Détection automatisée d'agrégats de cas de gastro-entérite aiguë et prévention du risque infectieux d'origine fécale porté par l'eau du robinet en France

■ P. BEAUDEAU[1], C. GALEY[1], L. RAMBAUD[1]

Mots-clés : base de données médico-tarifaires, détection automatisée d'agrégats, eau distribuée, épidémie, France, gastro-entérite, surveillance épidémiologique

Keywords: medical database, automated cluster detection, drinking water, outbreak, France, gastroenteritis, surveillance

Glossaire

La méthode automatisée de détection des agrégats décrite dans cet article met en évidence des agrégats limités à une commune et une semaine (**agrégat hebdomadaire**, AH). La consolidation consiste ici à agglomérer les agrégats hebdomadaires consécutifs observés sur une même commune en un seul agrégat, dit **agrégat consolidé** (AC). Un AH dure exactement une semaine et un AC une semaine ou plus. Il n'y a pas ici de coalition spatiale entre 2 AH concomitants touchant deux communes jointives.

Le risque (absolu) d'une maladie est la probabilité de contracter cette maladie. Un **risque relatif** (RR) associé à une exposition (ici habiter sur la commune cible) se définit comme le rapport des probabilités p_1/p_0, p_1 étant la probabilité de contracter la maladie quand on est exposé et p_0 la probabilité de contracter la maladie quand on n'est pas exposé (ici habiter ailleurs dans le département). Si l'exposition ne modifie pas le risque alors RR = 1.

Le **taux d'attaque** est la fraction totale de la population qui a été touchée par l'épidémie.

L'indicateur de survenue d'un événement (maladie, épidémie, etc.) est dit :

– **sensible** s'il est apte à repérer les vrais cas ; les vrais cas oubliés (non identifiés comme tels par l'indicateur) sont appelés « **faux négatifs** ». La sensibilité d'un indicateur se mesure et est estimée par la proportion de vrais cas identifiés par l'indicateur parmi les vrais cas (qu'il s'agit donc de dénombrer par ailleurs, avec une méthode de référence, par exemple) ;

– **spécifique** s'il n'étiquette pas de faux cas comme vrais, c'est-à-dire s'il n'engendre pas de fausses alertes ; les faux cas inclus à tort sont appelés « **faux positifs** ». La spécificité se mesure et est estimée par la proportion de non-cas identifiés comme tels parmi les non-cas.

Sensibilité et spécificité sont liées. Par exemple, en élargissant la définition d'un indicateur, on augmente sa sensibilité, mais on diminue sa spécificité. La définition d'un indicateur représente donc un compromis entre sa spécificité et sa sensibilité.

Introduction

La surveillance du risque d'origine fécale porté par l'eau du robinet se focalise traditionnellement sur les épidémies [BEAUDEAU *et al.*, 2008 ; BRUNKARD *et al.*, 2011]. Depuis le début des années 2000, l'Institut de veille sanitaire (InVS) recense ainsi annuellement en France en moyenne une à deux épidémies de gastro-entérites aiguës (GEA) d'origine hydrique, sur la base des investigations menées par ses représentants régionaux et par les agences régionales de santé [BEAUDEAU *et al.*, 2008]. La détection repose sur le signalement par les acteurs de terrain d'agrégats de cas de GEA, ou d'analyses microbiologiques non conformes de l'eau distribuée. Elle n'atteint pas l'exhaustivité et de nombreux épisodes épidémiques peuvent encore passer inaperçus. C'est notamment le cas des événements de faible ampleur qui sont difficilement identifiables par les acteurs de terrain.

Le développement de bases de données médico-tarifaires offre des opportunités pour la surveillance syndromique. L'exploitation de la base de données du Système national interrégimes de l'Assurance maladie (Sniir-AM) [TUPPIN *et al.*, 2010] permet ainsi de disposer d'une estimation du nombre de cas quotidiens de GEA médicalisés (GEAm) à l'échelle de la commune partout en France [BOUNOURE *et*

[1] Institut de veille sanitaire – 14, rue du Val-d'Osne – 94415 Saint-Maurice cedex. Courriel : p.beaudeau@invs.sante.fr

al., 2010]. Depuis 2009, cet indicateur est utilisé pour confirmer les épidémies suspectées par les opérateurs de terrain et, le cas échéant, pour les caractériser. En 2013, une méthode de détection automatisée d'agrégats (DAA) de cas de GEAm à l'échelle de la commune et de la semaine a été élaborée et testée sur trois départements pilotes.

Les objectifs de cet article sont de :

– présenter les données, la méthode et les résultats de l'étude pilote ;

– discuter la portée en santé publique de l'implémentation d'un tel système de surveillance épidémiologique au niveau national et ses implications possibles pour les exploitants et les autorités sanitaires locales.

1. Matériel et méthodes

1.1. Données

Le nombre de cas de GEAm a été estimé à partir des données de remboursement de médicaments tirées du Sniir-AM. Un algorithme [BOUNOURE *et al.*, 2010] permet de discriminer les ordonnances prescrites pour des cas de GEAm des autres ordonnances par le test de différents critères (combinaison de médicaments « incluants », absence de médicaments « excluants », durée de traitement estimée < 8 jours, délai de délivrance des médicaments < 2 jours). La méthode de discrimination est évaluée tous les ans et amendée en cas d'apparition de nouvelles spécialités ou de déremboursement d'un médicament clé.

Les cas de GEAm qui résultent du processus de discrimination sont étiquetés au jour de consultation du médecin et à la commune de résidence du cas. Les cas dits « touristes », définis par une distance entre la commune de résidence du bénéficiaire des soins et celle d'exercice du médecin prescripteur supérieure à 50 km, sont exclus de l'analyse.

La période d'étude s'étend du 1er janvier 2009 au 31 décembre 2012, soit 208 semaines. Les données comprises entre le 21 mars et le 1er mai 2011 sont manquantes. Quatre épidémies d'origine hydrique connues sont couvertes par les données.

Le secteur d'étude inclut trois départements, le Puy-de-Dôme (63), l'Isère (38) et la Gironde (33), soit 1 543 communes à dominante rurale (42 % de moins de 500 habitants et 81 % de moins de 2 000 habitants). La population totale s'élève à 3,215 millions d'habitants. Le nombre de pollutions accidentelles d'origine fécale de l'eau distribuée observées sur le secteur d'étude ne diffère pas sensiblement de la moyenne nationale, le Puy-de-Dôme se situant dans la moyenne, la Gironde étant moins touchée et l'Isère nettement plus [BEAUDEAU, 2008].

Nous avons utilisé pour la DAA des comptes communaux de cas de GEAm cumulés par semaine. Ce pas de temps s'accorde avec la durée des épidémies d'origine hydrique connues (1 à 3 semaines). Ce choix pallie aussi les variations quotidiennes d'incidence dues aux fluctuations d'activité professionnelle des médecins et des pharmaciens.

Les données de population (recensement 2009) ont été fournies par l'Institut national de la statistique et des études économiques.

1.2. Méthode

La détection d'agrégats de cas de GEAm est fondée sur la mise en évidence d'un nombre hebdomadaire de cas de GEAm observé (NCO) anormalement élevé par rapport au nombre de cas attendu (NCA), c'est-à-dire prévu sous l'hypothèse nulle d'absence d'agrégat.

Plusieurs méthodes de DAA sont disponibles, mais les plus courantes (carte de contrôle, moyennes historiques) ne sont pas adaptées à la nature spatialisée des données. Le scan spatio-temporel de Kulldorff [KULLDORFF *et al.*, 2005] permet d'identifier automatiquement l'étendue spatio-temporelle des agrégats. Il peut prendre en compte la structuration spatiale de réseaux de distribution d'eau (20 % des réseaux desservent plusieurs communes), mais la mise en œuvre de cette option nécessite des données fiables sur le contour des unités de distribution. Dans ce contexte, la méthode retenue s'est inspirée de la pratique épidémiologique de terrain.

Deux méthodes élémentaires ont été utilisées conjointement : un agrégat validé *in fine* doit être repéré par l'une et l'autre méthode. Pour la première (méthode A), le NCA est la médiane des taux d'incidence hebdomadaires des communes de plus de 500 habitants du département de la commune ciblée. L'hypothèse testée est donc que le taux d'incidence sur la commune cible pendant la semaine cible n'est pas significativement supérieur au taux médian observé dans le département. Dans la seconde

méthode (B, adaptée de MANSOTTE et BEAUDEAU [1999]), l'hypothèse testée est que le taux d'incidence sur la commune cible pendant la semaine cible n'a pas augmenté significativement plus vite que le taux médian observé dans l'ensemble des communes du département. L'augmentation se mesure entre une période témoin (semaines –5 à –2) et la semaine ciblée (semaine 0).

Un agrégat communal hebdomadaire de cas de GEAm (AH) est identifié par une méthode élémentaire s'il remplit trois conditions :

– un risque relatif de GEAm, estimé par le rapport entre le nombre de cas observés et le nombre de cas attendus au moins égal à 2 (RR ≥ 2). Des RR de l'ordre de 2 peuvent être en effet observés en dehors de tout facteur de risque d'ordre environnemental entre des populations communales (§ 3.1.2) ;

– un impact sanitaire, estimé par le nombre de cas excédentaires (NCO-NCA) > 5 cas ; ce critère vise à éviter l'inclusion massive de toxi-infections alimentaires collectives (Tiac) qui dominent parmi les agrégats de petite taille (§ 3.2.2) ;

– une probabilité $p < 10^{-5}$. Le test suppose que le nombre hebdomadaire de cas de GEAm sur la commune testée suit une loi de Poisson de moyenne égale NCA.

Chaque département est d'abord exploré avec l'une et l'autre méthode. Les AH détectés à la fois par la méthode A et par la méthode B sont validés (liste A∩B). Les AH consécutifs sur une même commune sont enfin agrégés pour constituer la liste des agrégats consolidés (AC).

Pour évaluer la sensibilité de la méthode, nous avons simulé 46 épidémies impliquant plus de cinq cas de GEAm que nous avons greffées sur des séries temporelles de comptes de cas de GEAm dans des communes n'ayant présenté préalablement aucun agrégat détecté. Deux profils d'épidémies ont été créés, correspondant à des durées de 1 et 3 semaines, et trois niveaux de taux d'attaque représentatifs des épidémies observées en France [BEAUDEAU *et al.*, 2008], ont été distingués (1, 3 et 6 %).

2. Résultats

L'exploration systématique du jeu de données (4 ans × 3 départements) a permis la détection de 826 AH par la méthode A et 543 par la méthode B (*tableau I*). Au total, 210 AH sont communs aux deux méthodes, correspondant à 193 AC. Les agrégats de durée de 2 semaines ou plus représentent 9 % (17/193) des AC totaux et la durée maximale observée est de 3 semaines.

Sur les 4 années de la période d'étude, 162 communes ont éprouvé un AC, soit 10,5 % des communes du secteur d'étude et 16 au moins deux AC, soit 1,5 % des communes. Le maximum observé est de quatre AC par commune.

Aucun agrégat ne survient dans des communes de moins de 100 habitants et 71 % des agrégats éclosent dans les communes entre 500 et 10 000 habitants (*tableau I*). La fréquence annuelle de détection d'un agrégat augmente avec la taille de la commune : 0,02 pour les 100-499, 0,03 pour les 500-1 999, 0,08 pour les 2 000-9 999. La méthode A tend à détecter

Taille (habitants)	Nombre total de communes	Méthode A		Méthode B		A∩B		Communes positives pour au moins un agrégat (A∩B)	
		Nombre agrégats	%	Nombre agrégats	%	Nombre agrégats	%	Nombre	%
[0, 100)	66	1	0	0	0	0	0	0	0,0
[100, 500)	590	81	13	47	9	30	16	27	4,6
[500, 2 000)	599	192	30	238	46	70	36	58	9,7
[2 000, 10 000)	242	171	27	182	35	67	35	60	24,8
[10 000, 50 000)	41	192	30	41	8	23	12	16	39,0
[50 000 et +	5	4	1	5	1	3	2	1	20,0
Total	1 543	641	100 %	513	100 %	193	100 %	162	10,5 %

Tableau I. Distribution des agrégats consolidés détectés par taille de commune (Puy-de-Dôme, Isère et Gironde, 2009-2012)

Nombre de cas de GEAm	Nombre d'agrégats consolidés	%	Total cas (A∩B)	%	Total cas (A)	%
[6, 10)	79	41	533	16	545	13
[10, 20)	68	35	932	28	872	20
[20, 50)	35	18	927	28	1 165	27
[50, 100)	9	5	584	17	495	12
[100 ou +	2	1	356	11	1 194	28
Total	193	100 %	3 332	100 %	4 271	100 %

Tableau II. Distribution des agrégats consolidés détectés par les méthodes A et B selon leur taille (Puy-de-Dôme, Isère et Gironde, 2009-2012)

plus de grands agrégats, notamment dans la catégorie 10 000-49 999 habitants. Sachant que ces agrégats excédentaires trouvés par la méthode A sont des faux positifs (voir *infra*), les deux méthodes ont une sensibilité équivalente.

Les 193 AC détectés totalisent 3 332 cas de GEAm excédentaires. 41 % impliquent moins de dix cas de GEAm *(tableau II)*. Ces petits agrégats représentent 16 % de l'ensemble des cas détectés. À l'opposé, 9 AC (6 % des AC) impliquent 50 cas et plus et représentent 28 % des cas totaux. Le plus grand AC détecté comprend 199 cas sur une durée de 2 semaines. Le nombre mensuel d'AC est de 4 en moyenne sur la zone d'étude, soit 1,3 AC par département et par an ; il culmine en décembre (9), atteint son minimum en mai (2) et fluctue entre 3 et 5 le reste de l'année. On n'observe pas de structure particulière dans la répartition spatiale des communes touchées par la survenue d'au moins un AC au sein des départements.

Quatre épidémies déjà connues (chacune > 30 cas GEAm, exemple en *figure 1*) sont identifiées par la DAA. Les simulations d'épidémies donnent une indication supplémentaire sur la sensibilité de la méthode A∩B. Le protocole utilisé ne permet pas d'établir la spécificité (programmé en 2015). Sur 46 épidémies simulées, 9 sont imputées par tirage au sort à des communes entre 100 et 500 habitants, 19 à des communes entre 500 et 2 000 habitants et 18 à des communes de 2 000 habitants et plus. La DAA en identifie 37 (se = 0,8). Les faux négatifs surviennent dans des communes de petite taille (8 sur 9 dans des communes < 2 000 habitants) ; ce sont des agrégats de petite taille (6 sur 9 sont de taille inférieure à 10 cas, contre 29 % des AC totaux) et majoritairement hivernaux (6 sur 9).

Nous avons inféré la fréquence des agrégats de GEAm observée dans l'étude au niveau national. Plus de 1 000 AC impliquant plus de 5 cas de GEAm (soit 30 000 cas de GEAm) pourraient être détectés annuellement en France, dont près de 100 AC impliquant 50 cas ou plus. Ces estimations sont des ordres de grandeur à valider dans le futur.

Si l'on se réfère à la définition symptomatique de la GEA [MAJOWICZ *et al.*, 2008] (en simplifiant : au moins trois selles liquides en 24 heures ou vomissements) et à une fréquence moyenne de consultation d'un médecin en cas de GEA en France de 33 % [VAN CAUTEREN *et al.*, 2012], on pourrait attendre un impact annuel des agrégats mis en évidence par la DAA de l'ordre de 100 000 cas de GEA (définition symptomatique).

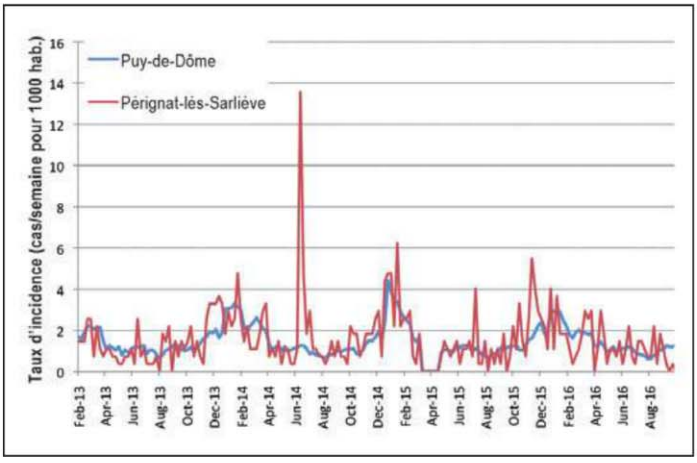

Figure 1. Taux d'incidence des GEAm dans le Puy-de-Dôme et sur la commune de Pérignat-lès-Sarliève (2009-2012). L'épidémie de Pérignat-lès-Sarliève (juin 2010) émerge distinctement du bruit de fond

3. Discussion

3.1. Perspectives méthodologiques

3.1.1. La spécificité prime sur la sensibilité

Les agrégats identifiés par la DAA sont destinés à être examinés sur le terrain (§ 3.2.2). L'enquête de terrain représente toutefois une mobilisation de moyens non négligeable et l'investigation de « faux positifs » sera mise au débit du système par le personnel en charge des investigations. Par ailleurs, le nombre d'agrégats identifiés (16 par département et par an en moyenne) pourrait dépasser les capacités d'investigation des opérateurs locaux, qui donnent la priorité probablement aux enquêtes par importance décroissante de l'impact. L'économie des moyens justifie donc de privilégier la spécificité du processus de recherche des agrégats à étudier plutôt que sa sensibilité.

On distingue cependant deux niveaux d'analyse de la spécificité :

– les agrégats identifiés peuvent-ils être dus au hasard ou résulter de biais méthodologiques, c'est-à-dire l'épidémie est-elle réelle (spécificité statistique) ?

– quelle est la probabilité que l'agrégat identifié soit attribuable à l'eau du robinet plutôt qu'à d'autres voies d'exposition (spécificité hydrique) : circulation orofécale par contact (mode de transmission qui domine durant l'épidémie virale hivernale), alimentation (Tiac), ou encore les baignades (voie d'exposition mineure en matière d'impact global).

On voit d'abord (§ 3.1.2) comment la recherche de la spécificité (notamment statistique) se décline dans le choix des options de la DAA. Cette orientation se prolonge dans l'investigation locale des agrégats découverts par la DAA (§ 3.2.2), avec le souci supplémentaire de reconnaître précocement les Tiac afin de diriger les moyens d'investigation sur le risque hydrique (recherche de la spécificité hydrique).

3.1.2. Les méthodes A et B sont complémentaires

Les critères de sélection des agrégats constituent une première barrière contre les faux positifs (spécificité statistique). Si l'on prend en compte la distribution de Poisson des comptes de cas hebdomadaires communaux et l'indépendance des tests, la multiplication des tests statistiques opérés engendre sur chacun des départements du secteur d'étude un faux positif tous les 4 ans ($p < 10^{-5}$ pour 100 000 essais environ). La sélection drastique opérée en ne retenant que l'intersection des agrégats positifs pour A et B (soit 18 % des agrégats hebdomadaires détectés par l'une ou l'autre méthode) contribue en premier à la spécificité de la méthode. Une analyse comparative des deux méthodes montre en effet que certaines situations à la source de faux positifs ne concernent qu'une seule des deux méthodes. Nous en donnons un exemple.

Le risque de faux positif engendré par les variations du taux d'incidence en fonction de l'âge mérite une attention particulière. D'une part, le risque de GEAm est en moyenne quatre fois plus élevé chez les enfants que chez les adultes, avec des variations saisonnières importantes de ce rapport ; d'autre part, le ratio du nombre d'adultes sur le nombre d'enfants varie de 2 à 10 dans les communes de 500 habitants ou plus incluses dans l'étude. Ces deux éléments se combinent pour engendrer une variabilité importante des taux d'incidence bruts (c'est-à-dire sans correction de l'effet de l'âge) entre communes. Ainsi, comparer directement l'incidence brute des GEAm dans une commune urbaine ou périurbaine jeune avec la médiane du département à dominante rurale et âgée (méthode A) engendre un RR qui peut avoisiner 2 en l'absence de tout facteur de risque environnemental. Ce biais se réalise effectivement avec la méthode A qui identifie quatre fois plus d'agrégats dans les communes de plus de 10 000 habitants que la méthode B (*tableau I*). La méthode B est par construction insensible à ce biais, et permet donc de corriger A au prix de l'émergence possible de faux négatifs. Une amélioration à prévoir pour mieux prévenir l'occurrence résiduelle de ce type de faux positifs est la standardisation sur l'âge des comptes communaux des cas de GEAm. Celle-ci permettrait aussi d'améliorer la sensibilité en relâchant la condition RR > 2 visant justement à prévenir ce type de faux positif.

Inversement, on peut pallier des défauts propres à la méthode B en jouant sur la complémentarité des deux méthodes. En s'appuyant sur une comparaison temporelle (période cible *vs* période témoin), la méthode B peut manquer la fin d'une épidémie qui se prolonge au-delà de 2 semaines, car pour le test de la semaine 3 de l'épidémie, la période témoin inclut par construction la semaine 1 de l'épidémie. Cette

inclusion surestime NCA et favorise l'émergence de faux négatifs en fin d'épidémie. Ainsi, 15 des AC détectés par la méthode A durent 3 semaines ou plus (maximum observé : 11 semaines), contre un seul pour la méthode B. Pour pallier la perte de sensibilité due à ce défaut, il est envisagé de caractériser l'impact et la durée des agrégats validés (communs à A et B) par l'estimation plus extensive produite par la méthode A *(tableau II)* plutôt que celle produite par A∩B. Sur le jeu de données de l'étude, le choix de cette option augmente de 28 % l'impact total en nombre de cas.

Il reste à examiner la spécificité des agrégats vis-à-vis de l'origine hydrique. Les causes les plus fréquentes d'agrégats non véhiculés par l'eau de distribution sont, d'une part, les Tiac et, d'autre part, les agrégats issus d'une transmission par contact et s'inscrivant principalement dans la montée en charge de l'épidémie de GEA hivernale.

Un peu plus de 1 200 Tiac sont rapportées chaque année en France au titre de la déclaration obligatoire [DELMAS *et al.*, 2006], mais ce nombre ne représente qu'une minorité des Tiac [HAEGHEBAERT *et al.*, 2001]. Les Tiac impliquent un nombre réduit de cas de GEA (quatre à cinq en moyenne, soit moins de deux cas de GEAm). Ce sont principalement les Tiac en restauration collective qui apparaissent parmi les agrégats repérés par la DAA, car la quasi-totalité des Tiac familiales est évincée avec l'exclusion des agrégats impliquant moins de cinq cas de GEAm (soit en moyenne 15 GEA).

Parmi les Tiac en restauration collective repérées par la DAA, une partie a déjà été rapportée au titre de la déclaration obligatoire et ces infections ne sont donc pas candidates à une nouvelle investigation. D'autres sont identifiables par la structure en âge de la population des cas au sein de l'agrégat. En effet, la majorité des établissements de restauration collective s'adressent à des classes d'âge particulières : enfants pour les cantines scolaires, adultes pour la restauration d'entreprise ou de maison de retraite. Le rapport des effectifs entre les cas enfants et adultes (1-15 ans *vs* ≥ 16 ans) au sein de l'agrégat pourrait ainsi aider à débusquer certaines Tiac en restauration collective et à mieux cibler les épidémies d'origine hydrique lors des investigations de terrain. Il restera cependant parmi les agrégats mis en évidence par la DAA des Tiac non identifiées préalablement comme telles, avec une fréquence qui devrait sensiblement diminuer quand la taille de l'agrégat augmente.

L'épidémie hivernale de GEA d'étiologie virale qui touche l'hémisphère Nord entre décembre et février se propage principalement par contact interpersonnel ou *via* des objets contaminés (mais aussi possiblement par l'eau du robinet ou les aliments). La dynamique de l'incidence est moins lissée au niveau local *(figure 1)* qu'au niveau départemental, avec l'éclosion d'agrégats communaux au niveau du front épidémique. Ces agrégats, qui se concentrent pendant le mois de décembre, ont été identifiés comme une source possiblement importante d'agrégats authentiques mais non liés à l'eau. Cette période pourrait concentrer les agrégats résiduels authentiques mais non liés à l'eau.

En résumé, l'étude pilote a démontré l'intérêt de la DAA, notamment le gain substantiel de sensibilité qu'il procure par rapport à l'approche conventionnelle. La spécificité est assurée par la complémentarité des deux méthodes. La spécificité « statistique » est perfectible par standardisation des données sur l'âge. La structure en âge des agrégats peut en outre aider au repérage des Tiac.

3.2. Portée de la surveillance des agrégats en santé publique

3.2.1. La question de l'impact

Depuis le XIX^{e} siècle, nul ne conteste le rôle direct que peut avoir l'eau dans les épidémies de typhoïde, de choléra et de nombreuses infections se traduisant principalement par les symptômes de la GEA. La transition épidémiologique observée au début du XX^{e} siècle a déplacé l'expression du risque. Les épidémies de typhoïde et de choléra ont disparu dans les pays développés. La question s'est alors posée du rôle de l'eau dans l'occurrence des cas sporadiques (dits encore endémiques) des infections bactériennes, virales ou protozoaires d'origine fécale.

Le risque endémique lié aux installations non conformes a été identifié à travers de rares études épidémiologiques [ZMIROU *et al.*, 1987] et le risque endémique attribuable aux installations conformes à la réglementation sanitaire est démontré pour