 1373-1379.

McQuigge, M. 2002 The Walkerton disaster and family physicians. *Can Fam Physician.* **48**, 1596-1597, 1605-1597.

Messner, M., Shaw, S., Regli, S., Rotert, K., Blank, V. & Soller, J. 2006 An approach for developing a national estimate of waterborne disease due to drinking water and a national estimate model application. *J Water Health.* **4 Suppl 2**, 201-240.

Morris, R. D., Naumova, E. N. & Griffiths, J. K. 1998 Did Milwaukee experience waterborne cryptosporidiosis before the large documented outbreak in 1993? *Epidemiology.* **9**, 264-270.

Morris, R. D., Naumova, E. N., Levin, R. & Munasinghe, R. L. 1996 Temporal variation in drinking water turbidity and diagnosed gastroenteritis in Milwaukee. *Am J Public Health.* **86**, 237-239.

Morrison, K., Charland, K., Okhmatovskaia, A. & Buckeridge, D. 2013 A Framework for Detecting and Classifying Outbreaks of Gastrointestinal Disease. *Online J Public Health Inform.* **5**,

Mostashari, F., Kulldorff, M., Hartman, J. J., Miller, J. R. & Kulasekera, V. 2003 Dead bird clusters as an early warning system for West Nile virus activity. *Emerg Infect Dis.* **9**, 641-646.

Mouly, D., Van Cauteren, D., Vincent, N., Vaissiere, E., Beaudeau, P., Ducrot, C. & Gallay, A. 2016 Description of two waterborne disease outbreaks in France: a comparative study with data from cohort studies and from health administrative databases. *Epidemiol Infect.* 144, 591-601

Mouly D, Goria S, Mounié M, Rambaud L, Beaudeau P, Gallay A, Ducrot C, Le Strat Y. Detection of waterborne disease outbreak using health administrative databases: a simulation-based study. *Plos-One* (accepted 25th october 2016, under review).

Murphy, H. M., Pintar, K. D., McBean, E. A. & Thomas, M. K. 2014 A systematic review of waterborne disease burden methodologies from developed countries. *J Water Health.* **12**, 634-655.

Muscatello, D. J., Churches, T., Kaldor, J., Zheng, W., Chiu, C., Correll, P. & Jorm, L. 2005 An automated, broad-based, near real-time public health surveillance system using presentations to hospital Emergency Departments in New South Wales, Australia. *BMC Public Health.* **5**, 141.

Naumova, E. N., Egorov, A. I., Morris, R. D. & Griffiths, J. K. 2003 The elderly and waterborne Cryptosporidium infection: gastroenteritis hospitalizations before and during the 1993 Milwaukee outbreak. *Emerg Infect Dis.* **9**, 418-425.

Nazareth, B., Stanwell-Smith, R. E., Rowland, M. G. & O'Mahony, M. C. 1994 Surveillance of waterborne disease in England and Wales. *Commun Dis Rep CDR Rev.* **4**, R93-95.

Nordin JD, Goodman MJ, Kulldorff M, Ritzwoller DP, Abrams AM, Kleinman K, et al. Simulated anthrax attacks and syndromic surveillance. *Emerg Infect Dis*. 2005;11(9):1394-8.

Noufaily, A., Enki, D. G., Farrington, P., Garthwaite, P., Andrews, N. & Charlett, A. 2013 An improved algorithm for outbreak detection in multiple surveillance systems. *Stat.Med.* **32**, 1206-1222.

Okhmatovskaia, A., Verma, A. D., Barbeau, B., Carriere, A., Pasquet, R. & Buckeridge, D. L. 2010 A simulation model of waterborne gastro-intestinal disease outbreaks: description and initial evaluation. *AMIA Annu Symp Proc.* **2010**, 557-561.

Payment, P. 1997 Epidemiology of endemic gastrointestinal and respiratory diseases: Incidence, fraction attribuable to tap water and costs to society. *Water Sci Technol.* **35**, 7-10.

Pons, W., Young, I., Truong, J., Jones-Bitton, A., McEwen, S., Pintar, K. & Papadopoulos, A. 2015 A Systematic Review of Waterborne Disease Outbreaks Associated with Small Non-Community Drinking Water Systems in Canada and the United States. *PLoS One.* **10**, e0141646.

Poullis, D. A., Attwell, R. W. & Powell, S. C. 2002 An evaluation of waterborne disease surveillance in the European Union. *Rev Environ Health.* **17**, 149-161.

Poullis, D. A., Attwell, R. W. & Powell, S. C. 2005 The characterization of waterborne-disease outbreaks. *Rev Environ Health.* **20**, 141-149.

Proctor, M. E., Blair, K. A. & Davis, J. P. 1998 Surveillance data for waterborne illness detection: an assessment following a massive waterborne outbreak of Cryptosporidium infection. *Epidemiol Infect.* **120**, 43-54.

Rambaud, L., Mouly, D., Schmitt, M., Kerrien, F. & Beaudeau, P. 2011 Utilisation des données de l'Assurance maladie pour valuer l'impact sanitaire d'une épidémie de gastro-entérites d'origine hydrique, Bourg-Saint-Maurice (Arc 1800), 2006. *Bull Epidemio Hebdo.* 339-343.

Richards, A. 2005 The Walkerton Health Study. *Can Nurse.* **101**, 16-21.

Risebro, H. L. & Hunter, P. R. 2007 Surveillance of waterborne disease in European member states: a qualitative study. *J Water Health.* **5 Suppl 1**, 19-38.

Ritter, L., Solomon, K., Sibley, P., Hall, K., Keen, P., Mattu, G. & Linton, B. 2002 Sources, pathways, and relative risks of contaminants in surface water and groundwater: a perspective prepared for the Walkerton inquiry. *J Toxicol Environ Health A.* **65**, 1-142.

Royston, P. & Altman, D. 1994 Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. *Journal of the Royal Statistical Society.* **43**, 429-467.

Salvadori, M. I., Sontrop, J. M., Garg, A. X., Moist, L. M., Suri, R. S. & Clark, W. F. 2009 Factors that led to the Walkerton tragedy. *Kidney Int Suppl.* S33-34.

Sauerbrei, W. & Royston, P. 1999 Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. *Journal of the Royal Statistical Society.* **162**, 71-94.

Schuster, C. J., Ellis, A. G., Robertson, W. J., Charron, D. F., Aramini, J. J., Marshall, B. J. & Medeiros, D. T. 2005 Infectious disease outbreaks related to drinking water in Canada, 1974-2001. *Can J Public Health.* **96**, 254-258.

Smith, A., Reacher, M., Smerdon, W., Adak, G. K., Nichols, G. & Chalmers, R. M. 2006 Outbreaks of waterborne infectious intestinal disease in England and Wales, 1992-2003. *Epidemiol Infect.* **134**, 1141-1149.

Smith, G. D. 2002 Commentary: Behind the Broad Street pump: aetiology, epidemiology and prevention of cholera in mid-19th century Britain. *Int J Epidemiol.* **31**, 920-932.

Smith, S., Elliot, A. J., Mallaghan, C., Modha, D., Hippisley-Cox, J., Large, S., Regan, M. & Smith, G. E. 2010 Value of syndromic surveillance in monitoring a focal waterborne outbreak due to an unusual Cryptosporidium genotype in Northamptonshire, United Kingdom, June - July 2008. *Euro Surveill.* **15**, 19643.

Tillett, H. E., De Louvois, J. & Wall, P. G. 1998 Surveillance of outbreaks of waterborne infectious disease: categorizing levels of evidence. *Epidemiol.Infect.* **120**, 37-42.

Tulchinsky, T. H., Burla, E., Clayman, M., Sadik, C., Brown, A. & Goldberger, S. 2000 Safety of community drinking-water and outbreaks of waterborne enteric disease: Israel, 1976-97. *Bull World Health Organ.* **78**, 1466-1473.

Tuppin, P., De Roquefeuil, L., Weill, A., Ricordeau, P. & Merliere, Y. 2010 French national health insurance information system and the permanent beneficiaries sample. *Rev.Epidemiol.Sante Publique.* **58**, 286-290.

Vakil, N. B., Schwartz, S. M., Buggy, B. P., Brummitt, C. F., Kherellah, M., Letzer, D. M., Gilson, I. H. & Jones, P. G. 1996 Biliary cryptosporidiosis in HIV-infected people after the waterborne outbreak of cryptosporidiosis in Milwaukee. *N Engl J Med.* **334**, 19-23.

Van Cauteren, D., De Valk, H., Vaux, S., Le Strat, Y. & Vaillant, V. 2012 Burden of acute gastroenteritis and healthcare-seeking behaviour in France: a population-based study. *Epidemiol.Infect.* **140**, 697-705.

Wang, H. & Chang, L. 2011 The Walkerton outbreak revisited at year 8: predictors, prevalence, and prognosis of postinfectious irritable bowel syndrome. *Gastroenterology.* **140**, 726-728; discussion 728-729.

Weibel, S. R., Dixon, F. R., B., W. R. & McCabe, L. J. 1964 Waterborne disease outbreaks 1946-60. *J. Am. Water Works Assoc.* **56**, 947-958.

Wheeler, J. G., Sethi, D., Cowden, J. M., Wall, P. G., Rodrigues, L. C., Tompkins, D. S., Hudson, M. J. & Roderick, P. J. 1999 Study of infectious intestinal disease in England: rates in the community, presenting to general practice, and reported to national surveillance. The Infectious Intestinal Disease Study Executive. *BMJ.* **318**, 1046-1050.

Wu, T. S., Shih, F. Y., Yen, M. Y., Wu, J. S., Lu, S. W., Chang, K. C., Hsiung, C., Chou, J. H., Chu, Y. T., Chang, H., Chiu, C. H., Tsui, F. C., Wagner, M. M., Su, I. J. & King, C. C. 2008 Establishing a nationwide emergency department-based syndromic surveillance system for better public health responses in Taiwan. *BMC Public Health.* **8**, 18.

Yang, Z., Wu, X., Li, T., Li, M., Zhong, Y., Liu, Y., Deng, Z., Di, B., Huang, C., Liang, H. & Wang, M. 2011 Epidemiological survey and analysis on an outbreak of gastroenteritis due to water contamination. *Biomed Environ Sci.* **24**, 275-283.

Zhou H & Lawson AB (2008). EWMA smoothing and Bayesian spatial modeling for health surveillance. *Stat Med* **27**, 5907-5928

Zhou, L., Singh, A., Jiang, J. & Xiao, L. 2003 Molecular surveillance of Cryptosporidium spp. in raw wastewater in Milwaukee: implications for understanding outbreak occurrence and transmission dynamics. *J Clin Microbiol.* **41**, 5254-5257.