# Réutilisation des données de vie réelle dans la gestion des épidémies d'arboviroses en territoire Caribéen. Application à la surveillance de la dengue en Martinique

Emmanuelle Sylvestre

UNIVERSITÉ DE RENNES 1

# THESE DE DOCTORAT DE

L'UNIVERSITE DE RENNES 1

ECOLE DOCTORALE N° 605
Biologie Santé
Spécialité : Analyse et traitement de l'information et des images médicales

Par
**Emmanuelle SYLVESTRE**

## Réutilisation des données de vie réelle dans la gestion des épidémies d'arboviroses en territoire Caribéen

Application à la surveillance de la dengue en Martinique

**Thèse présentée et soutenue à Rennes, le 5 avril 2022**
**Unité de recherche : Laboratoire Traitement du Signal et de l'Image, Equipe Données Massives en Santé**

**Rapporteurs avant soutenance** :

| | |
|---|---|
| Sandra BRINGAY | Professeur des Universités, Université de Montpellier |
| Anne-Sophie JANNOT | Maître de conférences des Universités – Praticien Hospitalier, Université Paris Cité |

**Composition du Jury :**

| | | |
|---|---|---|
| Président : | Grégoire FICHEUR | Professeur des Universités – Praticien Hospitalier, Université de Lille |
| Examinateurs : | Sandra BRINGAY | Professeur des Universités, Université de Montpellier |
| | Jacqueline DELOUMEAUX | Professeur associé – Praticien Hospitalier, Université des Antilles |
| | Anne-Sophie JANNOT | Maître de conférences des Universités – Praticien Hospitalier, Université Paris Cité |
| Dir. de thèse : | Marc CUGGIA | Professeur des Universités – Praticien Hospitalier, Université de Rennes 1 |
| Co-Dir. de thèse : | André CABIE | Professeur des Universités – Praticien Hospitalier, Université des Antilles |

# Remerciements

Je tiens à remercier particulièrement le Professeur Marc Cuggia de m'avoir encadrée durant cette thèse, mais surtout pour sa présence constante depuis de nombreuses années. Tes conseils et ton soutien même dans les moments les plus difficiles me sont précieux.

Je remercie également le Professeur André Cabié pour son encadrement malgré des circonstances extrêmement compliquées. Merci d'avoir su te rendre disponible quand j'en avais besoin.

Mes remerciements vont aussi aux membres du jury. Merci à mes rapporteurs, le Professeur Sandra Bringay et le Docteur Anne-Sophie Jannot de me faire l'honneur d'évaluer ce travail. Merci aux Professeurs Jacqueline Deloumeaux, et Grégoire Ficheur d'avoir accepté de participer à ce jury de thèse.

Je remercie également l'ensemble des personnes ayant participé de près ou de loin à ce travail : Guillaume, Elsa, Sandrine, Clarisse, Fatiha, René-Michel, Cédric, Boris et le Professeur Moustapha Dramé. Merci également à Manuel Etienne et Fabrice Malouines pour leur expertise et leur disponibilité.

Merci aussi aux membres de mon Comité de Suivi Individuel de thèse pour leur aide et leurs conseils : le Professeur Raymond Césaire, le Docteur Sahar Bayat-Makoei et le Docteur Cédric Arvieux.

Enfin, je remercie chaleureusement ma famille et mes amis pour leur soutien sans faille et leurs encouragements depuis toutes ces années.

# Table des matières

# Résumé

Les maladies infectieuses émergentes sont devenues un enjeu majeur de santé publique ces dernières années et sont à l'origine de nombreuses crises sanitaires touchant en majorité les régions tropicales et subtropicales telles que la Caraïbe. À l'heure actuelle, les systèmes de surveillance syndromique basés sur la collecte et l'analyse de données souffrent de plusieurs limites, notamment les délais entre un événement de santé et sa notification et le coût élevé de ces systèmes, alors que l'accélération de la propagation de ces pathologies demande des systèmes de plus en plus réactifs. Des alternatives basées sur les données de vie réelle, c'est-à-dire l'ensemble des données recueillies en dehors de la recherche clinique, sont étudiées depuis plusieurs années pour tenter de répondre aux enjeux de réactivité et de disponibilité des données. L'objectif de cette thèse est d'évaluer la place des données de vie réelle dans la surveillance syndromique et plus spécifiquement de la surveillance de la dengue dans la Caraïbe.

Cette thèse sur articles explore trois axes importants : une première partie fait un état des lieux et étudie les enjeux des systèmes de surveillance syndromique dans la Caraïbe. La seconde partie expose les verrous liés aux données de vie réelle. La troisième partie explore les verrous liés à l'exploitation des données à travers le cas d'usage de la surveillance de la dengue en Martinique. Enfin, la discussion s'intéresse aux atouts et contraintes des données de vie réelle dans la surveillance syndromique et leur place dans la préparation (*preparedness)* et la capacité de réponse des systèmes à répondre aux épidémies.

**Mots-clés :** Données de vie réelle ; Données massives en santé ; Surveillance syndromique ; Caraïbe ; Dengue

# Abstract

The incidence of emerging infectious diseases has increased considerably over the last few years, causing several health crises, especially in tropical and subtropical areas such as the Caribbean. Currently, traditional surveillance systems based on data collection are hampered by several limitations, especially delays between a case and its notification, and high costs despite their need to be responsive. Alternatives approaches based on real-world data, ie. data not collected in experimental conditions, have been studied for several years to help improve responsiveness and data availability of these systems. The aim of this thesis is to assess the role of real-world data in syndromic surveillance and more specifically dengue monitoring in the Caribbean.

This thesis explored three major axes: first, we conducted a state of the art on syndromic surveillance systems in the Caribbean and their limitations. Secondly, we identified the challenges related to real-world data. The third section explored the challenges related to data exploitation through the use case of dengue surveillance in Martinique. Finally, the discussion focused on the challenges and opportunities of real-world data in syndromic surveillance and their place in epidemic preparedness and response.

**Keywords :** Real-world data; Health Big Data; Syndromic surveillance; Caribbean; Dengue

# Productions scientifiques liées à la thèse

**Sylvestre E**, Thuny RM, Cécilia-Joseph E, Gueye P, Chabartier C, Brouste Y, Mehdaoui H, Najioullah F, Pierre-François S, Abel S, Cabié A, Dramé M. "Health informatics support for outbreak management: How to respond without an electronic health record?" *Journal of the American Medical Informatics Association.* 2020 Nov 1;27(11):1828-1829. doi: 10.1093/jamia/ocaa183.

**Sylvestre E**, Bouzille G, McDuffie M, Chazard E, Avillach P, Cuggia M. "A semi-automated approach for multilingual terminology matching: mapping the French version of the ICD-10 to the ICD-10 CM." *Studies In Health Technogy and Informatics*. 2020 June 16;270:18-22. doi: 10.3233/SHTI200114. PMID**:** 32570338

**Sylvestre E**, Joachim C, Cécilia-Joseph E, Bouzillé G, Campillo-Gimenez B, Cuggia M, Cabié A. "Data-driven methods for dengue prediction and surveillance using real-world and Big Data: A systematic review." *PLoS Neglected Tropical Diseases*. 7 janv 2022;16(1):e0010056.

**Sylvestre E**, Cécilia-Joseph E, Bouzillé G, Najioullah F, Etienne M, Malouines F, Rosine J, Julié S, Cabié A, Cuggia M. "Combining heterogenous real-world data sources to monitor dengue fever in Martinique. *Soumis à JMIR Public Health and Surveillance*

# Liste des abréviations

ATIH : Agence Technique de l'Information sur l'Hospitalisation

CDRN : Clinical Data Research Network

COVID-19 : Maladie à coronavirus 2019

CPRD : Clinical Practice Research Datalink

DENV : Virus de la dengue

DENV-1 : Virus de la dengue, sérotype 1

DENV-2 : Virus de la dengue, sérotype 2

DENV-3 : Virus de la dengue, sérotype 3

DENV-4 ; Virus de la dengue, sérotype 4

DIM : Département d'Information Médicale

DMP : Dossier Médical Partagé

DOMASIA : DOnnées MAssives en Santé et système d'Information Apprenant

DPI : Dossier Patient Informatisé

EDS : Entrepôt de Données de Santé

eHOP : Entrepôt HOPital

EHR4CR : Electronic Health Records for Clinical Research

HDH : Health Data Hub

HIMSS : Healthcare Information and Management Systems Society

HIPAA : Health Insurance Portability and Accountability Act

I2b2 : Informatics for Integrating Biology and the Bedside

ICD-10-CM : ICD-10 Clinical Modification

LTSI : Laboratoire Traitement du Signal et de l'Image

MCO : Médecine Chirurgie Obstétrique

OMOP : Observational Medical Outcomes Partnership

OMS : Organisation Mondiale de la Santé

PAHO : Pan American Health Organization

PCORI : Patient-Centered Outcomes Research Institute

PCORNET : PCORI-funded National Patient-Centered Clinical Research Network

PMSI : Programme de Médicalisation du Système d'Information

PSAGE : Programme de Surveillance, d'Alerte et de Gestion des Epidémies

SHRINE : Shared Health Research Information Network

SIG : Système d'Information Géographique

SNIIRAM : Système National d'Information Inter-Régimes de l'Assurance Maladie

SNDS : Système National des Données de Santé

SRAS : Syndrome Respiratoire Aigu Sévère

PPRN : Patient-Powered Research Network

RGPD : Règlement Général de Protection des Données

RiCDC : Réseau Interrégional des Centres de Données Cliniques

UMLS : Unified Medical Language System

# Liste des figures

# Première partie
# État des lieux et enjeux des systèmes de surveillance syndromique en région Caraïbe

# I. Les maladies infectieuses émergentes

Les maladies infectieuses émergentes sont devenues un enjeu majeur de santé publique ces dernières années. Elles sont à l'origine de nombreuses crises sanitaires mondiales, telles que le Syndrome Respiratoire Aigu Sévère (SRAS), en 2003, la grippe A (H1N1) en 2009, ou plus récemment, la maladie à coronavirus 2019 (COVID-19), classée en pandémie par l'Organisation Mondiale de la Santé (OMS) en mars 2020 (1). On définit comme « maladie infectieuse émergente » toutes les pathologies infectieuses d'identification récente ou ayant existé précédemment, mais dont l'incidence ou la zone géographique augmente rapidement (2). Dans le cas des maladies anciennes dont l'incidence est en augmentation après une période de déclin ou de contrôle, on parle alors de maladie ré-émergente (3).

Contrairement à d'autres pathologies, les maladies infectieuses émergentes sont imprévisibles et malgré de nombreuses avancées concernant la surveillance, le diagnostic, la thérapeutique et la vaccination, leur contrôle reste un défi à la fois de santé publique, mais aussi de stabilité économique (4). Par ailleurs, de nombreuses maladies émergentes sont des zoonoses, ce qui signifie qu'elles sont d'abord apparues chez des animaux avant d'être transmises à l'être humain. Ainsi, environ 75 % des maladies infectieuses émergentes des 40 dernières années proviennent de réservoirs animaux (5). Certaines régions, en particulier l'Asie, l'Amérique Centrale et Latine et l'Afrique tropicale sont plus vulnérables face au risque épidémiologique présenté par ces maladies (6).

Plusieurs facteurs peuvent expliquer l'émergence ou la réémergence d'une pathologie infectieuse, comme expliqué sur la figure de Spernovasilis, Tsiodras et & Poulakou (2022) (Figure 1) (7). Ces facteurs résultent des interactions entre les agents infectieux, les hôtes et l'environnement, facilitant l'évolution des agents infectieux dans des niches écologiques et leur permettant ainsi d'atteindre et de s'adapter à de nouveaux hôtes et de se disséminer plus facilement parmi ces derniers (3).

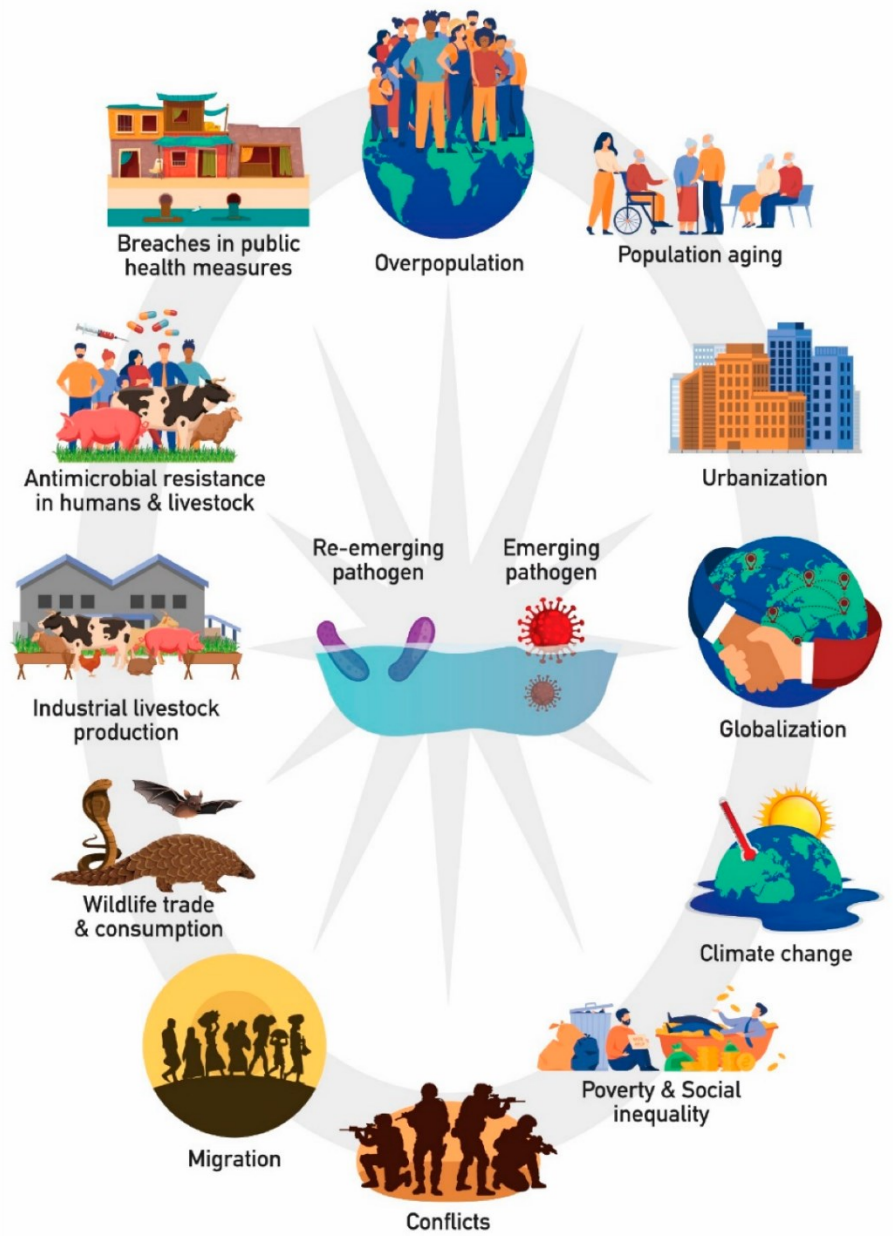



**Figure 1.** Facteurs précipitant l'incidence et la transmission des maladies infectieuses émergentes et ré-émergentes. (d'après Spernovasilis, Tsiodras et & Poulakou) (7)

Parmi ces maladies émergentes, nombre d'entre elles affectent de façon disproportionnée les populations les plus pauvres et certaines maladies tropicales fréquentes au sein de ces populations sont regroupées en « Pathologies tropicales négligées ». Ces pathologies correspondent à 17 maladies infectieuses particulièrement endémiques dans les régions tropicales et subtropicales (8).

Pour faire face aux futures épidémies de maladies infectieuses émergentes, un système de santé solide est indispensable, en particulier concernant la surveillance, la collecte d'informations, l'évaluation du risque, la capacité de communication entre les différents acteurs de Santé Publique, et la capacité du système à s'adapter aux nouvelles menaces.

# II. La dengue en Amérique Latine et dans la Caraïbe

La dengue fait partie des maladies tropicales négligées. Il s'agit de la principale arbovirose, causée par le virus de la dengue (DENV) et transmise aux humains par la piqûre de moustiques femelles infectés du genre Aedes : *Aedes aegypti* et *Aedes albopictus* (9). Ces moustiques sont aussi vecteurs des virus du chikungunya, du Zika et de la fièvre jaune. Il existe 4 sérotypes viraux (DENV-1, DENV-2, DENV-3 et DENV-4) capables de circuler en même temps dans les régions endémiques (10).

## 1. Épidémiologie et répartition géographique de la dengue

Selon l'OMS, seuls 9 pays avaient connu des épidémies graves de dengue avant 1970 (11), pourtant, l'incidence de cette maladie a été multipliée par 30 en 50 ans (12), avec la croissance démographique, l'urbanisation, l'augmentation des mouvements de population et le relâchement des programmes de contrôle vectoriels (13). Actuellement, la dengue est l'une des maladies vectorielles les plus fréquentes au monde et on estime qu'environ 4 milliards de personnes dans plus de 125 pays sont à risque d'infection, avec 390 millions de personnes touchées, 96 millions de cas symptomatiques et 20 000 morts par an (9,14). Les régions principalement touchées sont les régions tropicales et subtropicales (Asie du Sud-Est, Pacifique, Amérique). En Amérique Latine et aux Antilles (Figure 2), la morbidité et la mortalité de la maladie ont augmenté de façon exponentielle, passant d'environ 400 000 cas et un peu moins de 100 morts en 2000 à plus de 3 millions de cas et environ 1500 morts en 2019 (15,16). Entre 2000 et 2008, les Antilles représentaient 3,9% des cas de dengue déclarés à l'OMS en Amérique (17), la majorité étant déclarée par la Martinique, la Guadeloupe et la Guyane française. Dans les Antilles françaises (Martinique, Guadeloupe, Saint-Martin, Saint-Barthélemy), la dengue est endémo-épidémique, avec des recrudescences saisonnières tous les ans, de gravité variable. La Martinique a connu sa première grosse épidémie de dengue hémorragique en 1997-1998 (18), suivie par les épidémies de 2001-2002, 2005, 2007, 2013-2014 et 2019-2021 (19–21),

cette dernière étant considérée comme la plus longue jamais enregistrée depuis la mise en place de la surveillance de la dengue aux Antilles françaises.

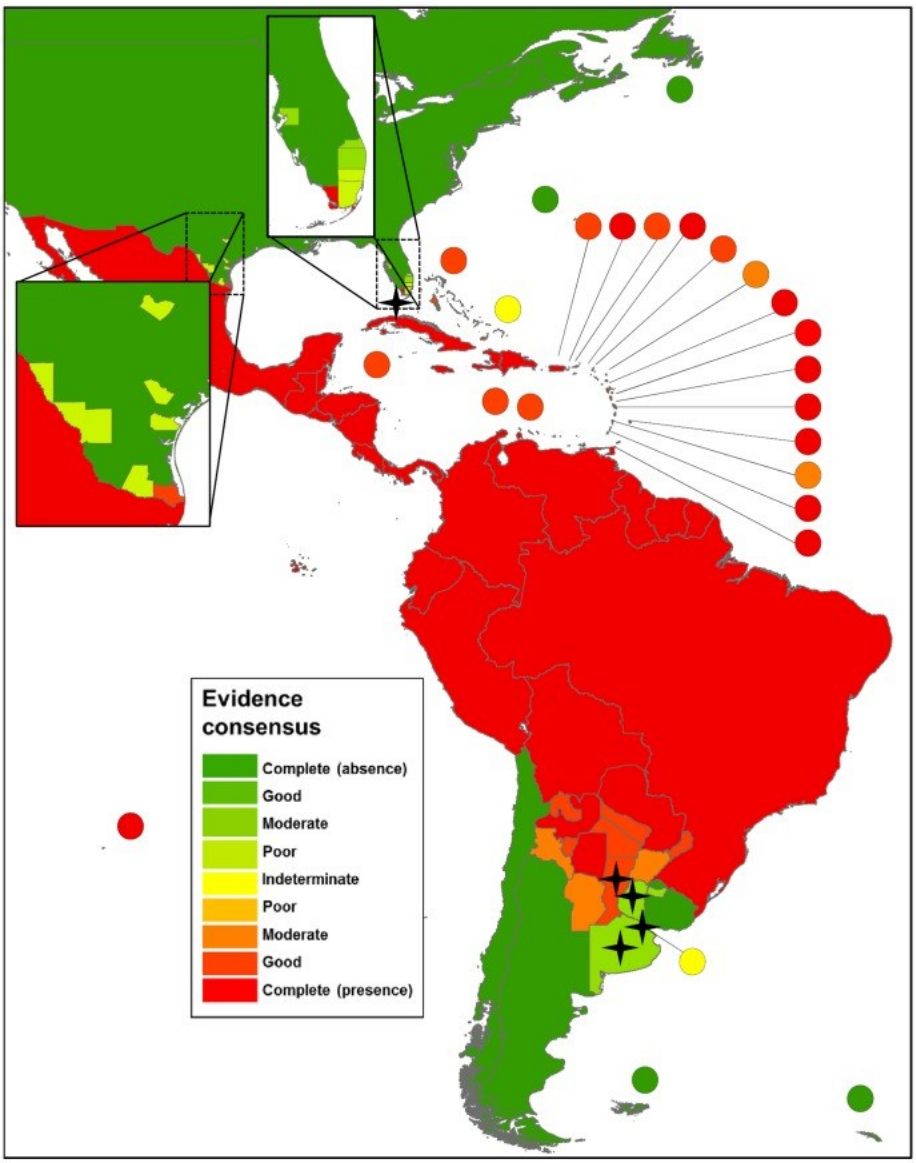



**Figure 2.** Répartition des transmissions du virus de la dengue sur le continent américain en 2012. En vert, les régions pour lesquelles il y a consensus sur l'absence de transmission du virus, en rouge, les régions pour lesquelles il y a consensus sur la présence du virus. (D'après Brady et al.) (22)

## 2. Clinique et classification

Les manifestations cliniques de la maladie sont variables, avec des évolutions imprévisibles. La majorité des patients est asymptomatique ou présente des symptômes d'allure grippale, qui apparaissent à la suite d'une période d'incubation de 4 à 10 jours après la piqûre d'un moustique

infecté et perdurent de 2 à 7 jours (Figure 3) (23). À la suite de cette phase fébrile, des formes sévères, notamment des formes hémorragiques pouvant aller jusqu'au décès, peuvent apparaître (phase critique). Cependant, les prises en charge précoces de la maladie peuvent réduire le risque de décès à moins de 1 % des formes sévères.

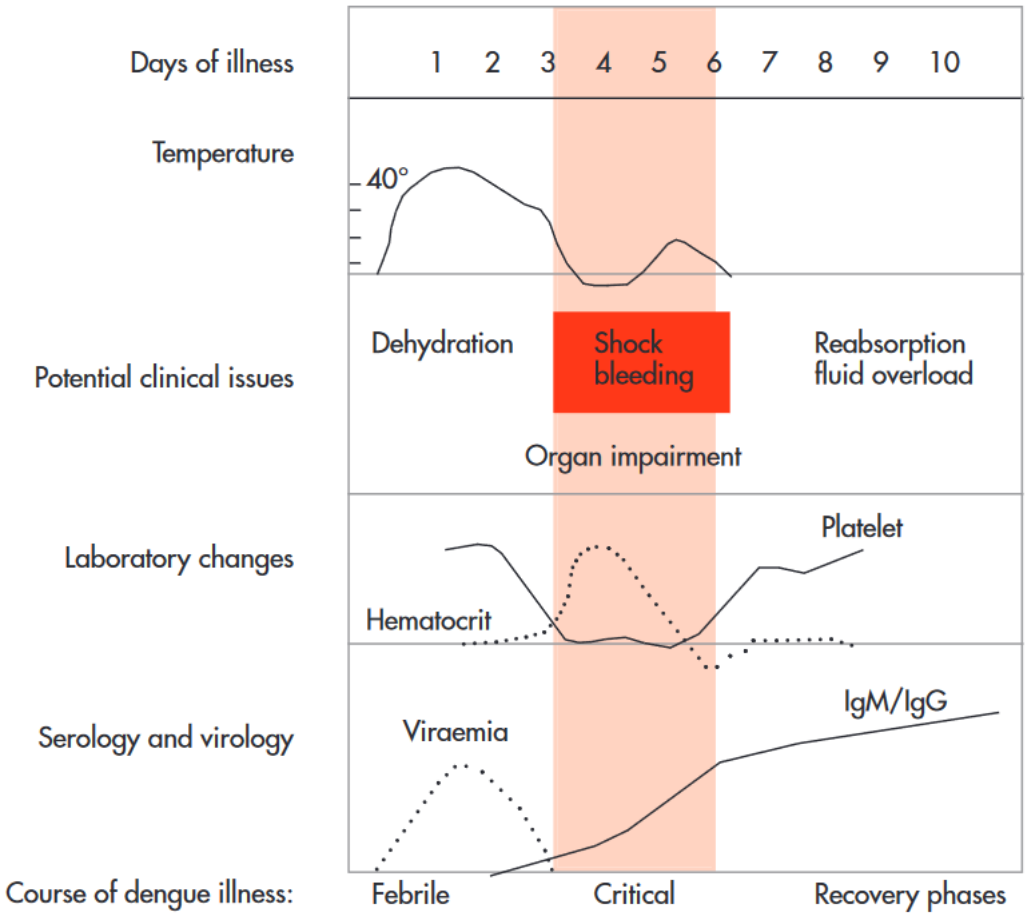



**Figure 3.** Evolution de la dengue (D'après les recommandations de l'OMS) *(23)*

Cette prise en charge repose sur la capacité à distinguer les formes bénignes des formes sévères et est d'autant plus capitale en période d'épidémie, où les services de santé peuvent se retrouver débordés par l'afflux de malades et la désorganisation des systèmes de soins (23).

Une première classification de l'OMS en 1997 groupait les infections symptomatiques en trois catégories : dengue classique (*dengue fever*, DF), dengue hémorragique (*dengue haemorragic fever*, DHF) et la dengue avec syndrome de choc (*dengue shock syndrome*, DSS) (24).

Face à la difficulté croissante rapportée par les cliniciens d'utilisation de cette classification pour trier les formes bénignes des formes sévères (25), une nouvelle classification, parue en

2009 a été proposée afin d'améliorer les prises en charge précoces. Elle distingue les formes de dengue, avec ou sans signe d'alerte, de la dengue sévère (Figure 4).

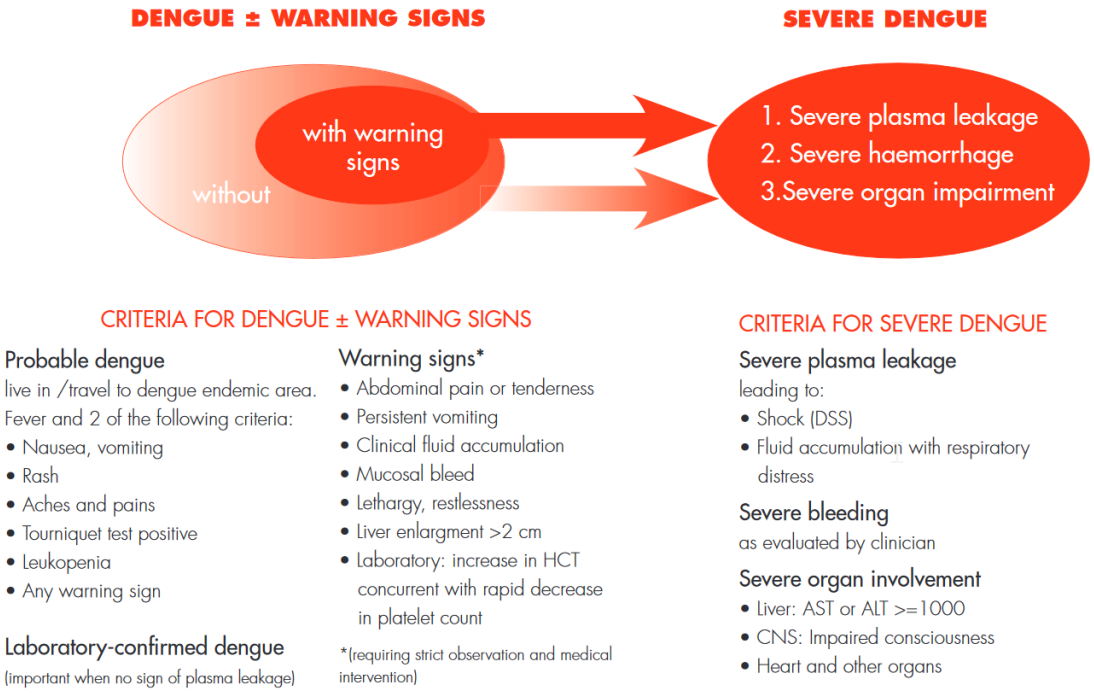



**Figure 4.** Nouvelle classification de la dengue par l'OMS en 2009 *(23)*.

Ces évolutions dans la classification de la maladie se retrouvent également dans la façon dont celle-ci est codée dans la 10ème édition de la Classification Internationale des Maladies (CIM-10), puisqu'avant la classification de 2009, la dengue pouvait être codée en « Dengue classique » (A90) ou « Fièvre hémorragique due au virus de la dengue » (A91) et depuis 2016, elle est désormais codée en « Dengue sans signe d'alerte » (A97.0), « Dengue avec signes d'alerte » (A97.1) ou « Dengue sévère » (A97.2).

La remontée précoce des notifications de cas de dengue est donc un élément crucial dans la gestion efficace des épidémies.

## 3. Stratégie de prévention et de lutte contre la dengue

La prévention et le contrôle des épidémies de dengue reposent sur deux piliers : la lutte antivectorielle et la surveillance syndromique des cas cliniques évocateurs (26).

La surveillance syndromique est basée sur une collecte active ou passive de données et permet de détecter précocement les épidémies, de surveiller la distribution temporelle et géographique

des cas de dengue et d'adapter la réponse d'un territoire à la situation épidémiologique. La surveillance entomologique, elle, permet de surveiller les changements dans la distribution géographique des moustiques et de mettre en place des interventions appropriées, de la simple éducation à l'éradication par insecticide (23).

Dans les Antilles françaises, la gestion des épidémies de dengue est sous la responsabilité de l'agence nationale de santé publique, Santé Publique France (27). Face à la recrudescence des épidémies de dengue dans la Caraïbe, Santé Publique France a mis en place dès 2006, avec l'ensemble des partenaires de chaque territoire, le « Programme de surveillance, d'alerte et de gestion des épidémies de dengue » (PSAGE dengue) (28). Ce programme regroupe à la fois les acteurs de la surveillance clinique et de la surveillance vectorielle et identifie différentes phases dans les épidémies de dengue (5 phases pour la Guadeloupe et la Martinique, 4 phases pour Saint-Martin et Saint-Barthélemy) (Figure 2)

| Phases et niveaux | Dénominations | Interprétation épidémiologique |
|---|---|---|
| Phase 1 | Transmission sporadique | Existence de cas sporadiques |
| Phase 2 – niveau 1 | Foyers isolés | Foyer(s) isolé(s) ou foyers sans lien(s) épidémiologique(s) |
| Phase 2 – niveau 2 | Circulation active du virus | Foyer(s) à potentiel évolutif ou foyers multiples avec lien(s) épidémiologique(s) entre eux |
| Phase 3 | Risque épidémique | Franchissement par les cas cliniquement évocateurs du niveau maximum attendu |
| Phase 4 – niveau 1<br>Phase 4 – niveau 2 | Epidémique<br>Epidémique à formes sévères | Epidémie confirmée (cf. critère épidémique d'alerte)<br>Epidémie avec fréquence élevée de formes sévères |
| Phase 5 | Retour à la normale | Dès le passage des cas cliniquement évocateurs en deçà du niveau maximum attendu et jusqu'au passage en phase de transmission sporadique, de foyers isolés ou de circulation active du virus |

**Figure 5.** Phases du PSAGE dengue en Guadeloupe et en Martinique (Point épidémiologique N°05/2021, Santé Publique France) (19)

L'Organisation panaméricaine de la Santé (Pan American Health Organization, PAHO), le bureau régional américain de l'OMS, a mis en place dès les années 1990, avec les réémergences de la dengue dans la région, de nombreuses résolutions visant à lutter contre *Aedes aegypti* sur

l'ensemble du continent américain. En 2016 une *Stratégie commune de prévention et de lutte contre les arboviroses* (29) a été adoptée pour la région avec, entre autres, les priorités suivantes :

- Renforcement des systèmes de surveillance vectoriels et syndromiques au niveau de chacun des États membres ;
- Renforcement des réseaux de communication et d'échanges entre les différents territoires et États de la région ;
- Renforcement des réseaux permettant de produire des données scientifiques sur l'ampleur, les tendances et les conséquences des arboviroses sur la région dans son ensemble.

L'amélioration des systèmes déjà en place est donc un enjeu de santé publique majeur pour l'ensemble des territoires américains, y compris les Antilles.

## III. Limites des systèmes de surveillance

Bien qu'efficaces, les systèmes de surveillance souffrent de défauts récurrents, en particulier le délai entre la notification d'un événement (cas clinique évocateur ou foyer de moustique) et sa remontée, et leur coût élevé lié au recueil, au traitement, à l'agrégation et à l'analyse des données collectées sur le terrain (30,31).

L'implémentation de ces mesures est variable d'un pays à l'autre, et les notifications de cas de dengue dans les Amériques restent sous-reportées à la PAHO, ce qui entraîne une sous-estimation de l'impact de la maladie dans la région et rend difficile l'anticipation de futures émergences (15,32,33).

Pour pallier ces limites, les scientifiques étudient des moyens d'améliorer les systèmes déjà en place, notamment grâce aux données de vie réelle. En épidémiologie, les données issues de l'activité Internet des utilisateurs, par leur capacité à produire des indicateurs quasi en temps réel, font partie des pistes étudiées depuis de nombreuses années pour aider à la prédiction de maladies transmissibles.

L'un des premiers outils utilisant les données d'Internet pour prédire les épidémies a été développé en 2008 par *Google* : le service *Google Flu Trends*, basé sur les requêtes des internautes américains afin d'estimer les taux d'incidence des épidémies de grippe aux États-

Unis (34). Cette étude a permis l'apparition d'une nouvelle discipline : l'infodémiologie, définie par Eysenbach comme l'étude de la distribution de l'information et de ses déterminants, en particulier provenant d'Internet, à visée de santé publique et de politique publique (35)

Depuis, les données issues du web sont régulièrement utilisées pour la surveillance de la grippe dans le monde (36,37). D'autres études ont même évalué le rôle de ces nouvelles sources de données dans la surveillance, voire la prédiction d'autres maladies infectieuses telle que la rougeole (38) ou la listériose (39). L'importance croissante des maladies émergentes et ré-émergentes a également permis d'analyser le rôle de ces sources dans le cas des maladies tropicales (40) et particulièrement dans la dengue, avec *Google Dengue Trends*, outil développé en 2014 par Gluskin, Johansson, Santillana et Brownstein (2014) (41).

Il faut noter que les données provenant d'Internet ne se limitent pas à *Google* (ou aux moteurs de recherche en général) : les réseaux sociaux (Twitter, Instagram, Facebook, Weibo en Chine…) de vidéos (YouTube) ou encore les sites d'information (*Google News*, Wikipedia) font partie des ressources exploitées dans ces études (40).

Pourtant, ces nouvelles approches, bien que prometteuses, présentent encore des limites. Premièrement, ces sources ne sont pas universelles (Twitter par exemple, qui est l'une des sources les plus utilisées dans les études, est absent dans plusieurs pays dont la Chine) et leur accessibilité varie en fonction du niveau de digitalisation des pays. Ensuite, la qualité de l'information varie d'une source à l'autre (une information vérifiée provenant d'un site d'information versus une publication sur Instagram). Enfin, la survenue d'un événement soudain, comme une épidémie, peut avoir une influence sur la production des données et donc biaiser les résultats (42). Il faut être capable de tenir compte de l'ensemble de ces limites lors de l'utilisation de ces sources de données afin de pouvoir dépasser la preuve de concept et les intégrer dans les stratégies de prévention des maladies émergentes et plus particulièrement de la dengue.

**Synthèse**

Les maladies émergentes sont devenues un enjeu majeur de santé publique depuis plusieurs années, notamment dans les régions tropicales telles que la Caraïbe. À l'heure actuelle, les systèmes de surveillance syndromique basés sur la collecte et l'analyse de données souffrent de plusieurs limites, notamment les délais entre un événement de santé et sa notification et le coût élevé de ces systèmes. Pourtant les épidémies sont de plus en plus fréquentes, avec des propagations très rapides. Par ailleurs, des alternatives basées sur les données de vie réelle sont étudiées depuis plusieurs années pour tenter de répondre aux enjeux de réactivité et de disponibilité des données.

# IV. Hypothèses et objectifs

Le potentiel des données de vie réelle a été étudié dans le cadre de système de surveillance syndromique de pathologies saisonnières telle que la grippe. Par ailleurs, les systèmes actuels de surveillance en région Caraïbe présentent des limites qu'il est nécessaire de lever afin d'améliorer leur réactivité future. L'hypothèse de départ de nos travaux est que les données de vie réelle seraient également un levier face aux limites du système de surveillance syndromique en région Caraïbe.

Néanmoins, l'utilisation de ces données se heurte à deux grands types de verrous liés à leur nature et leur exploitabilité effective.

## 1. Verrous liés à la nature des données

La *Stratégie commune de prévention et de lutte contre les arboviroses* (29) priorise à la fois le renforcement des systèmes de surveillance existants des territoires et la capacité des différents membres de la PAHO à communiquer entre eux et à partager leurs données afin de définir une surveillance globale de la région. Pour autant, on se heurte à différentes difficultés : les systèmes actuels sont hétérogènes, et ne peuvent pas toujours communiquer entre eux, les référentiels varient d'un territoire à l'autre et les sources de données de vie réelle sont multiples avec des caractéristiques propres à chacune. Afin de pouvoir appliquer cette stratégie dans la région, les pays et territoires doivent mettre en place des solutions permettant de lever ces obstacles. Il s'agit ici de mettre en place des systèmes interopérables, capables d'intégrer de

nouvelles données hétérogènes (hospitalières, Internet…) au sein de chaque territoire, mais aussi de communiquer ces données intégrées dans une région multilingue (Français, Anglais, Espagnol). Par ailleurs, ces solutions doivent pouvoir s'adapter à des systèmes d'information qui ne sont pas toujours matures en fonction des territoires concernés. Par maturité, nous entendons le degré de développement de ces systèmes, qui selon le modèle développé par HIMSS (Healthcare Information and Management Systems Society) inclut leurs dimensions techniques, fonctionnelles et organisationnelles (43).

## 2. Verrous liés à l'exploitation des données

Il existe de nombreuses sources de données de vie réelle et toutes ne sont pas pertinentes en fonction des territoires et des pathologies. Par ailleurs la qualité des données peut varier d'une source à l'autre et nécessite d'être évaluée. Les méthodes d'exploitation de ces données doivent s'adapter aux différents cas d'usage et il est nécessaire d'identifier celles qui peuvent s'appliquer à la surveillance de maladies vectorielles. Il faut également évaluer la pertinence des sources à disposition afin de les implémenter dans un système fonctionnel.

L'objectif principal de nos travaux est donc d'évaluer la place des données de vie réelle dans la surveillance syndromique et plus spécifiquement de la surveillance de la dengue dans la Caraïbe.

La suite du manuscrit se découpe en trois parties :

La deuxième partie aborde les caractéristiques des données de vie réelle et explorera les verrous liés à la nature de ces données. La troisième partie explore les verrous liés aux méthodes d'exploitation à travers un cas d'usage : la surveillance syndromique de la dengue en Martinique. Enfin, la quatrième partie comporte d'une part une discussion sur l'apport et les limites de nos travaux qui se sont principalement déroulés durant la crise sanitaire, et aborde d'autre part les leçons apprises et les perspectives que nous souhaitons poursuivre.

# Deuxième partie
# Verrous liés aux données de vie réelle

# Chapitre 1 : Les données de vie réelle

## I. Les sources de données

Les données de vie réelle désignent en santé l'ensemble des informations recueillies en dehors de la recherche clinique (44). Ces données sont générées lors des soins de routine d'un patient ou proviennent de sa vie quotidienne et reflètent donc la pratique courante, contrairement au cadre strictement réglementé des essais contrôlés randomisés (45). Parmi ces données de vie réelle, on retrouve donc les données massives en santé ou *Big Data.* Selon Weber, Mandl & Kohane (2014) (46) ces données sont extrêmement hétérogènes : elles peuvent être structurées ou non structurées, issues directement du soin ou des patients, et d'échelle variable allant de l'individu aux données agrégées.

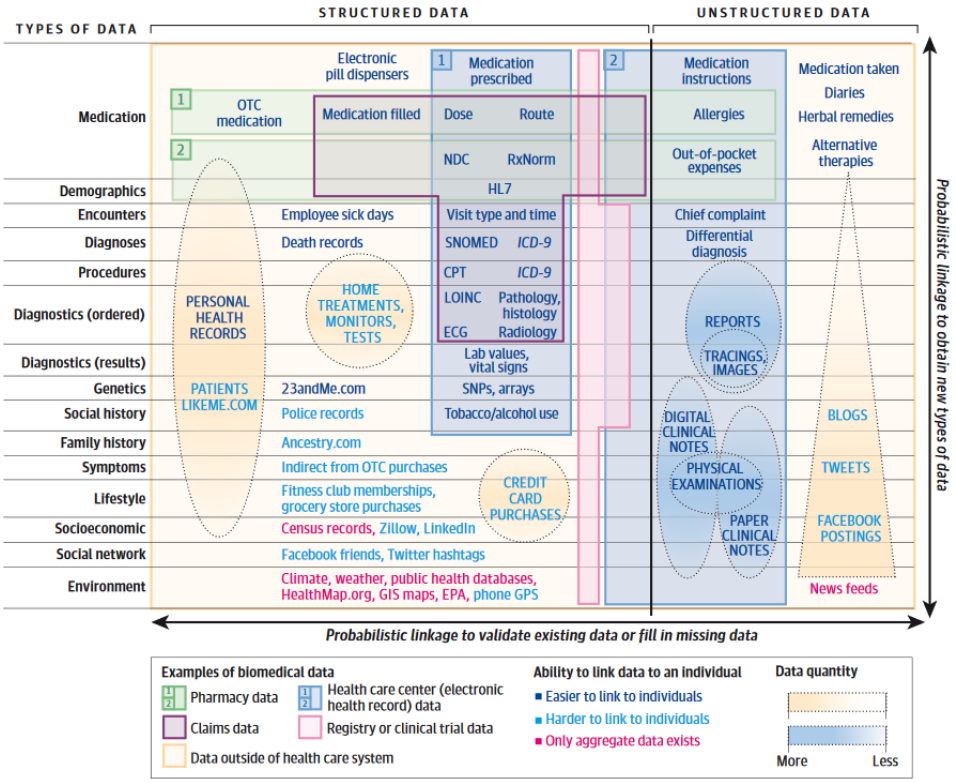



**Figure 6.** Ensemble des sources de données pouvant être liées à un individu pour une utilisation en santé (D'après Weber, Mandl & Kohane) *(46)*

Les données de vie réelle peuvent être collectées de manière spécifique, par exemple dans le cadre de la pharmacovigilance, pour évaluer les effets d'un traitement en pratique courante ou encore son effet sur la qualité de vie des patients (47). On les retrouve aussi dans la constitution de registres ou de cohortes, ou dans le cadre d'études épidémiologiques (45).

En France, on retrouve ces données à différentes échelles. Ainsi, les données issues directement du soin se retrouvent dans les hôpitaux au sein des Dossiers Patients Informatisés (DPI) et au sein des dossiers de soins dans les cabinets de médecins généralistes en ville. Cependant, elles ne sont pas ou peu centralisées, malgré de nombreuses tentatives depuis 2004. En effet, de nombreuses expérimentations ont été mises en œuvre : d'abord le « Dossier médical personnel » devenu en 2015 le « Dossier médical partagé » (DMP), dont le but est de mettre à disposition des professionnels de santé, et en particulier du médecin traitant, l'ensemble des informations médicales permettant la prise en charge optimale des patients en évitant les redondances et les pertes d'information au cours du parcours de soins (48). Depuis février 2022, « Mon espace santé » est le nouvel outil de centralisation des données de santé proposé par l'Assurance maladie à l'ensemble des assurés français (49). Cependant, cet outil privilégie l'utilisation de données non structurées telles que les comptes-rendus (d'hospitalisation, de consultation, d'imagerie…) ce qui soulève la question de l'interopérabilité et l'intégration de ces données, mais à ce jour le DMP reste au regard de la loi un outil pour le soin et aucun texte ne prévoit sa réutilisation dans le cadre de la recherche. Cette situation varie d'un pays à l'autre : ainsi, au Royaume-Uni, une partie des données issues des données de soins primaires est centralisée et structurée au sein de la base « Clinical Practice Research Datalink » (CPRD) depuis 1987, qui est régulièrement utilisée pour des projets de recherche sur données de vie réelle (50).

En revanche, concernant les données médico-administratives, l'ensemble des données issues à la fois de la prise en charge en ville et de la prise en charge hospitalière sont collectées au niveau national, via le Programme de Médicalisation du Système d'Information (PMSI) pour les hospitalisations et le Système National d'Information Inter-Régimes de l'Assurance Maladie (SNIIRAM) pour les remboursements des prescriptions en ville. Ces deux bases sont intégrées au sein du Système National des Données de Santé (SNDS), avec l'ensemble des causes médicales de décès et constituent l'une des bases médico-administratives les plus importantes au monde (51).

Par ailleurs, avec la démocratisation des smartphones et des objets connectés, les données produites par les patients (via les recherches Internet ou les applications mobiles de santé par exemple) sont devenues une nouvelle source permettant d'explorer la santé des patients en dehors des structures de soin (52) et ces données peuvent être utilisées pour la surveillance et la prévention d'épidémies (53).

Enfin, les données environnementales, liées par exemple au climat, à la pollution de l'air ou à l'exposition aux produits chimiques sont devenues une nouvelle source de données à explorer, compte tenu de l'influence des facteurs environnementaux sur l'incidence de certaines maladies (54,55). Contrairement aux données évoquées précédemment, ces données sont le plus souvent agrégées.

Afin d'exploiter les données de vie réelle et particulièrement des données massives en santé, il est donc indispensable de tenir compte de leurs caractéristiques (56).

# II. Caractéristiques des données massives

Le terme « données massives » ou *Big Data* est apparu avec la démocratisation d'Internet et l'explosion des données du web. Ces données se définissent traditionnellement par cinq dimensions principales : volume, vélocité, variété, véracité et valeur (57). Dans le domaine de la santé, elles présentent leurs propres spécificités : le volume et la vélocité ne correspondent pas toujours aux dimensions retrouvées dans les données du web. En revanche ces données sont extrêmement sensibles et leur exploitation exige un haut niveau de sécurité et de confidentialité.

## 1. Volume

Avec l'augmentation croissante des différentes sources de données et en particulier de l'Internet des Objets (via les objets connectés) et des réseaux sociaux (via les smartphones), le volume des données collectées est en croissance exponentielle qui nécessite une évolution constante des technologies de stockage. En santé, le volume de données varie en fonction de la source. Ainsi les données *-omics* ou d'imagerie représentent des volumes beaucoup plus importants que les données issues des DPI. Par ailleurs, le nombre de patients concernés par le traitement de données va également influencer leur volume.

## 2. Vélocité

La vélocité fait référence à la fois à la vitesse de production des données et à leur traitement. En particulier dans le cas des dispositifs médicaux générant des données, il faut être capable de les analyser quasiment en temps réel.

## 3. Variété

Avec la multiplicité des sources de données et des formats, les données disponibles ont des formats hétérogènes : structurées, semi-structurées, non structurées. Par ailleurs, une même information peut être exprimée de plusieurs façons (par exemple un code CIM-10 diagnostic structuré et un diagnostic non structuré dans un compte-rendu d'hospitalisation). Différentes approches sont par conséquent possibles (par exemple, la standardisation) pour pouvoir interpréter ces données de façon cohérente.

## 4. Véracité

La véracité fait référence à la qualité des données traitées. En effet, si les données collectées sont de qualité insuffisante, les conclusions générées à partir de leur analyse peuvent être imprécises, voire erronées. La définition la plus utilisée pour décrire la notion de qualité des données est leur « capacité à répondre aux besoins exprimés et implicites lorsqu'elles sont utilisées dans des conditions spécifiées » (58). Les problèmes liés à la qualité des données peuvent apparaître à différents niveaux :

- Au niveau de la source de données : les sources choisies ne sont pas fiables ou inconstantes ; les données peuvent être contradictoires ou dépassées ;
- Au niveau de la génération des données : une collecte manuelle peut générer des erreurs humaines ou des doublons ; la collecte peut être incomplète (données manquantes) ;
- Au niveau du traitement des données : la transmission et l'intégration des données peuvent être également source d'erreurs.

Dans le cadre de l'utilisation des données massives en santé, des données de mauvaise qualité peuvent affecter la prise en charge des patients, le contrôle constant de la qualité des données est donc indispensable à chacun des niveaux concernés (vérification et correction des sources, contrôle des méthodes de recueil et vérification des doublons, contrôle des processus d'intégration et de transformation des données).

Plusieurs indicateurs permettent d'évaluer la qualité des données de santé, en particulier : la précision, la complétude, la constance, et la cohérence (59).

## 5. Valeur

L'utilisation de ces données doit apporter une valeur ajoutée par rapport aux systèmes traditionnels de recueils de données. Dans le cas de la Santé, l'utilisation des données massives permet selon Pastorino et al (2019) (60) :

- d'augmenter les diagnostics précoces et l'efficacité des traitements ;
- d'améliorer la prévention des maladies en identifiant les facteurs de risque ;
- d'améliorer la pharmacovigilance et la sécurité des patients.

## 6. Sécurité et confidentialité

La sécurité et la confidentialité des données de santé sont des enjeux majeurs lors de leur exploitation. Dans l'idéal, ces données devraient être anonymisées (c'est-à-dire rendre impossible toute identification d'un individu de manière irréversible) afin d'assurer une véritable protection de la confidentialité. Cependant, dans le cadre des données de santé, l'anonymisation dénature ces données et donc limite les possibilités d'exploitation : elles vont être plutôt pseudonymisées, en remplaçant tous les éléments permettant l'identification directe des individus par des données indirectement identifiantes (61).

En ce qui concerne la sécurité des données, elle est assurée par leur stockage dans des environnements sécurisés, le contrôle des accès ainsi que leur traçabilité. Les mesures de sécurité les plus fréquentes sont le chiffrement des transferts ou l'authentification à plusieurs niveaux (62). D'autres approches basées sur les technologies blockchain (c'est-à-dire un stockage et une transmission de l'information décentralisées et tracées, l'ensemble étant sécurisé par cryptographie) sont envisagées mais à l'heure actuelle sont expérimentales (63).

Ces mesures de protection sont indispensables et sont soumises à des réglementations variant d'un pays à l'autre : en Europe, c'est le Règlement Général sur la Protection des Données (RGPD) (64) qui est le garant de la confidentialité des données, tandis qu'aux États-Unis c'est le Health Insurance Portability and Accountability Act (HIPAA) (65).

### Synthèse

Les données de vie réelle sont multi-échelles et hétérogènes mais complémentaires et l'intégration de ces différentes sources permet de faciliter l'échange d'informations et d'avoir une vision globale du système de soins et de la santé des patients. Les multiples dimensions de ces données doivent toujours être prises en compte lors des stratégies d'intégration afin de pouvoir les réutiliser et les exploiter de façon transversale.

Dans le chapitre 2, nous allons présenter les efforts d'intégration de données hétérogènes ainsi que l'intérêt de cette intégration dans le cadre de la surveillance syndromique.

# Chapitre 2 : Intégration des données hétérogènes

Le rapport du National Research Council Commitee (2011) (66) compare l'intégration des données de santé à la méthodologie utilisée dans les Systèmes d'Information Géographiques (SIG), composés de différentes couches d'information superposées. En effet, les SIG sont capables de mettre en relation des données très éloignées afin d'avoir une vision globale de notre environnement. Dans le cas de l'intégration des données de vie réelle, les données intégrées autour des patients permettent d'analyser les informations de santé autour des individus à différents niveaux, allant du génome, avec les données *-omics* à l'environnement avec l'exposome (Figure 7).

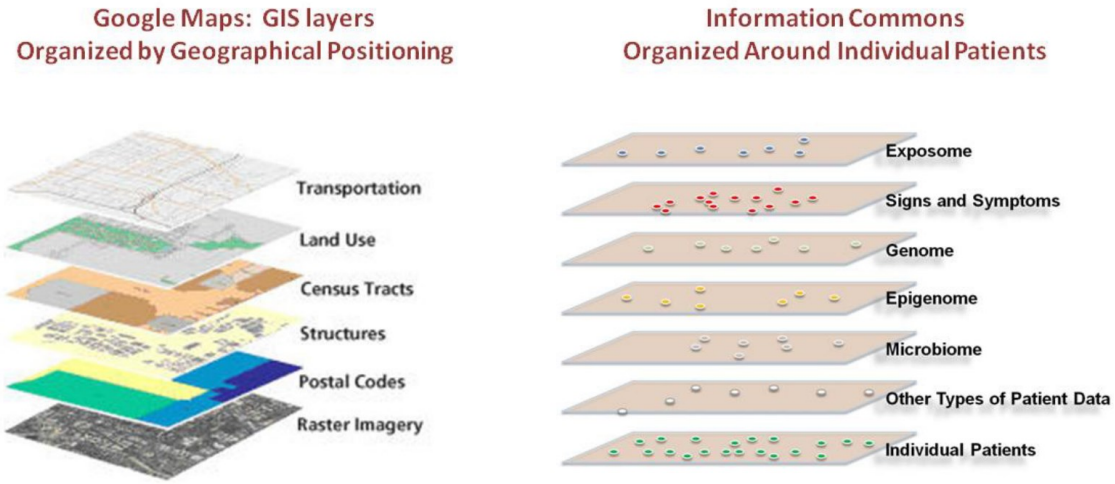



**Figure 7.** Organisation d'un système d'information centralisant les données de santé intégrées autour du patient. Cette méthodologie est similaire à celle des Systèmes d'Information Géographiques de type *Google Maps* (d'après le rapport du National Research Council Commitee) *(66)*.

L'intégration des données peut se faire à différentes échelles, en fonction des besoins :

- à l'échelle d'un établissement
- à l'échelle de plusieurs établissements, voire de plusieurs territoires ou pays

# I. Intégration à l'échelle d'un établissement

L'intégration des données au sein d'un même établissement de santé est le premier niveau d'intégration à mettre en place afin d'exploiter ces dernières. Depuis plusieurs années, les entrepôts de données de santé (EDS) permettent le décloisonnement des données du soin pour une réutilisation secondaire. L'outil le plus répandu est l'entrepôt open source i2b2 (Informatics for Integrating Biology and the Bedside) (67), développé à Harvard dans les années 2000 et basé sur le schéma en étoile (68) et une table centrale dite « table de faits » contenant l'ensemble des données atomiques. Cette table centrale est reliée aux tables de dimensions, chacune d'entre elles représentant un axe d'analyse. Cet ensemble forme le modèle « en constellation ». Cependant, d'autres initiatives existent, comme l'entrepôt eHOP (69,70), développé par l'équipe DOMASIA (DOnnées MAssives en Santé et système d'Information Apprenant) du Laboratoire Traitement du Signal et de l'Image (LTSI). La particularité de cet outil, le plus déployé dans les hôpitaux français, est qu'il tient compte du contexte dans lequel a été produite la donnée d'origine. Pour cela, cette technologie : (i) s'appuie sur les standards d'interopérabilité utilisés par les données sources ; (ii) intègre une dimension documentaire dans son modèle, ce qui permet de conserver le contexte reliant différentes données atomiques, et (iii) intègre des bases de connaissances métiers (par exemple Thériaque) pour permettre l'indexation et le traitement des données de faits.

# II. Intégration à l'échelle de plusieurs établissements ou territoires

L'intégration au sein d'un établissement constitue le premier niveau d'intégration de données de santé, mais de nombreuses initiatives vont au-delà de ce niveau d'intégration et de partage de données en proposant des solutions techniques permettant l'exploitation multicentrique d'entrepôts de données et la création de réseaux de données cliniques pour la recherche.

Sur le plan international, le réseau SHRINE (Shared Health Research Information Network), basé sur l'entrepôt i2b2 permet d'interroger de façon fédérée des entrepôts i2b2 hébergés dans différents hôpitaux (71,72). Il a été utilisé pour la constitution de cohortes (73,74), dans des essais thérapeutiques multicentriques (75) ou pour la constitution de registres nationaux (76).

En Europe, la plateforme EHR4CR (Electronic Health Records for Clinical Research), qui associait des partenaires de l'industrie pharmaceutique et des hôpitaux universitaires européens utilisait également des réseaux d'entrepôts issus de pays différents afin de faciliter le recrutement pour les essais thérapeutiques et une meilleure maîtrise de la notification des effets secondaires (77,78).

Ces avancées technologiques entraînent également des évolutions en termes de gouvernance, avec la création de consortiums pour faciliter le partage de ces données. Ainsi, aux Etats-Unis le « Patient-Centered Outcomes Research Institute » (PCORI) créé en 2013, est responsable du financement du « PCORI-funded National Patient-Centered Clinical Research Network » (PCORNet) (79,80), qui centralise les données des Dossiers Patients Informatisés (DPI) de plus de 100 millions d'américains provenant d'une centaine d'établissements hospitaliers. Le réseau PCORNet regroupe à la fois des « Clinical Data Research Networks » (CDRN) (81), qui sont des réseaux contenant des données provenant du soin (données du DPI, données de remboursement…) et des « Patient-Powered Research Networks » (PPRN) (82), qui sont eux basés sur des Patient-Reported Outcomes (PRO), c'est-à-dire des éléments de santé rapportés directement par les patients, notamment via des questionnaires ou des applications sur les smartphones (83). Ici, les établissements participants au réseau PCORNet utilisent un modèle commun de données (PCORnet Common Data Model) (84) afin de faciliter l'accessibilité des sources. D'autres modèles communs de données, comme le modèle OMOP (Observational Medical Outcomes Partnership) (85,86) sont adoptés par des consortiums afin d'améliorer l'intégration et le partage de données sur le plan international.

En France, le Système National des Données de Santé (SNDS), issu du chaînage de différentes bases médico-administratives existantes a été créé en 2017 afin de favoriser les études sur l'ensemble de la population française (51). Le SNDS contient à l'heure actuelle : les données de remboursement de l'Assurance Maladie (SNIIRAM), les données d'hospitalisation (PMSI) et les causes médicales de décès (base du CépiDC). Il contiendra à terme les données relatives au handicap et un échantillon de données provenant des organismes complémentaires d'Assurance Maladie. En 2019, le Health Data Hub (HDH) a été créé pour poursuivre cette mission d'intégration des différentes sources de données : il sert de guichet unique pour les demandes d'accès aux données du SNDS, il propose de mettre à disposition de la communauté un catalogue de données, il propose une plateforme sécurisée pour le stockage et le traitement des données et enfin il renforce les relations entre tous les acteurs du traitement de données en France (87).

Par ailleurs, la plateforme INSHARE (INtegrating and Sharing Health dAta for REsearch) a permis de regrouper au sein d'une architecture et d'une gouvernance commune des données et des partenaires provenant de différents types d'établissements : données provenant de l'entrepôt eHOP (CHU de Rennes et CHU de Brest), données provenant de registres (Registre général des cancers de Poitou-Charentes, Registre REIN) et données du SNDS (88). Cette plateforme a ensuite été à l'origine de différents cas d'usage (trajectoires de soins, surveillance d'effets secondaires…) (89,90) permettant une exploitation multicentrique de ces données.

On retrouve également à l'échelle interrégionale des réseaux d'EDS tels que le Réseau Interrégional des Centres de Données Cliniques (RiCDC) dans la région Grand-Ouest qui regroupe un ensemble d'hôpitaux et de Centres de Lutte Contre le Cancer utilisant l'entrepôt eHOP. Ce réseau regroupe à la fois une technologie commune, via le choix d'un même EDS et une gouvernance harmonisée puisque le pilotage se fait au niveau de l'interrégion. Il s'est doté d'une plateforme de traitement des données commune (le Ouest Data Hub) permettant l'intégration et l'exploitation multi-échelle des données provenant des entrepôts hospitaliers et d'autres sources de données externes (91).

## III. Apports et limites des EDS dans la surveillance syndromique

D'abord utilisés pour la recherche clinique et la création de cohortes, les EDS sont de plus en plus utilisés dans le cadre de la surveillance syndromique de pathologies infectieuses, grâce à leur capacité à fournir des données quasi en temps réel.

Par exemple, dans le cadre de la surveillance des épidémies saisonnières de grippe, les données issues des EDS sont fortement associées aux syndromes grippaux (92) et sont capables de prédire les taux d'incidence des syndromes grippaux, avec des performances supérieures aux modèles basés uniquement sur les données du web (93). Ils permettent également d'estimer le poids de l'épidémie de grippe sur les hôpitaux afin d'améliorer le pilotage quasi en temps réel en temps de crise (94).

La pandémie de COVID-19 a montré l'importance des outils numériques pour la surveillance en temps réel de phénomènes épidémiques (95). En ce qui concerne les EDS, cette crise sanitaire a permis la création de consortiums internationaux tels que le 4CE consortium (Consortium for Clinical Characterization of COVID-19 by EHR) (96) qui regroupe des

données issues de 96 hôpitaux dans 5 pays (États-Unis, France, Italie, Allemagne, Singapour). Les hôpitaux membres de ce consortium utilisent l'EDS i2b2 ou le modèle de données OMOP et des requêtes fédérées qui permettent la surveillance des cas, l'étude de trajectoires de patients ou encore la caractérisation de phénotypes (97).

En France, les consortiums « AP-HP COVID CDR Initiative. » et « AP-HP/Universities/Inserm COVID-19 research collaboration » ont également utilisé la technologie d'EDS afin de mener différentes études autour du COVID-19, et ont montré la plus-value de ces outils dans des contextes de crise sanitaire (98,99). L'EDS eHOP a été quant à lui utilisé pour développer un outil d'aide au diagnostic de la COVID au tout début de la pandémie (100).

Cependant, la réutilisation secondaire de données dans un contexte de crise sanitaire a montré ses limites. Ainsi, deux publications dans des revues majeures (*New England Journal of Medicine* (101) et *The Lancet* (102)) basées sur des DPI ont été rétractées, notamment à cause d'interrogations en termes de qualité des données. En effet, des données sources de qualité sont indispensables pour être capables de tirer des conclusions pertinentes à partir de leur analyse. Des éléments clés méthodologiques tels que le type de données, le processus de recueil, la complétude des données ou leur variabilité doivent être contrôlés dans toute réutilisation secondaire (103). Or, en période de crise, ces étapes peuvent être impactées, en particulier l'exhaustivité des données et le contrôle qualité après la collecte.

Par ailleurs, le déploiement et l'exploitation des EDS nécessitent des structures et des moyens, et cette réponse quasi immédiate des systèmes de santé face à la pandémie n'a été possible que dans des établissements et/ou territoires ayant des systèmes d'information matures (43), c'est-à-dire justement dotés d'EDS. D'une façon générale, le COVID-19 a été un accélérateur pour la transition numérique de nombreux territoire et les outils numériques ont été un atout dans la capacité des systèmes de santé à répondre rapidement à la crise (104), mais a aussi mis en évidence les disparités entre pays en matière de santé digitale. Ainsi, ce recours aux technologies du numérique a mis en évidence une fracture numérique à deux niveaux : premièrement, au niveau des accès aux technologies existantes qui sont très variables même au sein des pays les plus avancés, et deuxièmement au niveau de la maîtrise de ces outils digitaux par les individus, qu'ils soient professionnels de santé ou patients. Pourtant, de plus en plus d'experts s'accordent sur l'importance de ces outils numériques en situation d'épidémie (95).

Même s'ils ne sont pas entièrement matures, les systèmes d'information en santé doivent donc être capables d'être flexibles et réactifs afin de pouvoir répondre aux enjeux de surveillance et d'organisation au cours des épidémies. C'est le cas de l'Amérique du Sud et de la Caraïbe, qui sont des régions avec de grandes inégalités d'accès aux outils de santé digitale (105). Pourtant, ces régions endémo-épidémique pour de nombreuses maladies émergentes – dont la dengue– doivent pouvoir développer des outils simples et immédiats de réutilisation de données en attendant la maturité de leurs systèmes.

**Synthèse**

L'intégration des données de santé multi-sources est la première étape indispensable à leur exploitation. Elle peut se faire à plusieurs échelles : au niveau d'un établissement, d'un territoire, d'un pays et même de plusieurs pays.

Lorsque cette intégration est déjà existante dans des systèmes d'information matures, elle permet leur réutilisation quasi immédiate en cas d'épidémie. Quand ce n'est pas le cas, ces systèmes doivent malgré tout être capables d'être réactifs en situation de crise sanitaire.

Ce constat est celui du CHU de Martinique qui ne dispose pas encore d'EDS fonctionnel. Dans ce contexte, nous avons mis en place l'architecture d'un système de surveillance syndromique basé sur la réutilisation des données de vie réelle de l'hôpital. Ce système a ensuite été utilisé pour le développement et à la production de tableaux de bord de suivi à partir de l'outil. Ce travail a fait l'objet d'une publication qui est présentée ci-dessous.

## Article 1 : Health informatics support for outbreak management: How to respond without an electronic health record?

Cet article présente un processus d'intégration de données mis en place pendant une épidémie (ici, le COVID-19) au sein d'un système d'information hospitalier n'étant pas entièrement à maturité. Les données intégrées sont indispensables au suivi en temps réel à la fois pour les patients hospitalisés et les patients à domicile. Les deux plateformes décrites peuvent être mises en place avec des outils simples et permettent de lever le verrou de l'intégration de données en tenant compte de la fracture numérique présente dans les régions de la Caraïbe.

L’architecture développée pour la plateforme de suivi intra-hospitalier (COVCHUM) a pu ensuite être réutilisée pour la surveillance hospitalière de la dengue lors de l’accélération de l’épidémie de dengue en juillet 2020.

En ce qui concerne la plateforme de suivi à domicile COVID-SAMU (Figure 8), elle s’est révélée également utile pour l’identification de patients atteints de la dengue lors de la 2$^{e}$ vague de COVID-19 (qui était concomitante à l’épidémie de dengue) puisque la dengue et le COVID-19 ont des présentations cliniques initiales similaires.

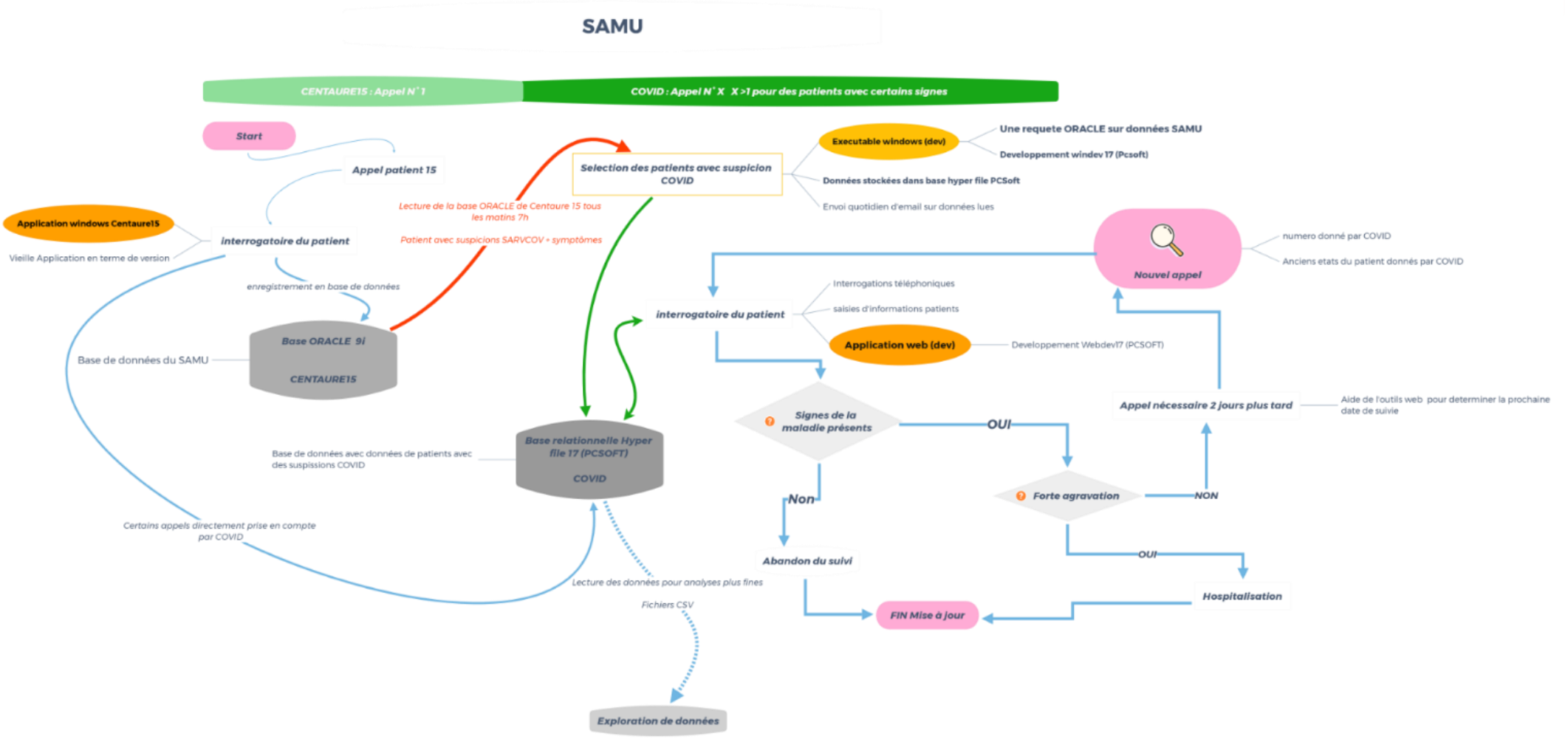
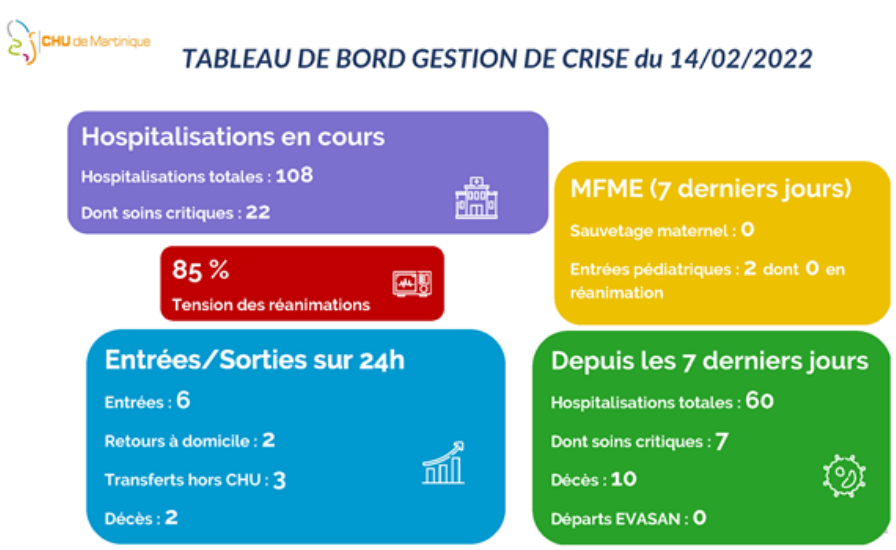
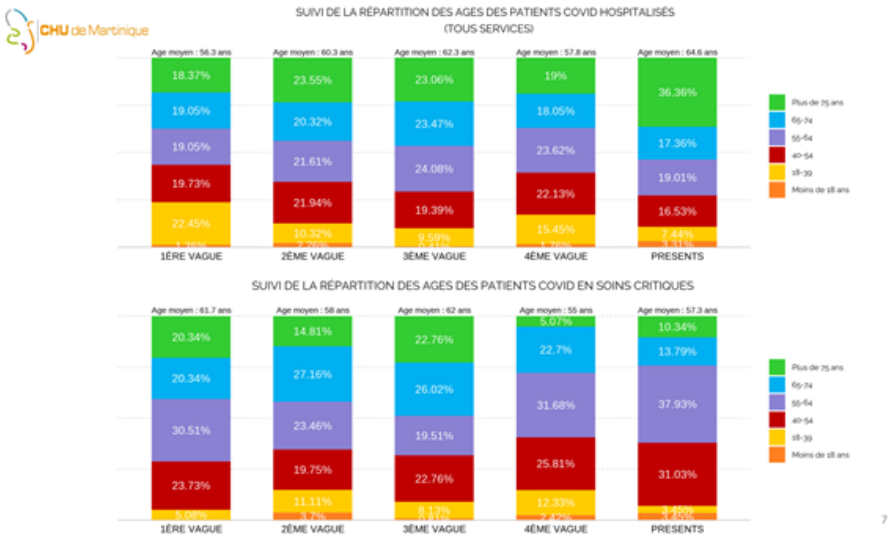
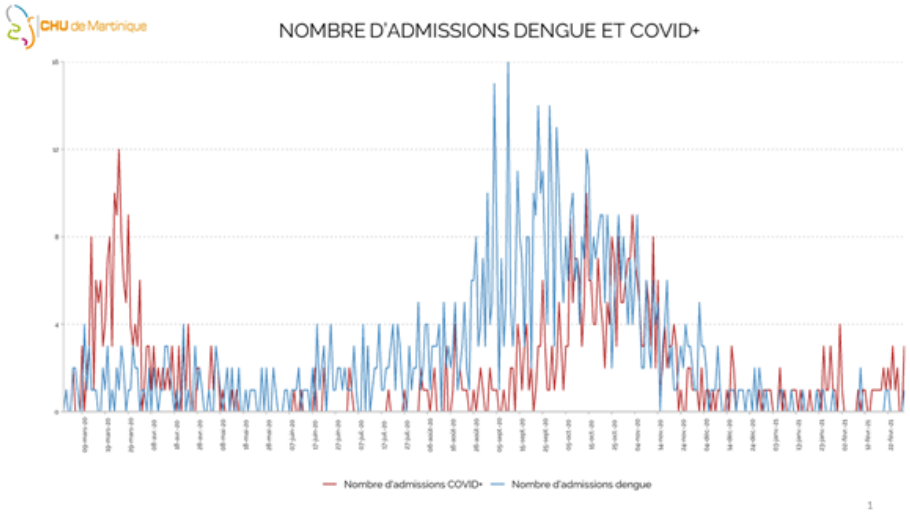
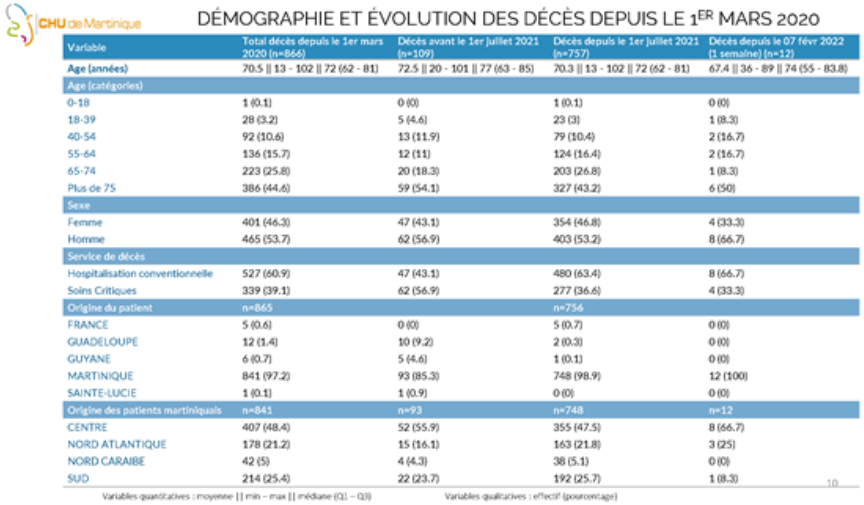



**Figure 8.** Architecture de la plateforme de suivi à domicile COVID-SAMU




| Variable | Total décès depuis le 1er mars 2020 (n=866) | Décès avant le 1er juillet 2021 (n=109) | Décès depuis le 1er juillet 2021 (n=757) | Décès depuis le 07 févr 2022 (1 semaine) (n=12) |
|---|---|---|---|---|
| **Age (années)** | 70.5 \|\| 13 - 102 \|\| 72 (62 - 81) | 72.5 \|\| 20 - 101 \|\| 77 (63 - 85) | 70.3 \|\| 13 - 102 \|\| 72 (62 - 81) | 67.4 \|\| 36 - 89 \|\| 74 (55 - 83.8) |
| Age (catégories) | | | | |
| 0-18 | 1 (0.1) | 0 (0) | 1 (0.1) | 0 (0) |
| 18-39 | 28 (3.2) | 5 (4.6) | 23 (3) | 1 (8.3) |
| 40-54 | 92 (10.6) | 13 (11.9) | 79 (10.4) | 2 (16.7) |
| 55-64 | 136 (15.7) | 12 (11) | 124 (16.4) | 2 (16.7) |
| 65-74 | 223 (25.8) | 20 (18.3) | 203 (26.8) | 1 (8.3) |
| Plus de 75 | 386 (44.6) | 59 (54.1) | 327 (43.2) | 6 (50) |
| Sexe | | | | |
| Femme | 401 (46.3) | 47 (43.1) | 354 (46.8) | 4 (33.3) |
| Homme | 465 (53.7) | 62 (56.9) | 403 (53.2) | 8 (66.7) |
| Service de décès | | | | |
| Hospitalisation conventionnelle | 527 (60.9) | 47 (43.1) | 480 (63.4) | 8 (66.7) |
| Soins Critiques | 339 (39.1) | 62 (56.9) | 277 (36.6) | 4 (33.3) |
| Origine du patient | n=865 | | n=756 | |
| FRANCE | 5 (0.6) | 0 (0) | 5 (0.7) | 0 (0) |
| GUADELOUPE | 12 (1.4) | 10 (9.2) | 2 (0.3) | 0 (0) |
| GUYANE | 6 (0.7) | 5 (4.6) | 1 (0.1) | 0 (0) |
| MARTINIQUE | 841 (97.2) | 93 (85.3) | 748 (98.9) | 12 (100) |
| SAINTE-LUCIE | 1 (0.1) | 1 (0.9) | 0 (0) | 0 (0) |
| Origine des patients martiniquais | n=841 | n=93 | n=748 | n=12 |
| CENTRE | 407 (48.4) | 52 (55.9) | 355 (47.5) | 8 (66.7) |
| NORD ATLANTIQUE | 178 (21.2) | 15 (16.1) | 163 (21.8) | 3 (25) |
| NORD CARAIBE | 42 (5) | 4 (4.3) | 38 (5.1) | 0 (0) |
| SUD | 214 (25.4) | 22 (23.7) | 192 (25.7) | 1 (8.3) |

Variables quantitatives : moyenne || min – max || médiane (Q1 – Q3)  Variables qualitatives : effectif (pourcentage)

**Figure 9.** Exemple de tableau de bord produit à partir de la plateforme COVCHUM

*Journal of the American Medical Informatics Association*, 27(11), 2020, 1828–1829
doi: 10.1093/jamia/ocaa183
Advance Access Publication Date: 6 August 2020
Correspondence

Correspondence

# Health informatics support for outbreak management: How to respond without an electronic health record?

**Emmanuelle Sylvestre,[1,2,3,4] René-Michel Thuny,[5] Elsa Cecilia-Joseph,[4] Papa Gueye,[6] Cyrille Chabartier,[7] Yannick Brouste,[8] Hossein Mehdaoui,[7] Fatiha Najioullah,[9,10] Sandrine Pierre-François,[3] Sylvie Abel,[3] André Cabié,[3,10,11] and Moustapha Dramé[12]**

[1]U1099, French Institute of Health and Medical Research, Rennes, France, [2]Laboratoire Traitement du Signal et de l'Image, Université de Rennes 1, Rennes, France, [3]Department of Infectious Diseases, Centre Hospitalier Universitaire de Martinique, Fort-de-France, Martinique, [4]Centre de Données Cliniques, Centre Hospitalier Universitaire de Martinique, Fort-de-France, Martinique, [5]Information Technology Department, Centre Hospitalier Universitaire de Martinique, Fort-de-France, Martinique, [6]SAMU de Martinique, Centre Hospitalier Universitaire de Martinique, Fort-de-France, Martinique, [7]Intensive Care Unit, Centre Hospitalier Universitaire de Martinique, Fort-de-France, Martinique, [8]Department of Emergency Medicine, Centre Hospitalier Universitaire de Martinique, Fort-de-France, Martinique, [9]Virology Laboratory, Centre Hospitalier Universitaire de Martinique, Fort-de-France, Martinique, [10]EA 4537, French Institute of Health and Medical Research, Fort-de-France, Martinique, [11]CIC-1424, Centre Hospitalier Universitaire de Martinique, French Institute of Health and Medical Research, Fort-de-France, Martinique, and [12]Department of Clinical Research and Innovation, Centre Hospitalier Universitaire de Martinique, Fort-de-France, Martinique

*Corresponding Author: Emmanuelle Sylvestre, MD, MSc, INSERM, U1099, F-35000 Rennes, France; emmanuelle.sylvestre@chu-martinique.fr

Received 26 June 2020; Editorial Decision 17 June 2020; Accepted 18 July 2020

**Key words:** digital divide, electronic health record, health informatics, COVID-19, pandemic

To the editor,

The world is facing an unprecedented health crisis in 2020 with the coronavirus disease 2019 (COVID-19) pandemic. Reeves et al's[1] article underlined the importance of the electronic health record (EHR) and health informatics in general to support outbreak management. They proposed several recommendations heavily based on the EHR to help hospitals improve their response in this unique situation.

This article is extremely relevant for the United States, as most American within the healthcare system have their data recorded electronically. According to the Office of the National Coordinator for Health Information Technology report, as of 2015, 96% of non-federal acute care hospitals and 78% of office-based physicians had adopted certified health information technology (IT).[2]

Thus, with a fully functioning EHR, the authors were able to implement screening tools to help proper triage, ordering tools for accelerated biology and imaging exams and even clinical decision support. All of those EHR enhancements followed COVID-19 monitoring guidelines set by institutions and were a major help for outbreak management.

The use of EHR as a potential public health tool has been studied for years,[3] and with the COVID-19 pandemic, many institutions worldwide have tried to leverage its full potential to accelerate their response.

However, some health institutions are still struggling to entirely digitize their health data. In France, according to the French Office of Health Care Supply, only 70% of hospitals have a fully functioning EHR.[4] Therefore, how can the 30% left still be efficient during this pandemic?

Hospitals can still rely on classic outbreak monitoring, based on manual reporting and contact tracing, but with the scale of COVID-19, these techniques have shown their limitations, especially regarding real-time transmission of information.

For this reason, after initially using manual outbreak managing techniques at the Martinique University Hospital, we decided to develop an alternative solution. Indeed, our hospital is the only academic hospital in Martinique, which is a French overseas territory located in the Caribbean. Moreover, Martinique University Hospital is one of the French hospitals without a fully functioning EHR.

Thus, the clinical informatics team and the COVID Crisis Team collaborated to develop and implement 2 simple managing tools. The Clinical Informatics Team included a medical informatics doctor, an epidemiologist, an engineer specialized in interoperability, and a scientist specialized in modeling. The COVID Crisis Team included 2 infectious disease physicians, an emergency room physician, an intensive care unit physician, and a bed manager.

The aim of these tools was to (1) be able to build and implement them quickly with limited resources; (2) use our existing developing tools, or open-source alternatives if not available; and (3) be able to create and distribute real-time reports.

We managed to build 2 databases in less than a week.

The first database (COVID-SAMU) is a triage database used for monitoring outpatient cases, with a phone call schedule based on national monitoring guidelines. The database has information on all outpatient cases, including their address, age, underlying diseases, and different symptoms. Sociodemographic data from patients with COVID-like symptoms are first automatically integrated from the hospital triage software. Then, we developed a web application in which each clinician can fill specific forms to monitor COVID symptoms and their evolution at the time of each phone call.

We decided to heavily rely on this form of outpatient monitoring rather than self-reporting (eg, based on a smartphone application) because of our population characteristics (Martinique is one of the oldest French territories).

The second database (COVCHUM) is for hospitalized patients. This database also integrates the few digitized data available (administrative data, reimbursement claims, and laboratory test reports). As for the COVID-SAMU database, we developed a web application and COVID-specific forms for clinicians. In this case, we needed to be able to integrate quickly the most important data for COVID monitoring despite the lack of interoperability between our different digitized systems.

Because our administrative data is fully digitized, we were able to link patients throughout the different systems with their hospital ID. Symptoms were mapped to the French International Classification of Diseases, 10th Edition when possible, and procedures were mapped to the French Procedure Terminology.

Both databases are implemented with WINDEV (PC Soft, Montpellier, France) because it allowed us to automatically integrate data from our hospital framework (all of our hospital software relies on the Oracle database management system [Oracle, Redwood Shores, CA]). We also used WEBDEV (PC Soft) to develop the web-based applications because we wanted to be able to deploy them hospital-wide in a very short time, even with a very small team. Finally, both COVID databases allow to perform queries using structured query language and extract structured data in comma-separated values form, which helps us create real-time reports.

We still wanted to comply as much as possible with health IT guidelines. As a result, we focused on interoperability, standardized terminologies, and automatic data collection when possible. We also implemented simple rule-based natural language processing algorithms to be able to extract unstructured data from clinical notes.

Despite our limited resources and our lack of an existing adequate informatics framework, we managed to implement relatively simple tools, which helped us improve our ability to rapidly respond to the evolving situation.

The EHR is an essential tool for COVID-19 management, but even without it, we can still develop alternative solutions that can tremendously help hospitals with limited resources and without state-of-the-art health IT. We should leverage these solutions to help reduce the impact of the digital divide in health care, especially in time of crisis.[5]

## FUNDING

This work was supported by the European Regional Development Fund ERDF and the Martinique Territorial Authority (CTM, Collectivité Territoriale de la Martinique) grant number MQ0019176.

## AUTHOR CONTRIBUTIONS

ES drafted the manuscript and helped designing the software. R-MT wrote the code for the software and reviewed the manuscript. EC-J helped designing the software and revised the manuscript. PG, CH, YB, HM, FN, SP-F, SA, and AC gave all clinical recommendations, tested the tool, and reviewed the manuscript. AC and MD helped with the software design and reviewed and revised the manuscript.

## CONFLICT OF INTEREST STATEMENT

None declared.

## REFERENCES

1. Reeves JJ, Hollandsworth HM, Torriani FJ, *et al*. Rapid response to COVID-19: health informatics support for outbreak management in an academic health system. *J Am Med Inform Assoc* 2020; 27 (6): 853–9. doi: 10.1093/jamia/ocaa037.
2. Office of the National Coordinator for Health Information Technology. 2018. Report to Congress: Annual Update on the Adoption of a Nationwide System for the Electronic Use and Exchange of Health Information Accessed June 3, 2020.
3. Mahmood S, Hasan K, Colder Carras M, *et al*. Global preparedness against COVID-19: we must leverage the power of digital health. *JMIR Public Health Surveill* 2020; 6 (2): e18980.
4. Direction Générale de l'Offre de Soins. Atlas des systèmes d'information hospitaliers. Ministère Solidar. Santé. 2020. https://solidarites-sante.gouv.fr/systeme-de-sante-et-medico-social/e-sante/sih/article/atlas-des-systemes-d-information-hospitaliers Accessed June 3, 2020.
5. Ramsetty A, Adams C. Impact of the digital divide in the age of COVID-19. *J Am Med Inform Assoc* 2020; 27 (7): 1147–8. doi: 10.1093/jamia/ocaa078.

# Chapitre 3 : Partage des données au sein de territoires multilingues

Dans le chapitre précédent, nous avons étudié les approches d'intégration de données à différentes échelles. Dans le chapitre suivant, nous étudierons les problématiques spécifiques liées à l'intégration de données dans un contexte international.

Dans le cadre de la *Stratégie commune de prévention et de lutte contre les arboviroses* (29), les membres de la PAHO sont appelés à communiquer entre eux et à partager leurs données afin de définir une surveillance globale de ces maladies dans la région.

Or, la Caraïbe est une région multilingue où se côtoient de nombreuses langues, les principales étant l'espagnol, le français et l'anglais. Par ailleurs, la plupart des îles de la Caraïbe parlent des variations de créoles hérités de l'anglais, du français et de l'espagnol.

Afin de répondre à ces exigences, les territoires doivent donc être capables non seulement d'intégrer la remontée des données de surveillance au sein des établissements, via les méthodologies d'intégration vues au chapitre précédent, mais aussi de consolider cette remontée sur le plan supranational, au sein de cette région multilingue. Les systèmes des différents territoires de la région doivent être capables d'être interopérables, malgré les différences de langage.

L'interopérabilité sémantique multilingue permet de transformer des mots, phrases ou textes provenant de langues différentes dans un format leur permettant de communiquer entre elles (106). Il existe plusieurs façons d'obtenir cette interopérabilité. Par exemple, l'analyse syntaxique et sémantique (*semantic parsing*) (107) va mettre en évidence la structure des phrases et le rôle sémantique de chacun de ses éléments, afin de transformer du langage naturel en éléments logiques (*logical forms*) pour faciliter la correspondance entre éléments de langues différentes. Une autre solution est l'alignement d'ontologies : dans ce cas, les concepts sont déjà structurés (via chacune des ontologies respectives) et la position du concept au sein de l'ontologie va également jouer un rôle dans l'alignement des ontologies (108). Cependant, ce processus est généralement limité à des ontologies conçues dans la même langue (le plus souvent l'anglais).

L'OMS a développé depuis 1948 la Classification Internationale des Maladies (CIM), qui est mondialement utilisée pour l'enregistrement de la morbidité et de la mortalité au niveau

mondial (109). Actuellement, la majorité des pays membres de l'OMS et de la PAHO utilisent également la CIM-10 pour le codage des diagnostics dans le cadre du financement des hôpitaux basé sur le « Diagnosis Related Group » (DRG) américain (110). Des adaptations de langue existent, ainsi que des modifications de la terminologie en elle-même, en particulier l'ICD-10-CM (ICD-10 Clinical Modification) qui correspond à l'adaptation américaine de la CIM-10.

La CIM-10 et l'ICD-10 CM étant largement utilisée dans la région, aligner les différentes variations linguistiques de ces dernières permettrait donc de standardiser les échanges entre les territoires.

Ce processus de standardisation est celui de l'alignement ontologique multilingue, qui permet d'aligner des ontologies, même lorsqu'elles sont issues de langues différentes, afin de les rendre interopérables (108).

**Synthèse**

Il existe différentes façons d'obtenir l'interopérabilité sémantique afin de communiquer une information entre des sources de langue différente dans des régions multilingues telles que la Caraïbe.

L'une des approches possibles est l'alignement ontologique multilingue. Pour répondre à cette problématique, nous avons développé une méthodologie d'alignement semi-automatisé entre deux terminologies de langues différentes. Ce travail a fait l'objet d'une publication présentée ci-dessous.

## Article 2 : A Semi-Automated Approach for Multilingual Terminology Matching: Mapping the French Version of the ICD-10 to the ICD-10 CM

L'article suivant présente une méthode d'alignement semi-automatisé de la CIM-10 de l'OMS (ici, la version française, mais cette méthodologie est applicable à toutes les langues de la CIM-10) avec l'ICD-10 CM. Cette méthode s'applique à la fois à l'alignement de la CIM-10 vers l' ICD-10 et de l'ICD-10 vers la CIM-10.

Ma contribution à ce travail a été de définir la méthodologie d'alignement, de concevoir l'algorithme d'alignement et d'évaluer cette méthode.

*Digital Personalized Health and Medicine*
*L.B. Pape-Haugaard et al. (Eds.)*

*doi:10.3233/SHTI200114*

# A Semi-Automated Approach for Multilingual Terminology Matching: Mapping the French Version of the ICD-10 to the ICD-10 CM

Emmanuelle SYLVESTRE[abc1], Guillaume BOUZILLÉ[ab], Michael McDUFFIE[d], Emmanuel CHAZARD[e] Paul AVILLACH[d], Marc CUGGIA[ab]
[a]*INSERM, LTSI UMR 1099, F-35000, Rennes France*
[b]*CHU Rennes, Centre de Données Cliniques, F-35000, Rennes France*
[c]*CHU Martinique, Centre de Données Cliniques, F-97200, Martinique France*
[d]*Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115 USA.*
[e]*Université de Lille, CHU Lille, CERIM EA2694, F-59000 Lille, France.*

**Abstract.** The aim of this study was to develop a simple method to map the French International Statistical Classification of Diseases and Related Health Problems, 10th revision (ICD-10) with the International Classification of Diseases, 10th Revision, Clinical Modification (ICD-10 CM). We sought to map these terminologies forward (ICD-10 to ICD-10 CM) and backward (ICD-10 CM to ICD-10) and to assess the accuracy of these two mappings. We used several terminology resources such as the Unified Medical Language System (UMLS) Metathesaurus, Bioportal, the latest version available of the French ICD-10 and several official mapping files between different versions of the ICD-10. We first retrieved existing partial mapping between the ICD-10 and the ICD-10 CM. Then, we automatically matched the ICD-10 with the ICD-10-CM, using our different reference mapping files. Finally, we used manual review and natural language processing (NLP) to match labels between the two terminologies. We assessed the accuracy of both methods with a manual review of a random dataset from the results files. The overall matching was between 94.2 and 100%. The backward mapping was better than the forward one, especially regarding exact matches. In both cases, the NLP step was highly accurate. When there are no available experts from the ontology or NLP fields for multi-lingual ontology matching, this simple approach enables secondary reuse of Electronic Health Records (EHR) and billing data for research purposes in an international context.

**Keywords.** ICD-10, Clinical terminologies, Interoperability, Multilingual matching

## 1. Introduction

The International Statistical Classification of Diseases and Related Health Problems, 10th revision (WHO-ICD-10) one of the most popular terminologies used in around the world.

1 Corresponding Author, *Faculté de médecine, Université Rennes 1, 2 Avenue du Professeur Léon Bernard 35043 Rennes Cedex 9, France*; E-mail: emmasyl@gmail.com.

It is a standard diagnostic terminology created and maintained by the World Health Organization (WHO) since 1990 for diagnostic coding[1]. The French healthcare system uses a French version of the WHO-ICD-10 (Classification Internationale des Maladies, 10e version, CIM-10) since 1997[2], while the United States created and implanted their own adaptation of the terminology (International Classification of Diseases, 10th Revision, Clinical Modification, ICD-10-CM) on October, 1st, 2015[3]. There are far more codes in the ICD-10 CM than in the CIM-10, even though they share the same common denominator: the WHO-ICD-10[3].

Multilingual ontology matching is the process of finding correspondences between ontologies of different languages to allow them to interoperate[4]. This enables secondary reuse of Electronic Health Records (EHR) and billing data from different healthcare systems for research purposes, especially if data from the United States is involved in the study. We can use two main strategies for multilingual ontology matching: direct and indirect alignment. The direct alignment is translation-based and uses external resources to help with translation, while the indirect alignment uses intermediary mappings between the source and target ontologies. Furthermore, mapping two ontologies can be an automated or manual process. Manual mapping is still the prevalent choice for ontology matching, but necessitates a large team of experts, is time-consuming, and is prone to errors[5]. On the other hand, automated approaches use public terminology resources such as the Unified Medical Language System (UMLS)[4] or Bioportal[6] but those sources are extremely incomplete outside of the English speaking world[7]. Therefore, when the purpose of a study is not the mapping itself but a necessary step to join databases, it should be possible to overcome the semantic interoperability issue by combining different automated matching techniques to conduct the study, even with limited resources or experts from the ontology matching field.

The aim of this study was to develop and evaluate a simple method to link the French ICD-10 (or any version of the WHO ICD-10) with the ICD-10 CM. We sought to map these terminologies forward (ICD-10 to ICD-10 CM) and backward (ICD-10 CM to ICD-10).

## 2. Methods

### *2.1. Terminology resources*

Since there were no direct mapping files for our study, we used all the intermediate mapping files available online. We used four data sources: i) the UMLS Metathesaurus, which integrates and assigns a unique identifier to synonymous concepts from several standard biomedical technologies (including ICD-10 CM, WHO-ICD-10 and a 1998 version of the CIM-10)[4], ii) Bioportal, which is a comprehensive repository of standard terminologies created by the National Center for Biomedical Ontology (NCBO), with an ontology alignment tool, called Lexical OWL Ontology Matcher (LOOM)[6], iii) the latest version of the CIM-10, which is mainly a translation of the WHO-ICD-10 with more children, and is publicly available on the national billing agency website (Agence Technique de l'Information sur l'Hospitalisation, ATIH)[2] and iv) different existing mapping files with forward and backward mapping such as: the General Equivalent Mapping (GEM) Files from the National Center for Health Statistics (NCHS)[8], the New-Zealand mapping files from the Ministry of Health (ICD-10 Australian Modification and ICD-9 Australian Modification)[9] and the mapping files between the

ICD-10 AM and ICD-10, from the Australian Consortium for Classification Development (ACCD)[10].

### *2.2. Mapping method*

We decided to map only the first 4-character codes of the CIM-10, because the WHO-ICD-10 is mainly 4-character codes, except for chapters XIII, XIX, XX, which means that after this position, there was a bigger risk of mapping codes with a different signification between the two languages.
Our strategy used three main steps (Figure 1). First, we used the NCBO API to retrieve the mapping between the ICD-10 CM and the WHO-ICD-10 and the partial mapping between the CIM-10 and the WHO-ICD-10 from the UMLS Metathesaurus and LOOM algorithm, which is available through the NCBO API. Then, we matched each terminology using the different mapping files mentioned above. And finally, we used manual review and natural language processing (NLP) to recognize labels between the two terminologies with custom R scripts. The NLP process was only used on unmatched codes after the two first steps. For each set, one native-speaking investigator from each language extracted the two or three main words of the ICD or CIM label, then they built together a translation dictionary and used rules-based NLP to match the labels and codes. For the remaining unmatched codes, we mapped them to their three-character codes parent. After all these steps, we considered the code unmatched if we could not find an exact or approximate match.

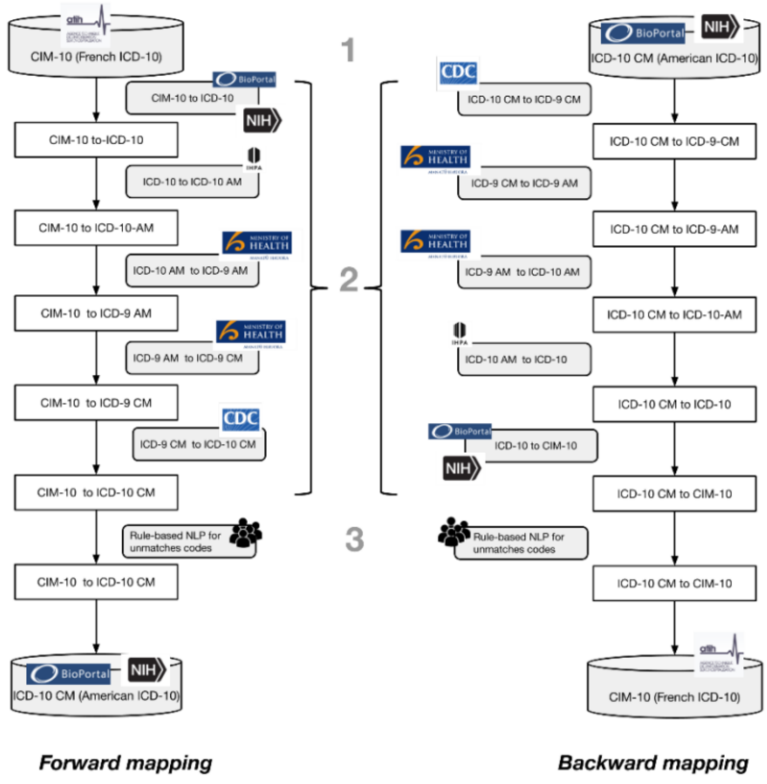



**Figure 1.** Forward and backward mapping methods

Since the automated mapping was based on official mapping files, we did not review those matched codes. However, we reviewed manually all NLP-based matches before confirming their match or un-match status. Uncertain pairs were reviewed again by two other medical investigators.

## 3. Results

The ICD-10 CM included 91,737 codes and the French ICD-10 included 39,928 codes, including the 3-charcters codes parents. The ICD-10 CM had 9835 (11%) of 4-characters codes while the ICD-10 had 12,345 (31%) of 4-character codes.
Among the 39,928 codes of the CIM-10, 8,477 (21.2%) were exact matches to the ICD-10 CM and 29,131 (73 %) were partial matches. 264 of those partial matches were based on the NLP-based step. There were 2,320 (5.8%) missing codes.
Among the 91,737 codes of the ICD-10 CM, 9,082 (9.9%) were exact matches and 82,655 (90.1 %) partial matches with no unmatched codes. 226 of those partial matches were based on the NLP-based step. There were no missing codes.
All the NLP-based matches were true positives with no false negatives. Overall, the backward mapping was 94.2%, while the forward mapping was 100%. (Table 1)

**Table 1.** Characteristics of all matches after the forward and backward mappings

| Mapping | Exact Match | Partial Match* | | | | No match |
|---|---|---|---|---|---|---|
| | | >4-characters code to one 4-character code | >4-characters code to more than one 4-characters code | 4-characters code to more than one 4-characters code | 4-characters codes (or more) to the 3-characters parent | |
| ICD-10 to ICD-10 CM | 8,477 (21.2%) | 4,832 (12.1%) | 23,345 (58.5%) | 932 (2.3%) | 22 (0.1%) | 2,320 (5.8%) |
| ICD-10CM to ICD-10 | 9,082 (9.9%) | 41,518 (45.3%) | 33,893 (36.9%) | 1,797 (2%) | 5,447 (5.9%) | 0 (0%) |

*The partial match includes the codes matched using the Natural Language Processing (NLP) step

## 4. Discussion

Our method showed a very high matching score, especially regarding the backward mapping (ICD-10 CM to ICD-10) with 100% of codes matched. However, the exact match was far better in the forward mapping (21% versus 10% in the backward). The manual evaluation confirmed the accuracy of the rules-based NLP algorithm.

Our study has several limitations. First, we only tested the method with two languages (French and English). However, we only had native speaking experts in French and English available for the manual review and most countries use either the ICD-10 CM, or a fairly close version of the WHO ICD-10 for reimbursement and billing[1]. Therefore, we knew that if our mapping process was accurate it would be relatively easy to adapt it to other languages. Second, our NLP was rather basic because we wanted first and foremost to use already existing reference files. Previous studies[11,12] showed that NLP based on modern machine-learning methods is the most pertinent method to match a diagnostic to an ICD-10 code or to translate a terminology to another language, but it is a very specific field with few experts, especially outside of the English language. The idea here was to propose an alternative when NLP specialists are not available to implement automatic translation and/or matching algorithms. Finally, most of our matches are partial. A majority of exact matches would have been ideal,

especially since the ICD-10 is not always very precise regarding some diagnoses[13], but since the WHO-ICD-10 and the ICD-10-CM have a vastly different number of codes[3], this outcome was predictable.

Manual mapping for multilingual ontology alignment is still the gold standard today, but it requires several experts of the healthcare and ontology fields and is a very time-consuming work that can take years[14]. This method cannot make the same claims of precision as official manual mapping files, but it could become a fairly quick and reliable process for international studies based on secondary reuse of EHR and billing data (including legacy data coded in ICD-9 CM, thanks to the GEM files from the NCHS[8]) without any ontology or translation experts.

## 5. Conclusion

Our study demonstrated that semi-automated mapping based on reference mapping files (in standard format) and basic NLP could be considered for secondary reuse of EHR and billing data from different countries when there are no existing reference files. Next, we would like to replicate this study with ICD-10 in other languages and use more automated NLP resources, like the Google Translate API, to confirm the accuracy of this method. This work could also serve as a basis for semi-automated mapping of the ICD-10 to the ICD-11, once the official ICD-10 to ICD-11 mapping files are available.

## References

[1] WHO | International Classification of Diseases (ICD) Information Sheet, *WHO*. (n.d.). https://www.who.int/classifications/icd/factsheet/en/.
[2] ATIH : Agence technique de l'information sur l'hospitalisation, (n.d.). http://www.atih.sante.fr/.
[3] L. Manchikanti, A.D. Kaye, V. Singh, and M.V. Boswell, The Tragedy of the Implementation of ICD-10-CM as ICD-10: Is the Cart Before the Horse or Is There a Tragic Paradox of Misinformation and Ignorance?, *Pain Physician*. **18** (2015) E485-495.
[4] US National Library of Medicine. Unified Medical Language System (UMLS): Metathesaurus, (n.d.). https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/index.html.
[5] C.T. Dos Santos, P. Quaresma, and R. Vieira, An API for multilingual ontology matching, in: 2010.
[6] P.L. Whetzel, N.F. Noy, N.H. Shah, P.R. Alexander, C. Nyulas, T. Tudorache, and M.A. Musen, BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications, *Nucleic Acids Res.* **39** (2011) W541-545.
[7] A. Doan, J. Madhavan, P. Domingos, and A. Halevy, Ontology matching: A machine learning approach, in: Handb. Ontol., Springer, 2004: pp. 385–403.
[8] ICD - ICD-10-CM - International Classification of Diseases, Tenth Revision, Clinical Modification, (2019). https://www.cdc.gov/nchs/icd/icd10cm.htm.
[9] Mapping between ICD-10 and ICD-9 | Ministry of Health NZ, (n.d.). http://www.health.govt.nz/nz-health-statistics/data-references/mapping-tools/mapping-between-icd-10-and-icd-9.
[10] Australian Consortium for Classification Development, The international statistical classification of diseases and related health problems, tenth revision, Australian modification (ICD-10-AM/ACHI/ACS): Mapping files, (n.d.). https://ace.ihpa.gov.au/Downloads.aspx.
[11] N.D. Hailu, K.B. Cohen, and L.E. Hunter, Ontology translation: A case study on translating the Gene Ontology from English to German, *Nat. Lang. Process. Inf. Syst. Int. Conf. Appl. Nat. Lang. Inf. Syst. NLDB Revis. Pap. Int. Conf. Appl. Nat. Lang. Info*. **8455** (2014) 33–38.
[12] A. Atutxa, A.D. de Ilarraza, K. Gojenola, M. Oronoz, and O. Perez-de-Viñaspre, Interpretable deep learning to map diagnostic texts to ICD-10 codes, *Int. J. Med. Inf.* **129** (2019) 49–59.
[13] Y.M. Mesfin, A.C. Cheng, A.H.L. Tran, and J. Buttery, Positive predictive value of ICD-10 codes to detect anaphylaxis due to vaccination: A validation study, *Pharmacoepidemiol. Drug Saf.* (2019).
[14] ICD-11 Content - Content Development Roadmap - SNOMED Confluence, (n.d.). https://confluence.ihtsdotools.org/display/CDR/ICD-11+Content.

# Troisième partie
# Verrous liés à l'exploitation des données

# Chapitre 1 : Méthodes pour la prédiction et la surveillance de la dengue

L'intégration des données de vie réelle n'est que la première étape dans le traitement de ces données. Grâce à la richesse apportée par l'utilisation de plusieurs sources, ce processus, s'il est correctement effectué, permet par la suite d'exploiter les données intégrées de façon transversale, en complémentarité des données traditionnelles collectées dans le cadre des essais thérapeutiques ou des études observationnelles.

Cependant, cette hétérogénéité des données implique également une variabilité dans les méthodes d'exploitation, chacune devant s'adapter aux types de données traitées.

Ici, le cas d'usage choisi était l'utilisation de données de vie réelle pour la surveillance (et potentiellement la prédiction) de la dengue en Martinique. Afin de sélectionner les sources de données et les éventuels algorithmes d'apprentissage machine pertinents, il fallait en amont identifier la littérature déjà publiée sur le sujet. En effet, nous souhaitions respecter les recommandations de l'OMS dans la prise en charge globale de la dengue, et donc exploiter à la fois les données de surveillance humaine et les données de surveillance vectorielle. Par ailleurs, nous souhaitions également évaluer l'apport potentiel des données de *Google*, dont la pertinence a déjà été évaluée pour la surveillance saisonnière de la grippe. La dengue en Martinique n'a pas une saisonnalité annuelle comme la grippe, mais la Martinique (et les Antilles en général) est une zone endémo-épidémique, c'est-à-dire que le virus circule toute l'année sur le territoire et cette circulation peut se transformer en épidémie d'une année à l'autre. Nous voulions donc évaluer l'apport des données du web pour une maladie au cycle de transmission extrêmement différent de celui de la grippe.

Pour cela, nous avons effectué une revue systématique afin d'identifier toutes les sources et les méthodologies basées sur les données de vie réelle qui ont déjà été utilisées et évaluées pour la surveillance et la prédiction de la dengue dans le monde. Nous souhaitions en particulier identifier le rôle potentiel des données vectorielles et des données issues de *Google*.

Cette revue incluant les publications sur 20 ans (de janvier 2000 au 31 août 2020) nous a permis d'identifier plusieurs éléments clés : i) l'utilisation des données de vie réelle dans la gestion de la dengue a surtout été étudiée en Asie, qui représentait 72% des études incluses ; ii) les données vectorielles restent très peu utilisées dans ce genre d'études ; iii) les données issues de moteurs de recherche type *Google* ou de réseaux sociaux sont intéressantes, mais restent peu utilisées

en concomitance avec les données cliniques ; iv) les modèles les plus largement utilisés sont les modèles de classification supervisée.

**Synthèse**

Les données de vie réelle étant hétérogènes, les méthodologies d'exploitation de ces données vont varier en fonction des cas d'usage.

Pour identifier les approches méthodologiques les plus pertinentes pour le cas d'usage qui est la surveillance syndromique de la dengue en Martinique, nous avons réalisé une revue systématique de la littérature.

Ce travail a fait l'objet d'une publication qui est présentée ci-dessous.

## Article 3 : Data-driven methods for dengue prediction and surveillance using real-world and Big Data: A systematic review

Ma contribution à cet article a été de définir la stratégie de recherche et rédiger le protocole de revue systématique en accord avec la méthodologie PRISMA, de l'enregistrer sur la base PROSPERO et d'identifier les articles éligibles.

Deux auteurs indépendants dont je faisais partie ont sélectionné les articles. En cas de désaccord sur la sélection des articles, un troisième auteur permettait de départager les deux auteurs. Une fois cette sélection effectuée, j'ai extrait les données pertinentes des articles et analysé l'ensemble des résultats.

PLOS NEGLECTED TROPICAL DISEASES

RESEARCH ARTICLE

# Data-driven methods for dengue prediction and surveillance using real-world and Big Data: A systematic review

**Emmanuelle Sylvestre[1,2]*, Clarisse Joachim[3,4], Elsa Cécilia-Joseph[2], Guillaume Bouzillé[1], Boris Campillo-Gimenez[1,5], Marc Cuggia[1], André Cabié[6,7,8]**

**1** Université de Rennes, CHU Rennes, INSERM, LTSI – UMR 1099, Rennes, France, **2** CHU Martinique, Centre de Données Cliniques, Martinique, France, **3** CHU Martinique, Pôle de Cancérologie Hématologie Urologie, Registre Général des Cancers de la Martinique, Martinique, France, **4** CHU Martinique, Pôle de Cancérologie Hématologie Urologie, Martinique Cancer Data Hub, Martinique, France, **5** Centre de Lutte Contre le Cancer Eugène Marquis, Rennes, France, **6** CHU Martinique, Infectious and Tropical Diseases Unit, Martinique, France, **7** CHU Martinique, INSERM, CIC-1424, Martinique, France, **8** PCCEI, Université de Montpellier, INSERM, EFS, Université Antilles, Montpellier, France

* emmanuelle.sylvestre@chu-martinique.fr





**Citation:** Sylvestre E, Joachim C, Cécilia-Joseph E, Bouzillé G, Campillo-Gimenez B, Cuggia M, et al. (2022) Data-driven methods for dengue prediction and surveillance using real-world and Big Data: A systematic review. PLoS Negl Trop Dis 16(1): e0010056. https://doi.org/10.1371/journal.pntd.0010056

**Editor:** Victor S. Santos, Universidade Federal de Alagoas - Campus Arapiraca, BRAZIL

**Received:** April 30, 2021

**Accepted:** December 6, 2021

**Published:** January 7, 2022



**Data Availability Statement:** All relevant data are within the manuscript and its Supporting information files.

**Funding:** The author(s) received no specific funding for this work.

**Competing interests:** The authors have declared that no competing interests exist.

## Abstract

### Background

Traditionally, dengue surveillance is based on case reporting to a central health agency. However, the delay between a case and its notification can limit the system responsiveness. Machine learning methods have been developed to reduce the reporting delays and to predict outbreaks, based on non-traditional and non-clinical data sources. The aim of this systematic review was to identify studies that used real-world data, *Big Data* and/or machine learning methods to monitor and predict dengue-related outcomes.

### Methodology/Principal findings

We performed a search in PubMed, Scopus, Web of Science and grey literature between January 1, 2000 and August 31, 2020. The review (ID: CRD42020172472) focused on data-driven studies. Reviews, randomized control trials and descriptive studies were not included. Among the 119 studies included, 67% were published between 2016 and 2020, and 39% used at least one novel data stream. The aim of the included studies was to predict a dengue-related outcome (55%), assess the validity of data sources for dengue surveillance (23%), or both (22%). Most studies (60%) used a machine learning approach. Studies on dengue prediction compared different prediction models, or identified significant predictors among several covariates in a model. The most significant predictors were rainfall (43%), temperature (41%), and humidity (25%). The two models with the highest performances were Neural Networks and Decision Trees (52%), followed by Support Vector Machine (17%). We cannot rule out a selection bias in our study because of our two main limitations: we did not include preprints and could not obtain the opinion of other international experts.

## Conclusions/Significance

Combining real-world data and *Big Data* with machine learning methods is a promising approach to improve dengue prediction and monitoring. Future studies should focus on how to better integrate all available data sources and methods to improve the response and dengue management by stakeholders.

## Author summary

Dengue is one of the most important arbovirus infections in the world and its public health, societal and economic burden is increasing. Although the majority of dengue cases are asymptomatic or mild, severe disease forms can lead to death. For this reason, early diagnosis and monitoring of dengue are crucial to decrease mortality. However, most endemic regions still rely on traditional monitoring methods, despite the growing availability of novel data sources and data-driven methods based on real-world data, *Big Data*, and machine learning algorithms. In this systematic review, we identified and analyzed studies that used these novel approaches for dengue monitoring and/or prediction. We found that novel data streams, such as Internet search engines and social media platforms, and machine learning methods can be successfully used to improve dengue management, but are still vastly ignored in real life. These approaches should be combined with traditional methods to help stakeholders better prepare for each outbreak and improve early responsiveness.

## Introduction

Dengue virus (DENV) is an arbovirus transmitted to humans by *Aedes aegypti* or *Aedes albopictus* female mosquitoes [1]. The incidence of dengue, the disease caused by DENV, has rapidly increased around the world in recent decades [2] due to population growth, urbanization, increased travel, and insufficient vector control [3]. The World Health Organization (WHO), considers dengue a major global public health challenge in the tropical and subtropical regions [4]. Today, dengue is one of the most important vector-borne diseases in the world and recent studies on its prevalence estimate that 3.9 billion people are at risk of transmission, with 390 million infections and 96 million symptomatic cases per year [1,5]. Although most infections are asymptomatic or are characterized by intense flu-like symptoms that last up to 10 days [6], severe forms of dengue hemorrhagic fever/dengue shock syndrome can also occur [7] and might lead to death. Mortality due to dengue can be greatly reduced by early diagnosis, appropriate clinical management [3,7].

Most dengue-endemic regions (mainly South-East Asia, the Americas, and the Pacific region) rely on traditional surveillance, based on hospital syndromic reporting and laboratory confirmation of a subset of cases to a central health agency [3,8]. The method is very accurate, but is hampered by its lack of responsiveness with substantial delays between a case and its notification [8], which can limit the health system ability/rapidity to put in place appropriate measures to avoid drastic consequences. Moreover, this traditional surveillance system is expensive, due to the time needed to aggregate and manually validate data [9]. These limitations have prompted researchers to investigate other solutions. Many studies have described alternative methods, such as mobile, digital and Internet-based systems, to efficiently crowdsource data from the community [3]. However, these approaches have not been translated yet

into the standard dengue management practice. Yet, they are relevant for all dimensions of dengue management, such as monitoring, clinical management, and dengue outbreak forecasting [3,8]. Over the years, scientists have developed statistical and machine learning models to reduce the reporting delays and monitor new cases in almost real-time, but also to accurately use non-traditional and non-clinical data sources (e.g. Internet search engines and social media platforms) to predict communicable disease outbreaks [10–13], including dengue. Many studies have proposed new strategies based on *Big Data* and machine learning models to improve dengue outbreak management. However, recent systematic reviews only examined the relevance and usefulness of Internet-based surveillance systems in emerging tropical disease management [8,14], and they did not focus specifically on dengue management. Furthermore, recent systematic reviews on dengue analyzed monitoring [15], vaccine efficacy [16], epidemiological trends [17,18], the overall disease burden [19–21] and clinical prognosis models [22], but they did not discuss these new methods to improve dengue management.

Therefore, the first aim of this systematic review was to identify and describe all real-world and *Big Data*-based methods used to monitor and predict/forecast dengue-related outcomes, regardless of the region and/or population. The second aim was to analyze several features of these studies, such as the data sources and their origin, the different outcome types (e.g. epidemiological and clinical outcomes), the chosen statistical methods, and their performance and variability based on the population and location.

## Methods

This systematic review was performed following the "Preferred Reporting Items for Systematic Reviews and Meta-Analyses" (PRISMA) guidelines [23]. Four reviewers (ES, CJ, AC and MC) developed the systematic review protocol. The literature search was performed in September 2020. The study protocol was registered on the PROSPERO registry of systematic reviews (ID: CRD42020172472).

### Eligibility criteria

The review focused on studies that used real-world data, *Big Data* and/or machine learning methods to monitor, predict and/or forecast dengue outbreaks or dengue-related outcomes (clinical or epidemiological). Studies from any country (also regions outside endemic regions) were included, without any language filter. Analyses could be performed on past or future data.

### Inclusion criteria

- Dengue diagnosis based on the standard WHO definition [7] valid at the time of the study
- Studies on humans, regardless of age, sex, and disease severity
- Studies using real-world data (including *Big Data)* (i.e. data not collected in experimental conditions) [24] for surveillance and/or prediction of dengue outbreaks.

### Exclusion criteria

- Studies without original data, such as reviews, editorials, guidelines and perspectives articles
- Randomized control trials, case series, and case reports
- Descriptive epidemiological studies without any modeling
- Studies on other arbovirus types (e.g. chikungunya, Zika virus disease)

- Studies exclusively on mosquitoes (without any human data) and *in vitro* studies
- Studies only on incidence using geographic information systems

## Search methodology

**Information sources and search strategy.** The literature search was carried out in MEDLINE (PubMed), Scopus and Web of Science between January 1, 2000 and August 31, 2020 to identify potentially eligible studies. MeSH terms and keywords were used to perform the queries. First, the MeSH term "Dengue" was combined with several other MeSH terms (e.g. Data mining, Big Data, Forecasting, Social media), using the Boolean operator AND. Then, a more specific combination of keywords was used for all databases: i) Dengue AND [Monitoring OR Surveillance] AND [Big Data OR Data mining OR Instagram OR Facebook OR Twitter OR Tweets OR Google OR Baidu OR Google Trends OR Social media OR Social network OR Internet], ii) Dengue AND [Prediction OR Forecasting OR Modeling OR Modelling] AND [Big Data OR Data mining OR Instagram OR Facebook OR Twitter OR Tweets OR Google OR Baidu OR Google Trends OR Social media OR Social network OR Internet], iii) Dengue AND [Big Data OR Data mining OR Instagram OR Facebook OR Twitter OR Tweets OR Google OR Baidu OR Google Trends OR Social media OR social network OR Internet].

Relevant articles were also searched in the grey literature, including French-language studies on HAL (Hyper Articles en Lignes) [25], which is an open archive where authors can deposit scholarly documents from all academic fields, *theses.fr* [26], which is the French open database for all ongoing and defended PhD theses in France, and the WHO Dengue Bulletin.

Finally, the references of the retained studies and of major dengue epidemiological review articles were screened to identify studies overlooked by the previous search strategies.

**Selection process.** Two independent authors (ES and CJ) screened the title and abstract to select relevant studies for the review. They read the full text of all studies that seemed to meet the eligibility criteria, or if the abstract was not explicit enough to make a decision. In case of disagreement, a third reviewer helped to reach a consensus (AC).

**Quality assessment, data collection, extraction, and analysis.** Two reviewers (ES and CJ) extracted data from the selected articles, including first and last authors, year of publication, study period, objectives, study population, methodology, model performance and evaluation, study site (S1 Text).

As reporting guidelines for machine learning models and real-world data studies are not available, each reviewer independently performed a quality assessment using quality assessment criteria described in previous review articles on these topics [27–29] (S1 Table). A narrative synthesis of all eligible studies was prepared using the following framework: i) data sources and outcomes, ii) statistical and machine learning methods, iii) evaluation metrics, and iv) study results.

All descriptive analyses from the extracted articles were performed using R version 3.6.3 [30].

## Results

Among the 2064 studies identified, 119 articles were included in this systematic review (Fig 1) [31–148]. Although the search time window was from January 1, 2000, the first included studies were published in 2008, and 67% of the eligible articles were published between 2016 and 2020 (Fig 2). The study populations were predominantly from South-East Asia (37%) and South America (22%). Among the 119 papers included, 77 (65%) were articles, and 42 (35%) were conference papers. On the basis of the Web of Science "Research Area" and the Scopus "Subject Area" classification, the topic of the selected articles were aggregated into eight

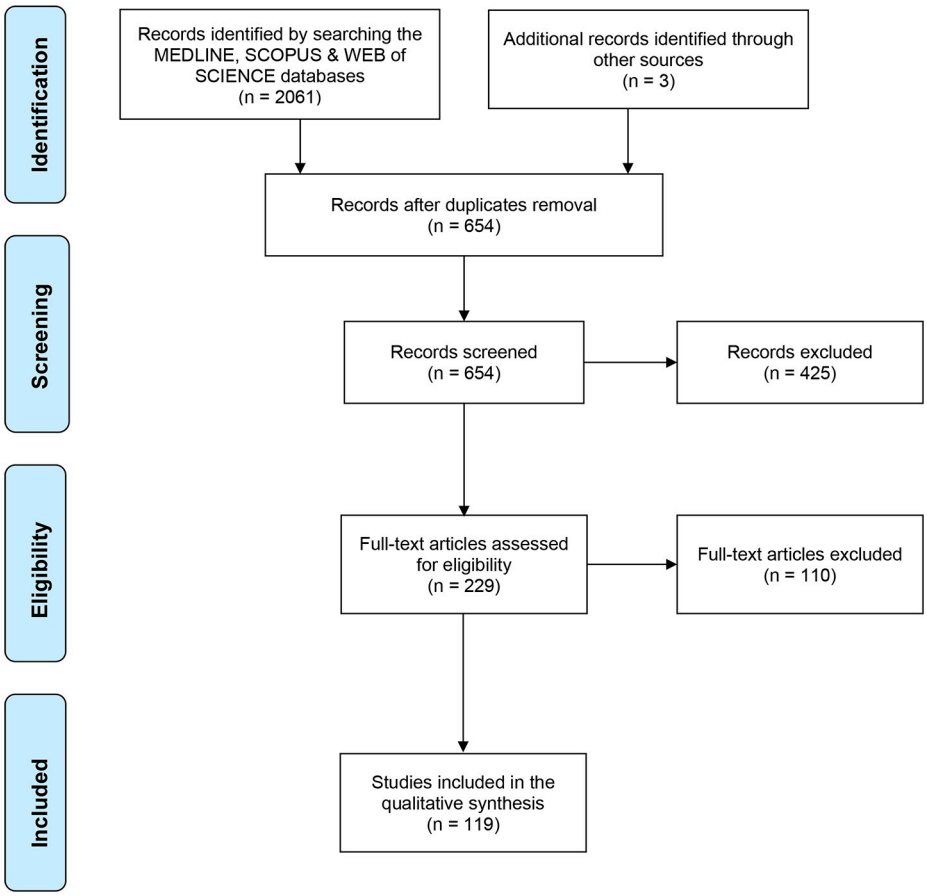



**Fig 1. PRISMA Flow Diagram describing the screening process for the systematic review.**

https://doi.org/10.1371/journal.pntd.0010056.g001

categories and three main themes: i) Information Technology & Science (52% of all articles), ii) Medicine (24%), and iii) Health Informatics, Public Health & Biology (24%) (Table 1). Conference papers were mainly classified in the "Information Technology & Science" category (39/42; 93%), whereas articles were more evenly distributed in the "Medicine" (28/77; 36%), "Health Informatics, Public Health & Biology" (26/77; 34%) and "Information Technology & Science" (23/77; 30%) themes (S2 Table). The complete list of all selected studies and their characteristics are in S3 Table.

## Data sources

All included studies, except one [68], used only retrospective data. Most articles had multiple and heterogeneous data sources. The most conventional data sources were: government agencies (n = 72, 46%) and medical institutions (e.g. hospitals/laboratories) (n = 30, 19%). The data retrieved from these sources included epidemiological data, climate and environmental data from meteorological departments, and clinical and biological data. Some studies also used open access data from the WHO or from databases of published studies (S3 Table).

Among the included studies, 47/119 (39%) used at least one novel data stream, such as Internet search engines and social networks [14]. Most of these studies (n = 41, 87%) were published after 2015. Google was the most frequently used Internet search engine (n = 19

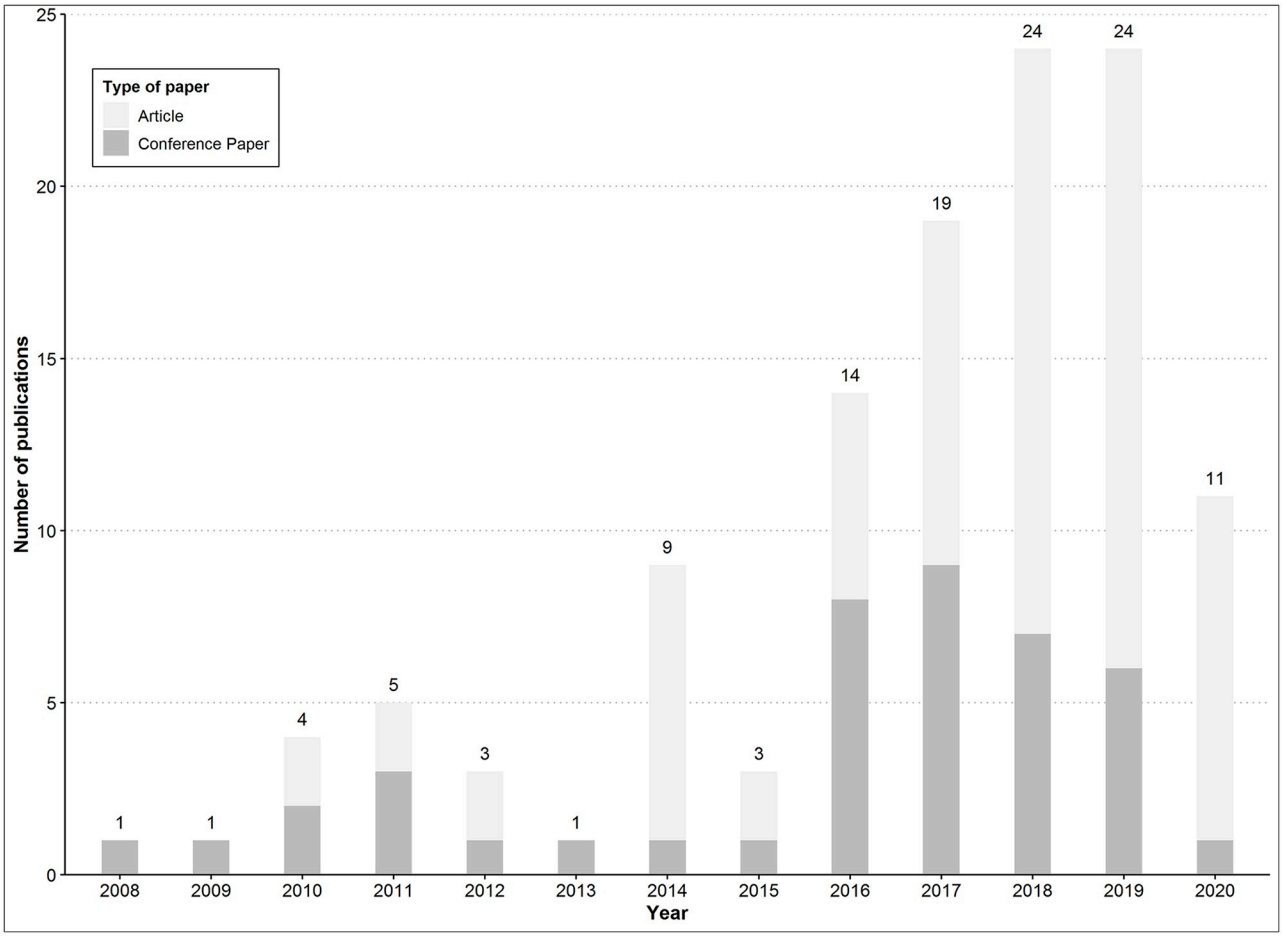



**Fig 2. Number of publications on dengue prediction and/or surveillance published between January 1, 2000 and August 31, 2020.**

https://doi.org/10.1371/journal.pntd.0010056.g002

studies) and Twitter the most frequently used social network (n = 18). Many studies based on novel data streams were research articles (n = 33, 70%), but the main theme, regardless of the study type (Conference paper or Article) varied depending on the data. Specifically, studies based on Google data were classified homogeneously into the three main themes. Conversely, studies that exploited social networks as data source were evenly distributed between Conference papers (n = 9) and Articles (n = 10), but only few of them were classified into the Medicine theme (Table 2).

Most studies used structured data, but 41 (34%) studies had an unstructured data source, such as Internet search-based queries or Twitter (Table 2). Among the 41 studies that used unstructured data, 28 (68%) did not develop their own pre-processing methods for these data sources, but simply used keywords related to their research. However, when studies used Natural-Language Processing (NLP)-based methods, they had a full pre-processing framework based on the NLP state-of-the-art recommendations.

Overall, studies that used non-conventional data relied less frequently on clinical data. Conversely, studies that used human data relied mostly on traditional sources, such as weather and environmental data. Moreover, genomic and vector data were vastly underused in combination with other sources, because only five studies using at least one of these sources were included in this systematic review. Data sources are detailed in Table 2.

**Table 1. Type, study population and themes of the selected studies.**

| | n | % |
|---|---|---|
| **Study type** | **119** | |
| Article | 77 | 65 |
| Conference paper | 42 | 35 |
| **Geographic region*** | | |
| Americas | | |
| Caribbean | 3 | 2 |
| North America | 3 | 2 |
| South America | 28 | 22 |
| Asia | | |
| East Asia | 16 | 13 |
| South-East Asia | 47 | 37 |
| South Asia | 27 | 21 |
| Australia | 1 | 1 |
| Worldwide | 2 | 2 |
| **Study main theme** | | |
| Information Technology & Science | **62** | **52** |
| Computer Science | 42 | 35 |
| Engineering | 10 | 8 |
| Science & Technology—Other Topics | 10 | 8 |
| Medicine | **28** | **24** |
| Infectious Diseases & Tropical Medicine | 20 | 17 |
| Medicine—Other Topics | 8 | 7 |
| Health Informatics, Public Health & Biology | **29** | **24** |
| Biology | 7 | 6 |
| Medical Informatics | 16 | 13 |
| Public Health | 6 | 5 |

*Some studies were carried out in more than one geographic regions

https://doi.org/10.1371/journal.pntd.0010056.t001

## Statistical methods

The main aim of the included studies was to predict a dengue-related outcome (n = 65, 55%), to assess the validity of data sources for dengue surveillance (n = 29, 24%), or both (n = 25, 21%). The most frequently chosen outcomes (for prediction and monitoring) were dengue incidence rate (n = 58, 49%), dengue diagnosis based on symptoms (n = 20, 17%), and dengue outbreaks (n = 18, 15%) (S4 Table).

Only one study [48] used NLP-based methods for dengue prediction or surveillance, but as a pre-treatment step to extract and format data for modelling.

The model choice was related to the study objectives (prediction/forecasting or validity of a data source for dengue monitoring). Overall, most studies compared the performances of different models and statistical methods. The most frequently used models, regardless of the study aim(s), were regression-based models (25%), followed by decision-tree models (18%), and artificial neural networks (15%). Most studies on dengue monitoring used correlation analyses to identify relevant variables and/or data sources. Correlation methods (Pearson correlation or Spearman correlation) were especially useful to assess the validity of novel data streams, such as Twitter and Internet search engines. Most studies that included machine-learning algorithms used supervised learning methods (69%). The models' characteristics are detailed in Table 3.

**Table 2. Data sources for dengue monitoring and prediction depending on the main theme.**

| Number of studies n(%)[a] | | Study main theme n (%) | |
|---|---|---|---|
| | IT | Med | PH |
| **119** | **62** | **28** | **29** |
| **Traditional data sources** | | | |
| Epidemiological and demographic data | 86 (72) | 42 (68) | 24 (86) | 20 (69) |
| Clinical and biological data | 33 (27) | 20 (32) | 3 (11) | 10 (34) |
| Genomic sequence data | 2 (1) | 1 (2) | 0 (0) | 1 (3) |
| Climate, environmental and geographic data | 45 (37) | 26 (42) | 12 (43) | 7 (24) |
| Vector data | 4 (3) | 1 (2) | 3 (11) | 0 (0) |
| **Novel data streams** | | | | |
| **Internet search engine data** | 25 (21) | 8 (13) | 11 (39) | 6 (21) |
| Baidu | 6 (5) | 2 (3) | 4 (14) | 0 (0) |
| Google | 19 (15) | 6 (9) | 7 (25) | 6 (20) |
| **Social media data** | 21 (17) | 14 (22) | 4 (14) | 3 (10) |
| Twitter | 18 (14) | 12 (19) | 4 (14) | 2 (6) |
| Other | 3 (2) | 2 (3) | 0 (0) | 1 (3) |
| **Other data sources** | 10 | 2 (3) | 3 (11) | 5 (17) |
| Cellphone | 2 | 2 (3) | 0 (0) | 0 (0) |
| HealthMap | 2 | 0 (0) | 1 (3) | 1 (3) |
| LeXisNexis | 2 | 0 (0) | 1 (3) | 1 (3) |
| Political stability | 1 | 0 (0) | 0 | 1 (3) |
| Wikipedia | 1 | 0 (0) | 1 (3) | 0 (0) |

[a] As most studies used several data sources, some articles are present several times.

IT: Information Technology & Science; Med: Medicine; PH: Health Informatics, Public Health & Biology

https://doi.org/10.1371/journal.pntd.0010056.t002

To evaluate and assess the performance of the chosen statistical methods and/or models, 71 studies (60%) used a machine learning approach and partitioned their data into a training set and a test set. Like for the models, the choice of evaluation metrics was closely related to the study aim(s). All articles used at least one metric, and most of them more than one. Overall, the most common metrics were based on a Confusion Matrix (53%), with Accuracy as the most used metric, followed by Recall or Sensitivity. Correlation-based metrics were used in 37% of studies, especially correlation coefficients (Pearson or Spearman, depending on the data source). The aim of most studies that used correlation metrics was to assess a data source for dengue monitoring (n = 37, 84% of the 44 studies with correlation metric). Error-based metrics were also commonly used (n = 35, 29% of all studies). Few studies used other metrics (n = 22, 18% of all studies) and only 9 studies (8%) did not used at least one metrics falling into the above categories. (Table 4).

## Study results

Among the 54 studies on surveillance, 37 (68%) assessed novel data streams, such as Internet search engines and social media, particularly Google (n = 16, 30%) and Twitter (n = 16, 30%). The most common traditional data source evaluated was climate, environmental and geographic data (n = 13/54; 24%) (S5 Table). All studies found a statistically significant association between the data source and the dengue-related outcome.

The aim of the studies on prediction (n = 90) could be categorized in two main groups: i) comparing different models to predict a dengue-related outcome, and ii) finding the

**Table 3. Statistical methods and models used in the selected studies depending on the study aim*.**

| Statistical methods | Prediction n (%) | Surveillance [a] n (%) | Prediction and surveillance n (%) | Totaln (%) |
|---|---|---|---|---|
| **Methods for statistical analysis** | **153** | **59** | **68** | **280** |
| Machine learning methods | 126 (82) | 27 (46) | 51 (75) | **204 (73)** |
| Supervised learning | 121 (79) | 21 (36) | 50 (74) | **192 (69)** |
| Unsupervised learning | 5 (3) | 6 (10) | 1 (1) | **12 (4)** |
| Other model types (including time series models) | 25 (16) | 9 (15) | 4 (6) | **38 (14)** |
| Correlation | 2 (1) | 23 (39) | 13 (19) | **38 (14)** |
| **Models for analyses** | **151** | **36** | **55** | **242** |
| Artificial neural networks | 31 (21) | 2 (6) | 3 (5) | **36 (15)** |
| Association rules | 3 (2) | 1 (3) | 0 (0) | **4 (2)** |
| Bayesian models | 12 (8) | 5 (14) | 3 (5) | **20 (8)** |
| Clustering | 5 (3) | 5 (14) | 1 (2) | **11 (5)** |
| Decision tree | 35 (23) | 2 (6) | 6 (11) | **43 (18)** |
| Regression model | 20 (13) | 9 (25) | 31 (56) | **60 (25)** |
| Support-vector machine | 17 (11) | 3 (8) | 7 (13) | **27 (11)** |
| Time series | 12 (8) | 1 (3) | 3 (5) | **16 (7)** |
| Other [b] | 16 (11) | 8 (22) | 1 (2) | **25 (10)** |

*As most studies used several models and/or statistical methods, some are listed several times.

[a] Studies evaluating a data source (traditional or novel data streams) for dengue monitoring

[b] Some models classified as "Other" are also included in the "Supervised learning" category

https://doi.org/10.1371/journal.pntd.0010056.t003

significant predictors among several covariates in a model. Twenty-two studies (24%) included tried to respond to both aims.

The most significant predictors were rainfall (22 models, 43% of 51 studies), temperature (21 models, 41% of 51 studies), and humidity (13 models, 25% of 51 studies). These predictors were also the most frequent in studies to predict dengue incidence rates or dengue outbreaks. Conversely, in studies on dengue diagnosis prediction, the most frequent predictors were fever (4 models, 66% of 6 studies), arthralgia/myalgia (3 models, 50% of 6 studies), platelet count (2 models, 33% of 6 studies), and white blood cell count (2 models, 33% of 6 studies) (Table 5).

Overall, in studies comparing different models, neural networks and decision trees gave the best performances and were the best models in 13 studies (52% of 54 studies), followed by support vector machine (9/54 studies, 17%). In studies to predict dengue incidence rates, regression-based models showed the highest performance (5/24 studies, 21%) (Table 6). The full list of models and predictors, depending on the outcome, is provided in S5 Table.

## Discussion

This systematic review showed that in the last 20 years, data-driven methods for dengue monitoring and prediction have become very popular, particularly in Asia where 72% of the included studies were performed. Very few studies were carried out outside Asia or the Americas, which is to be expected, because these are the two biggest dengue-endemic regions and 70% of the actual dengue burden is in Asia [149–151]. Studies in African countries were noticeably absent, although this continent also is a dengue-endemic region.

The most frequent data sources were conventional data traditionally used in dengue-related studies, such as case counts, climate, environmental, and clinical data. However, this review also highlighted the growing interest by the scientific community for novel Big Data streams for dengue surveillance and prediction [14,33,39–41,43,49,51–53,56,60,65,66,69–71,75–77,79–

**Table 4. Evaluation metrics used in the selected articles depending on their aim(s)*.**

| Evaluation metrics | Prediction n (%) | Surveillance [a] n(%) | Prediction and surveillance n(%) | Total n(%) |
|---|---|---|---|---|
| **Correlation metrics** | **8** | **22** | **18** | **48** |
| Correlation coefficient | 3 (38) | 16 (73) | 9 (50) | **28 (58)** |
| R-squared | 4 (50) | 5 (23) | 9 (50) | **18 (38)** |
| Other correlation metric | 1 (12) | 1 (5) | 0 (0) | **2 (4)** |
| **Error-based metrics** | **34** | **2** | **21** | **57** |
| Root mean square error | 14 (41) | 0 (0) | 9 (43) | **23 (40)** |
| Mean absolute error | 7 (21) | 0 (0) | 4 (19) | **11 (19)** |
| Mean absolute percentage error | 4 (12) | 0 (0) | 3 (14) | **7 (12)** |
| Mean squared error | 3 (9) | 0 (0) | 3 (14) | **6 (11)** |
| Other | 6 (18) | 2 (100) | 2 (10) | **10 (18)** |
| **Confusion matrix-based metrics** | **147** | **13** | **17** | **177** |
| Accuracy | 38 (26) | 6 (46) | 7 (41) | **51 (29)** |
| Recall/Sensitivity | 32 (22) | 2 (15) | 3 (18) | **37 (21)** |
| Specificity | 20 (14) | 0 (0) | 3 (18) | **23 (13)** |
| Precision/Positive predictive value | 17 (12) | 1 (8) | 1 (6) | **19 (11)** |
| F-score | 12 (8) | 2 (15) | 0 (0) | **14 (8)** |
| AUC and/or ROC curve[b] | 16 (11) | 1 (8) | 3 (18) | **20 (11)** |
| Kappa statistic | 5 (3) | 0 (0) | 0 (0) | **5 (3)** |
| Other | 7 (5) | 1 (8) | 0 (0) | **8 (5)** |
| **Other evaluation metrics** | **10** | **6** | **11** | **27** |
| **Number of articles using the evaluation metric** | **65** | **29** | **25** | **119** |
| Correlation metrics | 7 (11) | 19 (66) | 18 (72) | **44 (37)** |
| Error-based metrics | 19 (29) | 1 (3) | 15 (60) | **35 (29)** |
| Confusion matrix-based metrics | 47 (72) | 9 (31) | 8 (32) | **64 (54)** |
| Other evaluation metrics | 8 (12) | 6 (21) | 8 (32) | **22 (18)** |

*As studies used several metrics, some articles are listed more than once.

[a] Studies evaluating a data source (traditional data or novel data streams) for dengue monitoring

[b] AUC: Area Under the ROC Curve. ROC: Receiver Operating Characteristic

https://doi.org/10.1371/journal.pntd.0010056.t004

**Table 5. Most significant predictors for the three most frequently studied outcomes.**

| Number of studies n (%) | Dengue incidence rates n = 27 | Dengue outbreaks n = 9 | Dengue diagnosis n = 6 |
|---|---|---|---|
| **Significant predictors*** | | | |
| Rainfall | 14 (52) | 7 (78) | 0 (0) |
| Temperature | 14 (52) | 6 (67) | 0 (0) |
| Humidity | 9 (33) | 1 (11) | 0 (0) |
| Mosquito-related predictor | 0 (0) | 2 (22) | 0 (0) |
| Google search index | 4 (15) | 0 (0) | 0 (0) |
| Baidu search index | 3 (11) | 0 (0) | 0 (0) |
| Tweets | 3 (11) | 0 (0) | 0 (0) |
| Fever | 0 (0) | 0 (0) | 4 (66) |
| Arthralgia/myalgia | 0 (0) | 0 (0) | 3 (50) |
| Platelet count | 0 (0) | 0 (0) | 2 (33) |
| White blood cell count | 0 (0) | 0 (0) | 2 (33) |
| Other | 13 (48) | 6 (67) | 5 (83) |

*Most studies found several significant predictors

https://doi.org/10.1371/journal.pntd.0010056.t005

**Table 6. Model with the best performance for the three most frequently studied outcomes.**

| Number of studies | Dengue incidence rates n = 24 | Dengue outbreaks n = 9 | Dengue diagnosis n = 14 |
|---|---|---|---|
| **Best model** | | | |
| Artificial neural network | 4 (17) | 1 (11) | 4 (29) |
| Decision tree | 4 (17) | 2 (22) | 4 (29) |
| Support vector machine | 4 (17) | 1 (11) | 4 (29) |
| Regression model | 5 (21) | 1 (11) | 0 (0) |
| Time series | 3 (12) | 2 (22) | 0 (0) |
| Bayesian models | 2 (8) | 0 (0) | 0 (0) |
| Association rules | 1 (4) | 1 (11) | 0 (0) |
| Clustering | 0 (0) | 0 (0) | 1 (7) |
| Other | 1 (4) | 1 (11) | 1 (7) |

https://doi.org/10.1371/journal.pntd.0010056.t006

81,84,85,91,92,98,100,102,105,110–112,114,115,126,127,130,135–138]. Indeed, social media and Internet search engines have become widely accessible worldwide, and therefore they represented the most popular novel data streams in the included studies. The easy access to these sources facilitates the assessment of their influence on infectious disease surveillance and prediction [152–154]. This is particularly true for neglected tropical diseases, such as dengue, Zika virus disease and chikungunya, because of their reoccurrence and the massive increase of their incidence in recent years [155,156]. Moreover, harnessing these novel data streams can improve traditional dengue surveillance systems, because they allow the early detection of an outbreak, and thus can decrease delays between the actual dengue outbreak onset and the official case notifications [157,158]. In the case of dengue control, early response is especially important because it can influence the outbreak severity.

Our analysis also identified the underutilization of some data sources. Genomic data and vector-based data were exploited only in 6 of the 119 included studies [35,42,50,57,75,131], despite the importance of vector surveillance in dengue. Moreover, studies using genomic data were based only on human genome data, although scientists could easily access viral genome sequencing data, for instance via the European Virus Archive—GLOBAL (EVAg) [159]. EVAg aim is to offer access to viruses and to virus sequencing data (including dengue) to scientists, government agencies and academic institutions. None of the included studies made use of data provided by this archive. The lack of vector data is surprising because this type of information is crucial in dengue monitoring studies[160,161]. However, we could not evaluate publication bias, especially in the case of underused data sources. As all included studies on the pertinence of a data source found a significant association between the source and a dengue-related outcome, we cannot exclude that some data sources were not underused, but rather not relevant for dengue management. However, the nature of the underused data sources could suggest that there is a dichotomy between data sources and the objectives of dengue studies: the studies focus either on techniques for vector monitoring/prediction or on techniques for human surveillance/prediction, but rarely on both. This dichotomy was also observed within human surveillance and prediction studies. Specifically, health scientists seemed to rely mainly on traditional data, whereas information technology researchers focused more on non-traditional data (especially social networks). Thus, studies using hospital data for dengue prediction rarely leveraged other data sources, such as climate data. Conversely, studies based on non-traditional data sources rarely used human data, besides the official number of dengue case counts. This might be explained by the fact that clinical data are often hard to access for researchers, particularly outside the medical community, for legal and ethical reasons. Furthermore, a substantial number of the selected papers were conference papers from

Information Technology & Sciences Conferences rather than Medicine Conferences. This might reflect the lack of interactions between research teams focused on prediction and/or informatics and physicians and/or government agencies focused on infectious disease monitoring and management. Yet, this research field would greatly benefit from combining their complementary approaches/expertise. Nevertheless, the most commonly studied outcomes in these articles based on real-world data were dengue incidence rate, dengue outbreaks and dengue diagnosis because they need to assess the reliability of novel data streams compared with traditional data sources. As most studies could demonstrate that these sources and methods can complete traditional surveillance and prediction methods, stakeholders should be more aware of these alternative methodologies and novel data streams, and reach out to these highly specialized teams to optimize outbreak dynamic tracking and to improve data completeness and prediction model accuracy.

Most of the included studies relied on machine learning methods, particularly supervised learning models, to assess traditional and also novel data streams. These models were useful also for the analysis of traditional data sources, and allowed scientists to harness non-structured data with NLP methods [40,43,48,49,51–53,56,60,65,66,69–71,73,76,77,79–81,84,85,92,98,100,102,105,110–112,114,115,126,127,130,134–139]. Unsupervised learning models were not the method of choice in most studies, possibly because these studies wanted to identify relevant data sources and/or indicators for dengue monitoring and prediction. Indeed, unsupervised learning tends to be used to identify clusters with similar characteristics [162,163]. Studies that used these methods wanted to predict dengue diagnosis based on the patient clinical profiles or to assess the validity of novel data sources, such as Twitter. Moreover, this approach for dengue research is fairly recent: with the exception of one conference paper from 2011, all studies using unsupervised learning models were published after 2016. Similarly, most studies relying on NLP methods were published rather recently, especially after 2017 (35 of the 42 studies with NLP methods). These two observations suggest that unsupervised learning and NLP might become more prominent in dengue research. It is important to note that despite the use of real-world data, these statistical methods were employed to analyze only retrospective data (but for one study), making their pertinence in real conditions difficult to assess.

Evaluation metrics are crucial in real-world data studies because they help to determine whether the collected data are fit for the purpose (here, dengue surveillance and prediction) and to assess data quality and bias [164]. Although most of the included prediction studies used at least one of the gold standard metrics for information retrieval, such as precision (or positive predictive value) and recall (or sensitivity) [165], several articles employed only error-based metrics, such as root mean square error and mean absolute error. The choice of evaluation metrics is obviously related to the study objective, but even studies where information retrieval metrics could be calculated did not necessarily use them. Again, these methodological choices might be explained by the discrepancy between health scientists who prefer "traditional" modeling evaluation metrics and information technology scientists who focus on information retrieval metrics.

This study also highlighted that despite the variety of approaches to predict dengue outcomes, some factors are constantly relevant, regardless of the study period or country, such as weather-based predictors, artificial neural networks, and decision tree models. However, a consensus on universal models and data sources has not been reached and will probably be difficult to attain due to the complex nature of dengue transmission.

This review has two main weaknesses despite the systematic approach. First, we only searched for published articles and did not look for preprints. Second, besides the experts involved in this review, we could not obtain the opinion of other international experts due to

the infectious disease context of 2020 (COVID-19 and dengue outbreaks in many regions). Therefore, we may have missed relevant studies for the review. Finally, the definition of real-world data can vary according to the stakeholders' view. We had to choose one single definition for the reviewing process, but other definitions do exist. Therefore, we cannot rule out a selection bias in our study.

Overall, this review showed that combining novel real-world and *Big Data* sources with machine learning methods is a promising approach to improve dengue prediction and outbreak monitoring. These new approaches are especially relevant because they can help government agencies and experts to better prepare for each resurgence and better manage outbreaks. Their aim is not to replace existing systems, but to complement them, especially for reducing delays between outbreaks and reporting. Future studies should focus on better integrating all available data sources and methods to improve the stakeholders' response and to better understand dengue outbreaks.

## Supporting information

**S1 Checklist. PRISMA Checklist.**
(DOCX)

**S2 Checklist. PRISMA for Abstracts Checklist.**
(DOCX)

**S1 Text. Data extraction sheet.**
(DOCX)

**S2 Text. PROSPERO Protocol.**
(PDF)

**S1 Table. Quality assessment criteria.**
(DOCX)

**S2 Table. Themes associated with the included studies.**
(DOCX)

**S3 Table. Characteristics of studies included in the systematic review.**
(DOCX)

**S4 Table. Studied outcomes in dengue fever surveillance and prediction.**
(DOCX)

**S5 Table. Detailed study outcomes and results.**
(XLSX)

## Author Contributions

**Conceptualization:** Emmanuelle Sylvestre, Clarisse Joachim, Marc Cuggia, André Cabié.

**Data curation:** Emmanuelle Sylvestre, Clarisse Joachim.

**Formal analysis:** Emmanuelle Sylvestre, Elsa Cécilia-Joseph, Marc Cuggia, André Cabié.

**Supervision:** Guillaume Bouzillé, Boris Campillo-Gimenez, André Cabié.

**Visualization:** Emmanuelle Sylvestre, Elsa Cécilia-Joseph.

**Writing – original draft:** Emmanuelle Sylvestre.

**Writing – review & editing:** Emmanuelle Sylvestre, Clarisse Joachim, Elsa Cécilia-Joseph, Guillaume Bouzillé, Boris Campillo-Gimenez, Marc Cuggia, André Cabié.

## References


1. Bhatt S, Gething PW, Brady OJ, Messina JP, Farlow AW, Moyes CL, et al. The global distribution and burden of dengue. Nature. 2013; 496: 504–507. https://doi.org/10.1038/nature12060 PMID: 23563266
2. Waggoner JJ, Gresh L, Vargas MJ, Ballesteros G, Tellez Y, Soda KJ, et al. Viremia and Clinical Presentation in Nicaraguan Patients Infected With Zika Virus, Chikungunya Virus, and Dengue Virus. Clin Infect Dis Off Publ Infect Dis Soc Am. 2016; 63: 1584–1590. https://doi.org/10.1093/cid/ciw589 PMID: 27578819
3. Katzelnick LC, Coloma J, Harris E. Dengue: knowledge gaps, unmet needs, and research priorities. Lancet Infect Dis. 2017; 17: e88–e100. https://doi.org/10.1016/S1473-3099(16)30473-X PMID: 28185868
4. World Health Organization. Global strategy for dengue prevention and control, 2012–2020. Geneva, Switzerland: World Health Organization; 2012. http://apps.who.int/iris/bitstream/10665/75303/1/9789241504034_eng.pdf
5. Brady OJ, Gething PW, Bhatt S, Messina JP, Brownstein JS, Hoen AG, et al. Refining the global spatial limits of dengue virus transmission by evidence-based consensus. PLoS Negl Trop Dis. 2012; 6: e1760. https://doi.org/10.1371/journal.pntd.0001760 PMID: 22880140
6. Chan M, Johansson MA. The incubation periods of Dengue viruses. PloS One. 2012; 7: e50972. https://doi.org/10.1371/journal.pone.0050972 PMID: 23226436
7. World Health Organization. Dengue Guidelines for Diagnosis, Treatment, Prevention and Control. Special Programme for Research and Training in Tropical Diseases, editor. Geneva: World Health Organization; 2009.
8. Milinovich GJ, Williams GM, Clements ACA, Hu W. Internet-based surveillance systems for monitoring emerging infectious diseases. Lancet Infect Dis. 2014; 14: 160–168. https://doi.org/10.1016/S1473-3099(13)70244-5 PMID: 24290841
9. Madoff LC, Fisman DN, Kass-Hout T. A new approach to monitoring dengue activity. PLoS Negl Trop Dis. 2011; 5: e1215. https://doi.org/10.1371/journal.pntd.0001215 PMID: 21647309
10. Samaras L, Sicilia M-A, García-Barriocanal E. Predicting epidemics using search engine data: a comparative study on measles in the largest countries of Europe. BMC Public Health. 2021; 21: 100. https://doi.org/10.1186/s12889-020-10106-8 PMID: 33472589
11. Lu FS, Hattab MW, Clemente CL, Biggerstaff M, Santillana M. Improved state-level influenza nowcasting in the United States leveraging Internet-based data and network approaches. Nat Commun. 2019; 10: 147. https://doi.org/10.1038/s41467-018-08082-0 PMID: 30635558
12. Carneiro HA, Mylonakis E. Google trends: a web-based tool for real-time surveillance of disease outbreaks. Clin Infect Dis Off Publ Infect Dis Soc Am. 2009; 49: 1557–1564. https://doi.org/10.1086/630200 PMID: 19845471
13. Wilson K, Brownstein JS. Early detection of disease outbreaks using the Internet. CMAJ Can Med Assoc J. 2009; 180: 829–831. https://doi.org/10.1503/cmaj.090215 PMID: 19364791
14. Gianfredi V, Bragazzi NL, Nucci D, Martini M, Rosselli R, Minelli L, et al. Harnessing Big Data for Communicable Tropical and Sub-Tropical Disorders: Implications From a Systematic Review of the Literature. Front Public Health. 2018; 6. https://doi.org/10.3389/fpubh.2018.00090 PMID: 29619364
15. Runge-Ranzinger S, McCall PJ, Kroeger A, Horstick O. Dengue disease surveillance: an updated systematic literature review. Trop Med Int Health TM IH. 2014; 19: 1116–1160. https://doi.org/10.1111/tmi.12333 PMID: 24889501
16. da Silveira LTC, Tura B, Santos M. Systematic review of dengue vaccine efficacy. BMC Infect Dis. 2019; 19: 750. https://doi.org/10.1186/s12879-019-4369-5 PMID: 31455279
17. Gutierrez-Barbosa H, Medina-Moreno S, Zapata JC, Chua JV. Dengue Infections in Colombia: Epidemiological Trends of a Hyperendemic Country. Trop Med Infect Dis. 2020; 5.
18. Ramos-Castañeda J, Barreto Dos Santos F, Martínez-Vega R, Galvão de Araujo JM, Joint G, Sarti E. Dengue in Latin America: Systematic Review of Molecular Epidemiological Trends. PLoS Negl Trop Dis. 2017; 11: e0005224. https://doi.org/10.1371/journal.pntd.0005224 PMID: 28068335
19. Ahmed AM, Mohammed AT, Vu TT, Khattab M, Doheim MF, Ashraf Mohamed A, et al. Prevalence and burden of dengue infection in Europe: A systematic review and meta-analysis. Rev Med Virol. 2020; 30: e2093. https://doi.org/10.1002/rmv.2093 PMID: 31833169

20. Simo FBN, Bigna JJ, Kenmoe S, Ndangang MS, Temfack E, Moundipa PF, et al. Dengue virus infection in people residing in Africa: a systematic review and meta-analysis of prevalence studies. Sci Rep. 2019; 9: 13626. https://doi.org/10.1038/s41598-019-50135-x PMID: 31541167

21. Cafferata ML, Bardach A, Rey-Ares L, Alcaraz A, Cormick G, Gibbons L, et al. Dengue Epidemiology and Burden of Disease in Latin America and the Caribbean: A Systematic Review of the Literature and Meta-Analysis. Value Health Reg Issues. 2013; 2: 347–356. https://doi.org/10.1016/j.vhri.2013.10.002 PMID: 29702769

22. Dao Phuoc T, Khuong Quynh L, Vien Dang Khanh L, Ong Phuc T, Le Sy H, Le Ngoc T, et al. Clinical prognostic models for severe dengue: a systematic review protocol. Wellcome Open Res. 2019; 4: 12. https://doi.org/10.12688/wellcomeopenres.15033.2 PMID: 31448337

23. Moher D, Liberati A, Tetzlaff J, Altman DG, for the PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. BMJ. 2009; 339: b2535–b2535. https://doi.org/10.1136/bmj.b2535 PMID: 19622551

24. Sherman RE, Anderson SA, Dal Pan GJ, Gray GW, Gross T, Hunter NL, et al. Real-World Evidence—What Is It and What Can It Tell Us? N Engl J Med. 2016; 375: 2293–2297. https://doi.org/10.1056/NEJMsb1609216 PMID: 27959688

25. Baruch P. Open Access Developments in France: the HAL Open Archives System. Learn Publ. 2007; 20: 267–282. https://doi.org/10.1087/095315107X239636

26. Agence bibliographique de l'enseignement. Thèses. Agence bibliographique de l'enseignement supérieur (ABES); [cited 1 Apr 2021]. http://www.theses.fr

27. Aswi A, Cramb SM, Moraga P, Mengersen K. Bayesian spatial and spatio-temporal approaches to modelling dengue fever: a systematic review. Epidemiol Infect. 2019; 147. https://doi.org/10.1017/S0950268818002807 PMID: 30369335

28. Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. Ann Intern Med. 2015; 162: W1–W73. https://doi.org/10.7326/M14-0698 PMID: 25560730

29. Wang W, Kiik M, Peek N, Curcin V, Marshall IJ, Rudd AG, et al. A systematic review of machine learning models for predicting outcomes of stroke with structured data. PLOS ONE. 2020; 15: e0234722. https://doi.org/10.1371/journal.pone.0234722 PMID: 32530947

30. R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.r-project.org/index.html

31. Polwiang S. The time series seasonal patterns of dengue fever and associated weather variables in Bangkok (2003–2017). BMC Infect Dis. 2020; 20: 208. https://doi.org/10.1186/s12879-020-4902-6 PMID: 32164548

32. Xu J, Xu K, Li Z, Meng F, Tu T, Xu L, et al. Forecast of Dengue Cases in 20 Chinese Cities Based on the Deep Learning Method. Int J Environ Res Public Health. 2020; 17. https://doi.org/10.3390/ijerph17020453 PMID: 31936708

33. Rangarajan P, Mody SK, Marathe M. Forecasting dengue and influenza incidences using a sparse representation of Google trends, electronic health records, and time series data. PLoS Comput Biol. 2019; 15: e1007518. https://doi.org/10.1371/journal.pcbi.1007518 PMID: 31751346

34. Anno S, Hara T, Kai H, Lee M-A, Chang Y, Oyoshi K, et al. Spatiotemporal dengue fever hotspots associated with climatic factors in Taiwan including outbreak predictions based on machine-learning. Geospatial Health. 2019; 14. https://doi.org/10.4081/gh.2019.771 PMID: 31724367

35. Romero D, Olivero J, Real R, Guerrero JC. Applying fuzzy logic to assess the biogeographical risk of dengue in South America. Parasit Vectors. 2019; 12: 428. https://doi.org/10.1186/s13071-019-3691-5 PMID: 31488198

36. Mello-Román JD, Mello-Román JC, Gómez-Guerrero S, García-Torres M. Predictive Models for the Medical Diagnosis of Dengue: A Case Study in Paraguay. Comput Math Methods Med. 2019; 2019: 7307803. https://doi.org/10.1155/2019/7307803 PMID: 31485259

37. Stolerman LM, Maia PD, Kutz JN. Forecasting dengue fever in Brazil: An assessment of climate conditions. PloS One. 2019; 14: e0220106. https://doi.org/10.1371/journal.pone.0220106 PMID: 31393908

38. Macedo Hair G, Fonseca Nobre F, Brasil P. Characterization of clinical patterns of dengue patients using an unsupervised machine learning approach. BMC Infect Dis. 2019; 19: 649. https://doi.org/10.1186/s12879-019-4282-y PMID: 31331271

39. Husnayain A, Fuad A, Lazuardi L. Correlation between Google Trends on dengue fever and national surveillance report in Indonesia. Glob Health Action. 2019; 12: 1552652. https://doi.org/10.1080/16549716.2018.1552652 PMID: 31154985

40. Souza RCSNP, Assunção RM, Oliveira DM, Neill DB, Meira W. Where did I get dengue? Detecting spatial clusters of infection risk with social network data. Spat Spatio-Temporal Epidemiol. 2019; 29: 163–175. https://doi.org/10.1016/j.sste.2018.11.005 PMID: 31128626

41. Ramadona AL, Tozan Y, Lazuardi L, Rocklöv J. A combination of incidence data and mobility proxies from social media predicts the intra-urban spread of dengue in Yogyakarta, Indonesia. PLoS Negl Trop Dis. 2019; 13: e0007298. https://doi.org/10.1371/journal.pntd.0007298 PMID: 30986218

42. Davi C, Pastor A, Oliveira T, Neto FB de L, Braga-Neto U, Bigham AW, et al. Severe Dengue Prognosis Using Human Genome Data and Machine Learning. IEEE Trans Biomed Eng. 2019; 66: 2861–2868. https://doi.org/10.1109/TBME.2019.2897285 PMID: 30716030

43. Guo P, Zhang Q, Chen Y, Xiao J, He J, Zhang Y, et al. An ensemble forecast model of dengue in Guangzhou, China using climate and social media surveillance data. Sci Total Environ. 2019; 647: 752–762. https://doi.org/10.1016/j.scitotenv.2018.08.044 PMID: 30092532

44. Koh Y-M, Spindler R, Sandgren M, Jiang J. A model comparison algorithm for increased forecast accuracy of dengue fever incidence in Singapore and the auxiliary role of total precipitation information. Int J Environ Health Res. 2018; 28: 535–552. https://doi.org/10.1080/09603123.2018.1496234 PMID: 30016117

45. Carvajal TM, Viacrusis KM, Hernandez LFT, Ho HT, Amalin DM, Watanabe K. Machine learning methods reveal the temporal pattern of dengue incidence using meteorological factors in metropolitan Manila, Philippines. BMC Infect Dis. 2018; 18: 183. https://doi.org/10.1186/s12879-018-3066-0 PMID: 29665781

46. Baquero OS, Santana LMR, Chiaravalloti-Neto F. Dengue forecasting in São Paulo city with generalized additive models, artificial neural networks and seasonal autoregressive integrated moving average models. PloS One. 2018; 13: e0195065. https://doi.org/10.1371/journal.pone.0195065 PMID: 29608586

47. Chen Y, Chu CW, Chen MIC, Cook AR. The utility of LASSO-based models for real time forecasts of endemic infectious diseases: A cross country comparison. J Biomed Inform. 2018; 81: 16–30. https://doi.org/10.1016/j.jbi.2018.02.014 PMID: 29496631

48. Villanes A, Griffiths E, Rappa M, Healey CG. Dengue Fever Surveillance in India Using Text Mining in Public Media. Am J Trop Med Hyg. 2018; 98: 181–191. https://doi.org/10.4269/ajtmh.17-0253 PMID: 29141718

49. Guo P, Liu T, Zhang Q, Wang L, Xiao J, Zhang Q, et al. Developing a dengue forecast model using machine learning: A case study in China. PLoS Negl Trop Dis. 2017; 11: e0005973. https://doi.org/10.1371/journal.pntd.0005973 PMID: 29036169

50. Chatterjee S, Dey N, Shi F, Ashour AS, Fong SJ, Sen S. Clinical application of modified bag-of-features coupled with hybrid neural-based classifier in dengue fever classification using gene expression data. Med Biol Eng Comput. 2018; 56: 709–720. https://doi.org/10.1007/s11517-017-1722-y PMID: 28891000

51. Guo P, Wang L, Zhang Y, Luo G, Zhang Y, Deng C, et al. Can internet search queries be used for dengue fever surveillance in China? Int J Infect Dis IJID Off Publ Int Soc Infect Dis. 2017; 63: 74–76. https://doi.org/10.1016/j.ijid.2017.08.001 PMID: 28797591

52. Yang S, Kou SC, Lu F, Brownstein JS, Brooke N, Santillana M. Advances in using Internet searches to track dengue. PLoS Comput Biol. 2017; 13: e1005607. https://doi.org/10.1371/journal.pcbi.1005607 PMID: 28727821

53. Marques-Toledo C de A, Degener CM, Vinhal L, Coelho G, Meira W, Codeço CT, et al. Dengue prediction by the web: Tweets are a useful tool for estimating and forecasting Dengue at country and city level. PLoS Negl Trop Dis. 2017; 11: e0005729. https://doi.org/10.1371/journal.pntd.0005729 PMID: 28719659

54. Premaratne MK, Perera SSN, Malavige GN, Jayasinghe S. Mathematical Modelling of Immune Parameters in the Evolution of Severe Dengue. Comput Math Methods Med. 2017; 2017: 2187390. https://doi.org/10.1155/2017/2187390 PMID: 28293273

55. Jayasundara SDP, Perera SSN, Malavige GN, Jayasinghe S. Mathematical modelling and a systems science approach to describe the role of cytokines in the evolution of severe dengue. BMC Syst Biol. 2017; 11: 34. https://doi.org/10.1186/s12918-017-0415-3 PMID: 28284213

56. Li Z, Liu T, Zhu G, Lin H, Zhang Y, He J, et al. Dengue Baidu Search Index data can improve the prediction of local dengue epidemic: A case study in Guangzhou, China. PLoS Negl Trop Dis. 2017; 11: e0005354. https://doi.org/10.1371/journal.pntd.0005354 PMID: 28263988

57. Kesorn K, Ongruk P, Chompoosri J, Phumee A, Thavara U, Tawatsin A, et al. Morbidity Rate Prediction of Dengue Hemorrhagic Fever (DHF) Using the Support Vector Machine and the Aedes aegypti Infection Rate in Similar Climates and Geographical Areas. PloS One. 2015; 10: e0125049. https://doi.org/10.1371/journal.pone.0125049 PMID: 25961289

58. Dayama P, Sampath K. Dengue disease outbreak detection. Stud Health Technol Inform. 2014; 205: 1105–1109. PMID: 25160360

59. Sampath K, Dayama P. Predicting the operations alert levels for dengue surveillance and control. Stud Health Technol Inform. 2014; 205: 1100–1104. PMID: 25160359

60. Gluskin RT, Johansson MA, Santillana M, Brownstein JS. Evaluation of Internet-based dengue query data: Google Dengue Trends. PLoS Negl Trop Dis. 2014; 8: e2713. https://doi.org/10.1371/journal.pntd.0002713 PMID: 24587465

61. Flamand C, Fritzell C, Prince C, Abboud P, Ardillon V, Carvalho L, et al. Epidemiological assessment of the severity of dengue epidemics in French Guiana. PloS One. 2017; 12: e0172267. https://doi.org/10.1371/journal.pone.0172267 PMID: 28196111

62. Torres C, Barguil S, Melgarejo M, Olarte A. Fuzzy model identification of dengue epidemic in Colombia based on multiresolution analysis. Artif Intell Med. 2014; 60: 41–51. https://doi.org/10.1016/j.artmed.2013.11.008 PMID: 24388398

63. Buczak AL, Koshute PT, Babin SM, Feighner BH, Lewis SH. A data-driven epidemiological prediction method for dengue outbreaks using local and remote sensing data. BMC Med Inform Decis Mak. 2012; 12: 124. https://doi.org/10.1186/1472-6947-12-124 PMID: 23126401

64. Hoen AG, Keller M, Verma AD, Buckeridge DL, Brownstein JS. Electronic event-based surveillance for monitoring dengue, Latin America. Emerg Infect Dis. 2012; 18: 1147–1150. https://doi.org/10.3201/eid1807.120055 PMID: 22709430

65. Althouse BM, Ng YY, Cummings DAT. Prediction of dengue incidence using search query surveillance. PLoS Negl Trop Dis. 2011; 5: e1258. https://doi.org/10.1371/journal.pntd.0001258 PMID: 21829744

66. Chan EH, Sahai V, Conrad C, Brownstein JS. Using web search query data to monitor dengue epidemics: a new model for neglected tropical disease surveillance. PLoS Negl Trop Dis. 2011; 5: e1206. https://doi.org/10.1371/journal.pntd.0001206 PMID: 21647308

67. Faisal T, Taib MN, Ibrahim F. Neural network diagnostic system for dengue patients risk classification. J Med Syst. 2012; 36: 661–676. https://doi.org/10.1007/s10916-010-9532-x PMID: 20703665

68. Ibrahim F, Faisal T, Salim MIM, Taib MN. Non-invasive diagnosis of risk in dengue patients using bioelectrical impedance analysis and artificial neural network. Med Biol Eng Comput. 2010; 48: 1141–1148. https://doi.org/10.1007/s11517-010-0669-z PMID: 20683676

69. Syamsuddin M, Fakhruddin M, Sahetapy-Engel JTM, Soewono E. Causality Analysis of Google Trends and Dengue Incidence in Bandung, Indonesia With Linkage of Digital Data Modeling: Longitudinal Observational Study. J Med Internet Res. 2020; 22: e17633. https://doi.org/10.2196/17633 PMID: 32706682

70. Romero-Alvarez D, Parikh N, Osthus D, Martinez K, Generous N, Del Valle S, et al. Google Health Trends performance reflecting dengue incidence for the Brazilian states. BMC Infect Dis. 2020; 20: 252. https://doi.org/10.1186/s12879-020-04957-0 PMID: 32228508

71. Liu D, Guo S, Zou M, Chen C, Deng F, Xie Z, et al. A dengue fever predicting model based on Baidu search index data and climate data in South China. PloS One. 2019; 14: e0226841. https://doi.org/10.1371/journal.pone.0226841 PMID: 31887118

72. Musa SS, Zhao S, Chan H-S, Jin Z, He DH. A mathematical model to study the 2014–2015 large-scale dengue epidemics in Kaohsiung and Tainan cities in Taiwan, China. Math Biosci Eng MBE. 2019; 16: 3841–3863. https://doi.org/10.3934/mbe.2019190 PMID: 31499639

73. Messina JP, Brady OJ, Golding N, Kraemer MUG, Wint GRW, Ray SE, et al. The current and future global distribution and population at risk of dengue. Nat Microbiol. 2019; 4: 1508–1515. https://doi.org/10.1038/s41564-019-0476-8 PMID: 31182801

74. Titus Muurlink O, Stephenson P, Islam MZ, Taylor-Robinson AW. Long-term predictors of dengue outbreaks in Bangladesh: A data mining approach. Infect Dis Model. 2018; 3: 322–330. https://doi.org/10.1016/j.idm.2018.11.004 PMID: 30839927

75. Marques-Toledo CA, Bendati MM, Codeço CT, Teixeira MM. Probability of dengue transmission and propagation in a non-endemic temperate area: conceptual model and decision risk levels for early alert, prevention and control. Parasit Vectors. 2019; 12: 38. https://doi.org/10.1186/s13071-018-3280-z PMID: 30651125

76. Verma M, Kishore K, Kumar M, Sondh AR, Aggarwal G, Kathirvel S. Google Search Trends Predicting Disease Outbreaks: An Analysis from India. Healthc Inform Res. 2018; 24: 300–308. https://doi.org/10.4258/hir.2018.24.4.300 PMID: 30443418

77. Ho HT, Carvajal TM, Bautista JR, Capistrano JDR, Viacrusis KM, Hernandez LFT, et al. Using Google Trends to Examine the Spatio-Temporal Incidence and Behavioral Patterns of Dengue Disease: A

Case Study in Metropolitan Manila, Philippines. Trop Med Infect Dis. 2018; 3. https://doi.org/10.3390/tropicalmed3040118 PMID: 30423898

78. Phakhounthong K, Chaovalit P, Jittamala P, Blacksell SD, Carter MJ, Turner P, et al. Predicting the severity of dengue fever in children on admission based on clinical features and laboratory indicators: application of classification tree analysis. BMC Pediatr. 2018; 18: 109. https://doi.org/10.1186/s12887-018-1078-y PMID: 29534694

79. Strauss RA, Castro JS, Reintjes R, Torres JR. Google dengue trends: An indicator of epidemic behavior. The Venezuelan Case. Int J Med Inf. 2017; 104: 26–30. https://doi.org/10.1016/j.ijmedinf.2017.05.003 PMID: 28599813

80. Nsoesie EO, Flor L, Hawkins J, Maharana A, Skotnes T, Marinho F, et al. Social Media as a Sentinel for Disease Surveillance: What Does Sociodemographic Status Have to Do with It? PLoS Curr. 2016; 8. https://doi.org/10.1371/currents.outbreaks.cc09a42586e16dc7dd62813b7ee5d6b6 PMID: 28123858

81. Liu K, Wang T, Yang Z, Huang X, Milinovich GJ, Lu Y, et al. Using Baidu Search Index to Predict Dengue Outbreak in China. Sci Rep. 2016; 6: 38040. https://doi.org/10.1038/srep38040 PMID: 27905501

82. Ximenes R, Amaku M, Lopez LF, Coutinho FAB, Burattini MN, Greenhalgh D, et al. The risk of dengue for non-immune foreign visitors to the 2016 summer olympic games in Rio de Janeiro, Brazil. BMC Infect Dis. 2016; 16: 186. https://doi.org/10.1186/s12879-016-1517-z PMID: 27129407

83. Mohamad Mohsin MF, Abu Bakar A, Hamdan AR. Outbreak detection model based on danger theory. Appl Soft Comput. 2014; 24: 612–622. https://doi.org/10.1016/j.asoc.2014.08.030 PMID: 32362801

84. Puengpreeda A, Yhusumrarn S, Sirikulvadhana S. Weekly Forecasting Model for Dengue Hemorrhagic Fever Outbreak in Thailand. Eng J-Thail. 2020; 24: 71–87. https://doi.org/10.4186/ej.2020.24.3.71

85. Amin S, Uddin MI, Hassan S, Khan A, Nasser N, Alharbi A, et al. Recurrent Neural Networks With TF-IDF Embedding Technique for Detection and Classification in Tweets of Dengue Disease. Ieee Access. 2020; 8: 131522–131533. https://doi.org/10.1109/ACCESS.2020.3009058

86. Manogaran G, Lopez D, Chilamkurti N. In-Mapper combiner based MapReduce algorithm for processing of big climate data. Future Gener Comput Syst- Int J Escience. 2018; 86: 433–445. https://doi.org/10.1016/j.future.2018.02.048

87. Agarwal N, Koti SR, Saran S, Kumar AS. Data mining techniques for predicting dengue outbreak in geospatial domain using weather parameters for New Delhi, India. Curr Sci. 2018; 114: 2281–2291. https://doi.org/10.18520/cs/v114/i11/2281-2291

88. Manogaran G, Lopez D. A Gaussian process based big data processing framework in cluster computing environment. Clust Comput- J Netw Softw Tools Appl. 2018; 21: 189–204. https://doi.org/10.1007/s10586-017-0982-5

89. Jahangir I, Abdul-Basit, Hannan A, Javed S. Prediction of Dengue Disease Through Data Mining by Using Modified Apriori Algorithm. Proceedings of the 4th Acm International Conference of Computing for Engineering and Sciences (icces'2018). New York: Assoc Computing Machinery; 2018.

90. Husin NA, Alharogi A, Mustapha N, Hamdan H, Husin UA. Early Self-Diagnosis of Dengue Symptoms Using Fuzzy and Data Mining Approach. In: Nifa F a. A, Lin CK, Hussain A, editors. Proceedings of the 3rd International Conference on Applied Science and Technology (icast'18). Melville: Amer Inst Physics; 2018. p. 020048.

91. Anggraeni W, Pramudita G, Riksakomara E, Radityo PW, Samopa F, Pujiadi, et al. Artificial Neural Network for Health Data Forecasting, Case Study: Number of Dengue Hemorrhagic Fever Cases in Malang Regency, Indonesia. 2018 International Conference on Electrical Engineering and Computer Science (icecos). New York: Ieee; 2018. pp. 207–212.

92. Dennison Livelo E, Cheng C. Intelligent Dengue Infoveillance Using Gated Recurrent Neural Learning and Cross-Label Frequencies. 2018 Ieee International Conference on Agents (ica). New York: Ieee; 2018. pp. 2–7.

93. Wiratmadja II, Salamah SY, Govindaraju R. Healthcare Data Mining: Predicting Hospital Length of Stay of Dengue Patients. J Eng Technol Sci. 2018; 50: 110–126. https://doi.org/10.5614/j.eng.technol.sci.2018.50.1.8

94. Arafiyah R, Hermin F. Data mining for dengue hemorrhagic fever (DHF) prediction with naive Bayes method. 1st International Conference of Education on Sciences, Technology, Engineering, and Mathematics (ice-Stem). Bristol: Iop Publishing Ltd; 2018. p. 012077.

95. Abuhamad HIS, Abu Bakar A, Zainudin S, Sahani M, Ali ZM. Feature Selection Algorithms for Malaysian Dengue Outbreak Detection Model. Sains Malays. 2017; 46: 255–265. https://doi.org/10.17576/jsm-2017-4602-10

96. Manivannan P, Devi PI. Dengue Fever Prediction Using K-Means Clustering Algorithm. 2017 Ieee International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (incos). New York: Ieee; 2017.

97. Dharmawardana KGS, Lokuge JN, Dassanayake PSB, Sirisena ML, Fernando ML, Perera AS, et al. Predictive Model for the Dengue Incidences in Sri Lanka Using Mobile Network Big Data. 2017 Ieee International Conference on Industrial and Information Systems (iciis). New York: Ieee; 2017. pp. 278–283.

98. Espina K, Estuar MRJE. Infodemiology for Syndromic Surveillance of Dengue and Typhoid Fever in the Philippines. In: CruzCunha MM, Varajao JEQ, Rijo R, Martinho R, Peppard J, SanCristobal JR, et al., editors. Centeris 2017—International Conference on Enterprise Information Systems / Projman 2017—International Conference on Project Management / Hcist 2017—International Conference on Health and Social Care Information Systems and Technologies, Centeri. Amsterdam: Elsevier Science Bv; 2017. pp. 554–561.

99. Rahim NF, Taib SM, Abidin AIZ. Dengue Fatality Prediction Using Data Mining. J Fundam Appl Sci. 2017; 9: 671–683. https://doi.org/10.4314/jfas.v9i6s.52

100. Klein GH, Neto PG, Tezza R. Big Data and social media: surveillance of networks as management tool. Saude E Soc. 2017; 26: 208–217. https://doi.org/10.1590/S0104-12902017164943

101. Kerdprasop N, Kerdprasop K. Remote Sensing Based Modeling of Dengue Outbreak with Regression and Binning Classification. 2016 2nd Ieee International Conference on Computer and Communications (iccc). New York: Ieee; 2016. pp. 46–49.

102. Anggraeni W, Aristiani L. Using Google Trend Data in Forecasting Number of Dengue Fever Cases with ARIMAX Method Case Study: Surabaya, Indonesia. Proceedings of 2016 International Conference on Information & Communication Technology and Systems (icts). New York: Ieee; 2016. pp. 114–118.

103. Mathulamuthu SS, Asirvadam VS, Dass SC, Gill BS, Loshini T. Predicting Dengue Incidences Using Cluster Based Regression on Climate Data. 2016 6th Ieee International Conference on Control System, Computing and Engineering (iccsce). New York: Ieee; 2016. pp. 245–250.

104. Rahmawati D, Huang Y-P. Using C-support Vector Classification to Forecast Dengue Fever Epidemics in Taiwan. In: Wang WJ, Lee PJ, Er MJ, Jeng JT, editors. 2016 International Conference on System Science and Engineering (icsse). New York: Ieee; 2016.

105. Missier P, Romanovsky A, Miu T, Pal A, Daniilakis M, Garcia A, et al. Tracking Dengue Epidemics Using Twitter Content Classification and Topic Modelling. In: Casteleyn S, Dolog P, Pautasso C, editors. Current Trends in Web Engineering, Icwe 2016 International Workshops. Cham: Springer International Publishing Ag; 2016. pp. 80–92.

106. Abeyrathna MP a. R, Abeygunawrdane DA, Wijesundara R a. a. V, Mudalige VB, Bandara M, Perera S, et al. Dengue Propagation Prediction using Human Mobility. 2nd International Mercon 2016 Moratuwa Engineering Research Conference. New York: Ieee; 2016. pp. 156–161.

107. Fathima AS, Manimeglai D. Analysis of Significant Factors for Dengue Infection Prognosis Using the Random Forest Classifier. Int J Adv Comput Sci Appl. 2015; 6: 240–245.

108. Tazkia RAK, Narita V, Nugroho AS. Dengue Outbreak Prediction for GIS based Early Warning System. 2015 International Conference on Science in Information Technology (ICSITech). New York: Ieee; 2015. pp. 121–125.

109. Wu Y, Lee G, Fu X, Hung T. Detect climatic factors contributing to dengue outbreak based on wavelet, support vector machines and genetic algorithm. In: Ao SI, Gelman L, Hukins DWL, Hunter A, Korsunsky AM, editors. World Congress on Engineering 2008, Vols I-Ii. Hong Kong: Int Assoc Engineers-Iaeng; 2008. pp. 303-+.

110. Salam N, Deeba F, Qadir F, Al-Hijli F, Al-Otaibi YN. Analysis of Correlation between Google Search Trends and Dengue Outbreaks from India. J Clin Diagn Res. 2019; 13: LC13–LC15. https://doi.org/10.7860/JCDR/2019/42611.13304

111. Chire Saire JE. Building Intelligent Indicators to Detect Dengue Epidemics in Brazil using Social Networks. OrjuelaCanon AD, editor. 2019 Ieee Colombian Conference on Applications in Computational Intelligence (colcaci). New York: Ieee; 2019.

112. Swain S, Seeja KR. Analysis of Epidemic Outbreak in Delhi Using Social Media Data. In: Kaushik S, Gupta D, Kharb L, Chahal D, editors. Information, Communication and Computing Technology. Singapore: Springer-Verlag Singapore Pte Ltd; 2017. pp. 25–34.

113. Saravanan N, Gayathri V. Classification of Dengue Dataset Using J48 Algorithm and Ant Colony Based Aj48 Algorithm. Proceedings of the International Conference on Inventive Computing and Informatics (icici 2017). New York: Ieee; 2017. pp. 1062–1067.

114. Carlos MA, Nogueira M, Machado RJ. Analysis of Dengue Outbreaks Using Big Data Analytics and Social Networks. 2017 4th International Conference on Systems and Informatics (icsai). New York: Ieee; 2017. pp. 1592–1597.

115. Ye X, Li S, Yang X, Qin C. Use of Social Media for the Detection and Analysis of Infectious Diseases in China. Isprs Int J Geo-Inf. 2016; 5: 156. https://doi.org/10.3390/ijgi5090156

116. Li W, Chen Y. Risk Factor Identification and Spatiotemporal Diffusion Path During the Dengue Outbreak. In: Weng Q, Gamba P, Xian G, Chen JM, Liang S, editors. 2016 4rth International Workshop on Earth Observation and Remote Sensing Applications (EORSA). New York: Ieee; 2016.

117. Srilekha G, Anupama B. Prediction of Dengue Outbreaks with Big Data using Machine Learning. GEDRAG Organ Rev. 2020; 33. https://doi.org/10.37896/GOR33.02/073

118. Ganthimathi M, Thangamani M, Mallika C, Prasanna Balaji V. Prediction of dengue fever using intelligent classifier. Int J Emerg Trends Eng Res. 2020; 8: 1338–1341. https://doi.org/10.30534/ijeter/2020/65842020

119. Kumar NK, Sikamani KT. Prediction of chronic and infectious diseases using machine learning classifiers-A systematic approach. Int J Intell Eng Syst. 2020; 13: 11–20. https://doi.org/10.22266/IJIES2020.0831.02

120. Guiyab RB. Development of prediction models for the dengue survivability prediction: An integration of data mining and decision support system. Int J Innov Technol Explor Eng. 2019; 8: 2199–2205. https://doi.org/10.35940/ijitee.J9411.0881019

121. Chovatiya M, Dhameliya A, Deokar J, Gonsalves J, Mathur A. Prediction of dengue using recurrent neural network. 2019. pp. 926–929. https://doi.org/10.1109/icoei.2019.8862581

122. Kerdprasop K, Kerdprasop N, Chansilp K, Chuaybamroong P. The Use of Spaceborne and Oceanic Sensors to Model Dengue Incidence in the Outbreak Surveillance System. Lect Notes Comput Sci Subser Lect Notes Artif Intell Lect Notes Bioinforma. 2019;11619 LNCS: 447–460.

123. Link H, Richter SN, Leung VJ, Brost RC, Phillips CA, Staid A. Statistical models of dengue fever. Commun Comput Inf Sci. 2019; 996: 175–186. https://doi.org/10.1007/978-981-13-6661-1_14

124. Arafiyah R, Hermin F, Kartika IR, Alimuddin A, Saraswati I. Classification of Dengue Haemorrhagic Fever (DHF) using SVM, naive bayes and random forest. 2018.

125. Mishra S, Tripathy HK, Panda AR. An improved and adaptive attribute selection technique to optimize Dengue fever prediction. Int J Eng Technol. 2018; 7: 480–486. https://doi.org/10.14419/ijet.v7i2.29.13802

126. Wu C-H, Kao S-C, Kan M-H. Knowledge discovery in open data of dengue epidemic. 2017.

127. Albinati J, Meira W Jr, Pappa GL, Teixeira M, Marques-Toledo C. Enhancement of epidemiological models for dengue fever based on twiter data. 2017. pp. 109–118.

128. Zhu G, Hunter J, Jiang Y. Improved Prediction of Dengue Outbreak Using the Delay Permutation Entropy. 2017. pp. 828–832.

129. Zainudin Z, Shamsuddin SM. Predictive analytics in Malaysian dengue data from 2010 until 2015 using BigML. Int J Adv Soft Comput Its Appl. 2016; 8: 18–30.

130. Milinovich GJ, Avril SMR, Clements ACA, Brownstein JS, Tong S, Hu W. Using internet search queries for infectious disease surveillance: Screening diseases for suitability. BMC Infect Dis. 2014; 14. https://doi.org/10.1186/s12879-014-0690-1 PMID: 25551277

131. Ongruk P, Siriyasatien P, Kesorn K. New key factors discovery to enhance dengue fever forecasting model. Adv Mater Res. 2014; 931–932: 1457–1461. https://doi.org/10.4028/www.scientific.net/AMR.931-932.1457

132. Balasundaram A, Bhuvaneswari PTV. Comparative study on decision tree based data mining algorithm to assess risk of epidemic. 2013. pp. 390–396.

133. Wu Y, Lee G, Fu X, Soh H, Hung T. Mining weather information in dengue outbreak: Predicting future cases based on wavelet, SVM and GA. Lect Notes Electr Eng. 2009;39 LNEE: 483–494.

134. Zhang Y, Ibaraki M, Schwartz FW. Disease surveillance using online news: Dengue and zika in tropical countries. J Biomed Inform. 2020; 102. https://doi.org/10.1016/j.jbi.2020.103374 PMID: 31911171

135. Souza RCSNP. Detecting spatial clusters of infection risk with geo-located social media data. 2018.

136. Coberly JS, Fink CR, Elbert Y, Yoon I-K, Velasco JM, Tomayao AD, et al. Tweeting Fever: Can Twitter Be Used to Monitor the Incidence of Dengue-Like Illness in the Philippines? JOHNS HOPKINS APL Tech Dig. 2014; 32: 12.

137. Gomide J, Veloso A, Meira Jr. W, Almeida V, Benevenuto F, Ferraz F, et al. Dengue surveillance based on a computational model of spatio-temporal locality of Twitter. 2011.

138. Fang Z-H, Tzeng J-S, Chen CC, Chou T-C. A study of machine learning models in epidemic surveillance: Using the query logs of search engines. 2010. pp. 1438–1449.

139. Souza J, Leung CK, Cuzzocrea A. An Innovative Big Data Predictive Analytics Framework over Hybrid Big Data Sources with an Application for Disease Analytics. Adv Intell Syst Comput. 2020;1151 AISC: 669–680.

140. Yogapriya P, Geetha P. Dengue disease detection using K-means, hierarchical, kohonen-SOM clustering. Int J Innov Technol Explor Eng. 2019; 8: 904–907. https://doi.org/10.35940/ijitee.J9066.0881019

141. Adias Sabara M, Somantri O, Nurcahyo H, Kurnia Achmadi N, Latifah U, Harsono. Diagnosis classification of dengue fever based on Neural Networks and Genetic algorithms. 2019.

142. Jongmuenwai B, Lowanichchai S, Jabjone S. Comparision using data mining algorithm techniques for predicting of dengue fever data in northeastern of Thailand. 2019. pp. 532–535.

143. Balasaravanan K, Prakash M. Detection of dengue disease using artificial neural network based classification techniquetion. Int J Eng Technol. 2018; 7: 13–15. https://doi.org/10.14419/ijet.v7i1.3.8978

144. Acosta Torres J, Oller Meneses L, Sokol N, Balado Sardiñas R, Montero Díaz D, Balado Sansón R, et al. Decision tree technique applied to the clinical method in the dengue diagnosis. Rev Cuba Pediatr. 2016; 88: 441–453.

145. Soonthornphisaj N, Thitiprayoonwongse D. Knowledge discovery on dengue patients using data mining techniques. 2016. pp. 371–375.

146. Fathima SA, Hundewale N. Comparitive analysis of machine learning techniques for classification of arbovirus. 2012. pp. 376–379.

147. Fathima S, Hundewale N. Comparison of Classification Techniques-SVM and Naives Bayes to predict the Arboviral Disease-Dengue. In: Chen B, Chen J, Chen X, Chen Y, Cho YR, Cui J, et al., editors. 2011 Ieee International Conference on Bioinformatics and Biomedicine Workshops. Los Alamitos: Ieee Computer Soc; 2011. pp. 538–539.

148. Long ZA, Abu Bakar A, Razak Hamdan A, Sahani M. Multiple attribute frequent mining-based for dengue outbreak. Lect Notes Comput Sci Subser Lect Notes Artif Intell Lect Notes Bioinforma. 2010;6440 LNAI: 489–496.

149. Stanaway JD, Shepard DS, Undurraga EA, Halasa YA, Coffeng LE, Brady OJ, et al. The Global Burden of Dengue: an analysis from the Global Burden of Disease Study 2013. Lancet Infect Dis. 2016; 16: 712–723. https://doi.org/10.1016/S1473-3099(16)00026-8 PMID: 26874619

150. Zeng Z, Zhan J, Chen L, Chen H, Cheng S. Global, regional, and national dengue burden from 1990 to 2017: A systematic analysis based on the global burden of disease study 2017. EClinicalMedicine. 2021; 32: 100712. https://doi.org/10.1016/j.eclinm.2020.100712 PMID: 33681736

151. World Health Organization. A Global Brief on Vector-Borne Diseases. Geneva: World Health Organization; 2014. https://apps.who.int/iris/bitstream/handle/10665/111008/WHO_DCO_WHD_2014.1_eng.pdf

152. Brownstein JS, Freifeld CC, Madoff LC. Digital Disease Detection—Harnessing the Web for Public Health Surveillance. N Engl J Med. 2009; 360: 2153–2157. https://doi.org/10.1056/NEJMp0900702 PMID: 19423867

153. Choi J, Cho Y, Shim E, Woo H. Web-based infectious disease surveillance systems and public health perspectives: a systematic review. BMC Public Health. 2016; 16: 1238. https://doi.org/10.1186/s12889-016-3893-0 PMID: 27931204

154. Velasco E, Agheneza T, Denecke K, Kirchner G, Eckmanns T. Social media and internet-based data in global systems for public health surveillance: a systematic review. Milbank Q. 2014; 92: 7–33. https://doi.org/10.1111/1468-0009.12038 PMID: 24597553

155. Leta S, Beyene TJ, De Clercq EM, Amenu K, Kraemer MUG, Revie CW. Global risk mapping for major diseases transmitted by Aedes aegypti and Aedes albopictus. Int J Infect Dis. 2018; 67: 25–35. https://doi.org/10.1016/j.ijid.2017.11.026 PMID: 29196275

156. Monaghan AJ, Sampson KM, Steinhoff DF, Ernst KC, Ebi KL, Jones B, et al. The potential impacts of 21st century climatic and population changes on human exposure to the virus vector mosquito Aedes aegypti. Clim Change. 2018; 146: 487–500. https://doi.org/10.1007/s10584-016-1679-0 PMID: 29610543

157. Reis BY, Kohane IS, Mandl KD. An epidemiological network model for disease outbreak detection. PLoS Med. 2007; 4: e210. https://doi.org/10.1371/journal.pmed.0040210 PMID: 17593895

158. Thacker SB, Qualters JR, Lee LM, Centers for Disease Control and Prevention. Public health surveillance in the United States: evolution and challenges. MMWR Suppl. 2012; 61: 3–9.

159. Romette JL, Prat CM, Gould EA, de Lamballerie X, Charrel R, Coutard B, et al. The European Virus Archive goes global: A growing resource for research. Antiviral Res. 2018; 158: 127–134. https://doi.org/10.1016/j.antiviral.2018.07.017 PMID: 30059721

160. Dos Reis IC, Gibson G, Ayllón T, de Medeiros Tavares A, de Araújo JMG, da Silva Monteiro E, et al. Entomo-virological surveillance strategy for dengue, Zika and chikungunya arboviruses in field-caught Aedes mosquitoes in an endemic urban area of the Northeast of Brazil. Acta Trop. 2019; 197: 105061. https://doi.org/10.1016/j.actatropica.2019.105061 PMID: 31194961

**161.** Jones R, Kulkarni MA, Davidson TMV, Team R-LR, Talbot B. Arbovirus vectors of epidemiological concern in the Americas: A scoping review of entomological studies on Zika, dengue and chikungunya virus vectors. PLOS ONE. 2020; 15: e0220753. https://doi.org/10.1371/journal.pone.0220753 PMID: 32027652

**162.** Ajin VW, Kumar LD. Big data and clustering algorithms. 2016 International Conference on Research Advances in Integrated Navigation Systems (RAINS). 2016. pp. 1–5.

**163.** Dave M, Gianey H. Different clustering algorithms for Big Data analytics: A review. 2016 International Conference System Modeling Advancement in Research Trends (SMART). 2016. pp. 328–333.

**164.** Pearl J. Causality: models, reasoning, and inference. Cambridge University Press; 2009.

**165.** Manning CD, Raghavan P, Schütze H. Introduction to Information Retrieval. 2009; 569.

# Chapitre 2 : Pertinence des sources de données de vie réelle dans la surveillance de la dengue

Le chapitre précédent nous a permis d'identifier les sources de données potentiellement pertinentes dans la surveillance de la dengue. En particulier, il nous a permis d'identifier la sous-identification des données vectorielles dans ce genre d'études et l'intérêt des données issues de moteurs de recherche tels que *Google*.

Cependant, cette revue systématisée nous a aussi montré que la plupart des études évaluant l'apport des données de vie réelle dans la dengue ont été réalisées en Asie, et quasiment aucune d'entre elles n'avait été réalisée dans la Caraïbe. Une grande partie des pays et territoires de la région sont des îles et sont donc beaucoup plus petits que les régions ayant déjà évalué ces méthodes. Or, l'utilisation des données de *Google* demande un volume minimal afin d'être exploitables.

Nous souhaitions donc évaluer la pertinence de ces sources, et en particulier pour *Google,* les éventuelles adaptations méthodologiques à mettre en place pour un petit territoire tel que la Martinique ou les îles voisines. Nous souhaitions également évaluer leur capacité potentielle à anticiper les épidémies, en réduisant les délais de notifications de cas.

Cette étude a montré que les données du PMSI, les interventions entomologiques hebdomadaires de l'équipe de démoustication et les recherches de mots clés associés à la dengue dans *Google* étaient corrélées à l'augmentation du taux de positivité des PCR dengue et précédaient cette augmentation entre 2 et 5 semaines selon les sources.

**Synthèse**

Les données vectorielles sont sous-utilisées dans les études de surveillance syndromique de la dengue. Par ailleurs, la plupart des études existantes sont réalisées dans des grands territoires avec des volumes de données importants.

C'est pour cela que nous avons évalué la pertinence de différentes sources de données de vie réelle pour la surveillance de la dengue (y compris les données de surveillance vectorielle) dans un petit territoire tel que la Martinique.

Ce travail a fait l'objet d'un article soumis à JMIR Public Health and Surveillance.

## Article 4 : Combining heterogenous real-world data sources to monitor dengue fever in Martinique

Ma contribution à cette étude a été d'intégrer les données, selon un modèle de données commun, de définir la méthodologie d'évaluation de ces données et de les analyser.

## Original Paper

Emmanuelle Sylvestre[12*], MD, MSc; Elsa Cécilia-Joseph[2], PhD; Guillaume Bouzillé[1], MD, PhD; Fatiha Najioullah[3], PharmD, PhD; Manuel Etienne[4], PhD; Fabrice Malouines[4], Jacques Rosine[5], Sandrine Julié[6], MD; André Cabié[789], MD, PhD; Marc Cuggia[1], MD, PhD

[1] Université de Rennes, CHU Rennes, INSERM, LTSI – UMR 1099, F-35000, Rennes, France;

[2]CHU Martinique, Centre de Données Cliniques, F-97200, Martinique, France;

[3]CHU Martinique, Laboratoire de Virologie, F-97200, Martinique, France;

[4]Collectivité Territoriale de la Martinique – Agence Régionale de Santé, Centre de Démoustication et de Recherche en Entomologie, F-97200, Martinique, France.

[5]Santé Publique France, Cellule Martinique, Saint-Maurice, France

[6]CHU Martinique, Service de Santé Publique – Département d'Information Médicale, F-97200, Martinique, France;

[7]CHU Martinique, Infectious and Tropical Diseases Unit, F-97200, Martinique, France;

[8]CHU Martinique, INSERM, CIC-1424, F-97200, Martinique, France

[9]PCCEI, Université de Montpellier, INSERM, EFS, Université Antilles, Montpellier, France

* Corresponding author

E-mail: emmanuelle.sylvestre@chu-martinique.fr

# Combining heterogenous real-world data sources to monitor dengue fever in Martinique


## Abstract

**Background:** Traditionally, dengue prevention and control rely on vector control programs and reporting of symptomatic cases to a central health agency. However, case reporting is often delayed, and the true burden of dengue disease is often underestimated. Moreover, some countries do not have routine control measures in place for vector control. Therefore, researchers are constantly assessing novel sources of data to improve traditional surveillance systems. These studies are mostly carried out in big territories, and rarely in smaller endemic regions, such as Martinique and the Lesser Antilles.

**Objective:** The aim of this study was to determine whether heterogenous real-world data sources could help to reduce reporting delays and improve dengue monitoring in Martinique island, a small endemic region.

**Methods:** Heterogenous data sources (hospitalization data, entomological data, and Google Trends) and dengue surveillance reports for the last 14 years (January 2007 to February 2021) were analyzed to identify associations with dengue outbreaks and their time lags.

**Results:** Dengue hospitalization rate was the variable most strongly correlated with the increase in dengue positivity rate by RT-PCR (Pearson's correlation coefficient = 0.70) with a time lag of -3 weeks. Weekly entomological interventions also were correlated with the increase in dengue positivity rate by RT-PCR (Pearson's correlation coefficient = 0.59) with a time lag of -2 weeks. The most correlated query from Google Trends was the "Dengue" topic restricted to the Martinique region (Pearson's correlation coefficient = 0.637) with a time lag of -3 weeks.

**Conclusions:** Real-word data are valuable data sources for dengue surveillance in smaller territories. Many of these sources precede the increase of dengue cases of several weeks, and therefore can help to improve the ability of traditional surveillance systems to provide an early response in dengue outbreaks. All these sources should be better integrated to improve the early response to dengue outbreaks, and vector-borne diseases in general, in smaller endemic territories.



**Keywords:** dengue; surveillance; real-word data; *Big Data*; Caribbean; dengue-endemic region

# Introduction

Dengue is one of the most important vector-borne diseases in the world with 390 million infections, 96 million symptomatic cases, and 20,000 estimated deaths per year in more than 125 countries [1,2]. The disease is mostly endemic in tropical and sub-tropical regions (i.e., South-East Asia, the Americas, and the Pacific) with 4 billion people at risk [3]. In Latin America and the Caribbean, morbidity and mortality have increased from 400,519 cases and 92 deaths in 2000 to more than 3,1 million cases and 1,534 deaths in 2019 [4,5]. Dengue prevention and control in these regions rely on two main approaches: vector control programs and traditional surveillance, based on passive detection of symptomatic cases (inpatient and outpatient) [4,6]. Although both approaches are effective, they are expensive and suffer from delays between case occurrence and case reporting. Furthermore, some countries do not have routine control measures for vector control [7], and national epidemiological surveillance systems tend to underestimate the true burden of dengue disease [8].

In Martinique, a French overseas territory in the Lesser Antilles, health authorities have launched the "Monitoring, warning and management of dengue outbreaks program" (Programme de surveillance, d'alerte et de gestion des épidémies de dengue, PSAGE) in which vector control and traditional surveillance are combined. PSAGE identifies five main stages in dengue outbreaks: i) sporadic transmission, ii) dengue clusters with or without an epidemiological link, iii) epidemic risk when the number of symptomatic cases is above the expected threshold, iv) dengue outbreak, and v) back to normal. Vector surveillance still plays a role in this system; however, the change of PSAGE stage is mainly based on the number of symptomatic cases identified by general practitioners who are part of the French Sentinel Network surveillance system [9,10].

Surveillance systems are a key public health tool to detect early cases of emerging infectious diseases, to prevent outbreaks [11] among populations, and to put in place measures to reduce transmission [12]. Traditional surveillance systems are often expensive due to the time and resources required to process the data collected from public health networks [13]. To improve these systems and reduce the delay between diagnosis and reporting, researchers have been evaluating novel data sources, especially real-world data (i.e., data not collected in experimental conditions [14]) such as emergency department

visits, mobile data, or Internet-based systems [15–18]. So far, most of these studies have been carried out in large territories or countries, such as Brazil or Mexico [19,20].

The aim of this study carried out in Martinique was to investigate whether heterogenous real-world data sources could help to reduce reporting delays and improve dengue monitoring in a smaller endemic region.

# Methods

## Data sources

We used several sources of data that were routinely collected during the study period (from January 1, 2007, to February 28, 2021): epidemiological surveillance reports from the French national Public Health Agency (Santé Publique France), reimbursement claims, laboratory data from Martinique University Hospital, entomological data from the Martinique Mosquito Control and Entomology Research Center (Centre de Démoustication et de Recherche en Entomologie, CEDRE), and Relative Search Volumes (RSV) from Google Trends. All used data were anonymous.

### *Epidemiological surveillance data*

We obtained weekly dengue surveillance reports from the French Public Health Agency. These reports are based on data collected by general practitioners from the French Sentinel Network. They also provide the official start and end dates of each dengue outbreak, and the weekly PSAGE stage during the outbreak. These reports are not continuously published, but only if the dengue risk level is above stage 1 (i.e., the baseline stage).

### *Clinical and laboratory data*

We obtained i) inpatient data (age and diagnoses associated with dengue disease or dengue symptoms, see below), ii) administrative data (outpatient medical consultations, hospitalizations, and Emergency department visits), and iii) laboratory data: dengue virus (DENV) detection by reverse transcription-real time polymerase chain reaction (RT-PCR).

All included diagnoses were coded using the French version of the International Classification of Diseases, 10th edition (ICD-10), and were:

- Dengue or severe dengue
- Possible coding errors associated with dengue: fever, unspecified viral hemorrhagic fever
- Severity symptoms: hemorrhage, shock, dehydration
- Thrombocytopenia
- Hepatic symptoms: hepatitis, hepatomegaly, hepatic failure, elevation of transaminase
- Neurologic symptoms: encephalitis, encephalopathy

We selected these diagnoses with the help of infectious disease physicians. The relevant ICD-10 codes are listed in Multimedia Appendix 1.

For the laboratory data, we used the DENV positivity rate by RT-PCR. Laboratory results concerned both inpatients and outpatients because the Martinique University Hospital is the reference center for DENV screening by RT-PCR in Martinique.

### *Entomological data*

We used data from the CEDRE surveillance database, such as the weekly number of entomologic interventions and where they were carried out. This agency manages entomological surveillance and vector-control in Martinique and collects data on each of their intervention.

### *Google relative search volumes*

We used data from Google Trends [21] that provides real-time and archived information on Google queries from 2004 onwards. These queries are normalized by Google as RSV by dividing the total search volume for a query in a geographical location by the total number of queries in that region at a given point in time [22]. We used this tool to retrieve information on online interest for keywords associated with dengue during our selected time frame (January 2007 to February 2021). We based our methodology to retrieve Google Trends data on previously published methodology frameworks indicating that

Google Trends data should be retrieved for exactly the same period as the other data under study, and as a single dataset rather than as individual queries for each year [23]. As data for our study period were only available in monthly intervals, we considered that the interest was constant over each week of the month for each query.

For data retrieval, we selected the relevant keywords with experts in the field. Normally, all spelling variations should be included in the research to limit the risk of missing data. However, in our case, combining all possible spelling variants of some keywords in a single query was impossible and an error message from Google indicated that the available data were not enough. Nevertheless, we could retrieve results using the "topic" option from Google that includes various keywords associated with a category.

As Martinique (and the other islands in the Lesser Antilles) are small regions, we tried two strategies to explore the geographic region of our keywords: we selected "Martinique" as the region in the tool, and we added "Martinique" as a keyword in our query, with the region selected as "worldwide". Moreover, we selected our keywords in three different languages (French, English, and Spanish) because the Lesser Antilles are a multilingual region.

## Data processing

Most of the information in the CEDRE database was unstructured. Therefore, we used rule-based natural language processing methods to process the data and extract the relevant information for our study, especially to address standardization.

## Statistical analysis

Between 2007 and 2021, four dengue outbreaks were recorded in Martinique: from August 20, 2007 to January 14, 2008; from February 22, 2010 to October 25, 2010; from July 22, 2013 to April 14, 2014, and from November 18, 2019 to February 8, 2021. The fourth dengue outbreak was the biggest in Martinique in the last 20 years.

During the same period there was one chikungunya outbreak in 2014, one Zika virus outbreak in 2016, the first coronavirus disease 2019 (COVID-19) wave in March 2020, and

the second COVID-19 wave from September to December 2020. The last dengue outbreak was concomitant with the second COVID-19 wave. Therefore, the PSAGE stages did not vary much over the years, which makes difficult to study correlations with this categorical variable. Therefore, we needed to find a good continuous estimator for times-series analyses. To this aim, we assessed the DENV RT-PCR positive rate performance for PSAGE stage prediction. First, we divided the RT-PCR data between a training set and a test set, and then evaluated the performance of a logistic regression model to predict the PSAGE stages. The metrics used were accuracy, specificity, precision, recall, F-Score, Area Under Curve (AUC), and Receiver Operating Characteristic (ROC) curve. Then, to investigate the association between RT-PCR positive rate and each data source we plotted their time series. Lastly, for each source, we estimated the Pearson correlation coefficient (r) and the cross-correlations between the weekly data and the DENV RT-PCR positive rate. We performed all statistical analyses with R version 4.1.0 [24].

# Results

## RT-PCR positive rate performance

The performance of the DENV RT-PCR positive rate to predict the PSAGE stage was good, particularly for predicting the sporadic transmission (stage 1) and the outbreak stages (stage 4) (see metrics in Table 1).

**Table 1. DENV RT-PCR positive rate and PSAGE stage prediction**

| Metrics | PSAGE stage 1 or 5 | PSAGE stage 2 | PSAGE stage 3 | PSAGE stage 4 |
|---|---|---|---|---|
| Accuracy | 0.787 | 0.878 | 0.953 | 0.867 |
| Specificity | 0.509 | 1 | 1 | 0.944 |
| Precision | 0.786 | NA[a] | NA[a] | 0.65 |
| Recall | 0.929 | 0 | 0 | 0.50 |
| F-score | 0.852 | NA[a] | NA[a] | 0.565 |
| AUC[b] | 0.787 | 0.594 | 0.791 | 0.889 |

[a] Not enough data available to build a prediction model for these stages
[b] Area Under the Curve

## Hospital data

We normalized all hospital data to plot the time series in order to take into account the different scales. As children and adults can be affected differently depending on the dengue infection type (primary versus secondary), we stratified our datasets based on the ward type (adult or pediatric).

### *Administrative data*

We normalized all administrative data as follows:

$$y = \frac{(x - \min(x))}{(\max(x) - \min(x))}$$

where $x$ is the weekly number of hospitalizations/consultations/emergency visits, min the minimum and max the maximum values observed in the dataset.

Adult hospitalizations and emergency department visits were significantly associated with the DENV RT-PCR positive rate. We also observed a significant cross-correlation, at -3 and -5 weeks, indicating that the increase in emergency department visits preceded the increase in DENV RT-PCR positive rate by 3 to 5 weeks. Table 2 shows the correlations and cross-correlations between administrative data and DENV detection rate by RT-PCR.

**Table 2. Correlations and cross-correlations between administrative data and DENV RT-PCR positive rate**

| Data | Correlation | P-value | Confidence interval | Max cross-correlation[a] | Time lag[b] |
|---|---|---|---|---|---|
| **Hospitalizations** | | | | | |
| Total | -0.066 | .07 | [-0.137;0.006] | -0.091 | -5 weeks |
| Adults | -0.095 | **.0099** | [-0.165; -0.023] | -0.097 | -2 weeks |
| Children | 0.067 | .06 | [-0.004;0.139] | 0.118 | -8 weeks |
| **Emergency department visits** | | | | | |
| Total | 0.111 | **.0024** | [0.039;0.181] | **0.169** | **-5 weeks** |
| Adults | 0.181 | ***P*<.001** | [0.11;0.25] | **0.216** | **-3 weeks** |
| Children | 0.046 | .21 | [-0.025;0.118] | **0.107** | **-5 weeks** |
| **Consultations** | | | | | |

| | | | | | |
|---|---|---|---|---|---|
| Total | -0.065 | .08 | [-0.137;0.007] | -0.067 | -2 weeks |
| Adults | -0.061 | .09 | [-0.133;0.0105] | -0.097 | -5 weeks |
| Children | -0.046 | .21 | [-0.118;0.026] | -0.087 | -5 weeks |

[a] Maximum cross-correlation
[b] Time lag that results in the maximum cross-correlation

### *Inpatient data*

We normalized the inpatient data as the percentage of each diagnosis among all diagnoses for that year. The percentage of dengue diagnoses among inpatients was significantly associated with the increase in DENV RT-PCR positive rate. We also detected a significant cross-correlation, at the time lag of -3 weeks, indicating that the increase in dengue diagnoses among hospitalized people preceded the increase in DENV RT-PCR positive rates by 3 weeks (Table 3).

**Table 3. Correlations between dengue diagnoses inpatients and DENV RT-PCR positive rate**

| Data | Correlation | P-value | Confidence interval | Max cross-correlation[a] | Time lag[b] |
|---|---|---|---|---|---|
| **Percentage of dengue diagnoses** | | | | | |
| Total | 0.704 | ***P*<.001** | [0.665;0.738] | **0.710** | **-3 weeks** |
| Adults | 0.698 | ***P*<.001** | [0.659; 0.733] | **0.703** | **-3 weeks** |
| Children | 0.672 | ***P*<.001** | [0.631;0.701] | **0.675** | **-3 weeks** |

[a] Maximum cross-correlation
[b] Time lag that results in the maximum cross-correlation

Concerning dengue-related symptoms, thrombocytopenia and liver involvement in adults and in children were associated with DENV RT-PCR positive rate. The significant cross-correlation, at time lags ranging between -2 and -5 weeks, indicated that the increase in thrombocytopenia and liver dysfunction signs preceded the increase in DENV RT-PCR positive rates by 3 to 5 weeks (Table 4).

**Table 4. Correlations between dengue symptoms among inpatients and dengue RT-PCR positive rate**

| Data | Correlation | P-value | Confidence interval | Max cross-correlation[a] | Time lag[b] |
|---|---|---|---|---|---|
| **Symptoms** | | | | | |
| Total | 0.077 | **.035** | [0.005;0.148] | 0.081 | -4 weeks |
| Adults | 0.071 | .054 | [-0.001;0.142] | 0.071 | 0 weeks |
| Children | 0.093 | **.011** | [0.021;0.16] | 0.127 | -4 weeks |
| **Coding errors** | | | | | |
| Total | -0.096 | **.009** | [-0.167;-0.024] | -0.098 | -1 week |
| Adults | -0.072 | **.050** | [-0.143;-1.12 x $10^{-4}$] | -0.086 | -2 weeks |
| Children | -0.043 | .236 | [-0.115;0.0285] | -0.043 | 0 weeks |
| **Symptom severity** | | | | | |
| Total | 0.105 | **.004** | [0.033;0.175] | 0.105 | 0 weeks |
| Adults | 0.068 | .065 | [-0.004;0.139] | 0.068 | 0 weeks |
| Children | 0.263 | ***P*<.001** | [0.195;0.329] | **0.279** | **-4 weeks** |
| **Thrombocytopenia** | | | | | |
| Total | 0.281 | ***P*<.001** | [0.213;0.346] | **0.289** | **-2 weeks** |
| Adults | 0.235 | ***P*<.001** | [0.166;0.302] | **0.242** | **-2 weeks** |
| Children | 0.269 | ***P*<.001** | [0.201;0.335] | **0.288** | **-4 weeks** |
| **Liver dysfunction symptoms** | | | | | |
| Total | 0.152 | ***P*<.001** | [0.081;0.222] | **0.179** | **-5 weeks** |
| Adults | 0.123 | ***P*<.001** | [0.0517;0.193] | **0.153** | **-5 weeks** |
| Children | 0.152 | ***P*<.001** | [0.081;0.222] | **0.147** | **-5 weeks** |
| **Neurologic symptoms** | | | | | |
| Total | -0.028 | .438 | [-0.100;0.0435] | -0.061 | -7 weeks |
| Adults | -0.045 | .216 | [-0.117;0.0265] | -0.068 | -7 weeks |
| Children | 0.029 | .426 | [-0.0427;0.101] | 0.034 | -6 weeks |

[a] Maximum cross-correlation

[b] Time lag that results in the maximum cross-correlation

The weekly hospitalization rates for dengue and thrombocytopenia during the study period are shown in Figure 1.

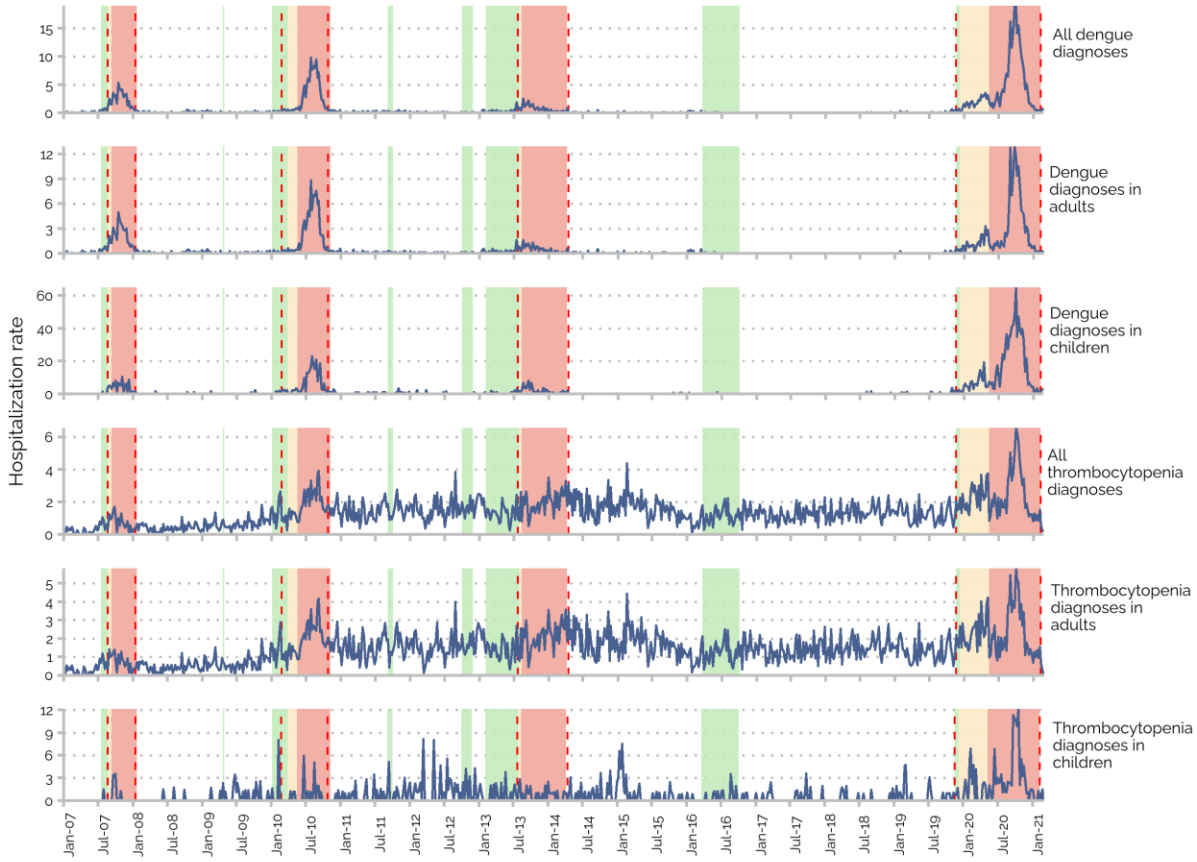



**Figure 1. Weekly hospitalization rates for dengue and thrombocytopenia during the different PSAGE stages from January 2007 to February 2021.**
Blue curves: weekly hospitalization rates for the indicated ICD-10 diagnoses. Green areas: PSAGE stage 2 (i.e. dengue clusters). Yellow areas: PSAGE stage 3 (epidemic risk). Red areas: PSAGE stage 4 (dengue outbreak). Red dashed lines: official dates of the dengue outbreaks that are decided retrospectively by the French Public Health Agency at the end of each outbreak.

## Entomological data

The weekly number of entomological interventions was significantly ($p=4.12 \times 10^{-7}$) associated with DENV RT-PCR positive rate (r=0.591; 95% CI, 0.542-0.636). They were also significantly cross-correlated (0.627 at -2 weeks), indicating that their increase preceded the increase in DENV RT-PCR positive rate by 2 weeks.

We did not find any significant correlation or cross-correlation between the intervention zones and the RT-PCR positive rate.

## Google Relative Search Volumes

Several Google keywords were significantly associated with the DENV RT-PCR positive rate. Overall, this association was stronger for the simplest combination of keywords, without spelling variations, especially for the keywords "dengue symptoms". We could not assess some keyword combinations because of the lack of data. Furthermore, when Google Trends provided "Topics", the results outperformed those obtained using manual combinations of keywords that included spelling, language, or accent variations. Keywords not restricted to the geographical region of "Martinique" (by using the Geographical region feature or by adding the keyword "Martinique" to the query) were not significantly associated with the DENV RT-PCR positive rate. We obtained the strongest significant cross-correlation using the topic "dengue" in the Martinique region (0.643 at the time lag of -3 weeks). This indicated that that an increase of queries for this term in the Martinique region preceded the increase in DENV RT-PCR positive rate by 3 weeks (Table 5). Conversely, we did not find any significant cross-correlation within meaningful time lag values for the term "mosquito" and its different spelling and language variations. All correlations between Google Trends keywords and DENV RT-PCR positive rates are listed in Multimedia Appendix 2.

**Table 5. Strongest correlations between Google Trends keywords and DENV RT-PCR positive rate**

| Keywords | Correlation | P-value | Confidence interval | Max cross-correlation[a] | Time lag[b] |
|---|---|---|---|---|---|
| **Dengue** | | | | | |
| Keywords "dengue" + dingue" and region: Martinique | 0.597 | ***P*<.001** | [0.548;0.641] | **0.598** | **- 1 week** |
| Keywords "dengue" + "martinique" | 0.534 | ***P*<.001** | [0.480;0.583] | **0.611** | **- 6 weeks** |
| Dengue Topic and region Martinique | 0.637 | ***P*<.001** | [0.591;0.677] | **0.643** | **- 3 weeks** |
| **Dengue symptoms** | | | | | |
| Keyword "symptome dengue" and region Martinique | 0.412 | ***P*<.001** | [0.351;0.47] | **0.435** | **-3 weeks** |

| | | | | | |
|---|---|---|---|---|---|
| **Mosquito** | | | | | |
| Keyword "mosquito" with various French spellings and region: Martinique | 0.200 | ***P*<.001** | [0.130;0.268] | 0.200 | 0 weeks |
| **Aedes** | | | | | |
| Keywords "aedes" and region Martinique | 0.339 | ***P*<.001** | [0.273;0.401] | **0.369** | **-3 weeks** |
| Aedes topic and region Martinique | 0.214 | ***P*<.001** | [0.591;0.677] | **0.304** | **-7 weeks** |

[a] Maximum cross-correlation

[b] Time lag that results in the maximum cross-correlation

The weekly estimates for non-hospital data during the study period are displayed in Figure 2.

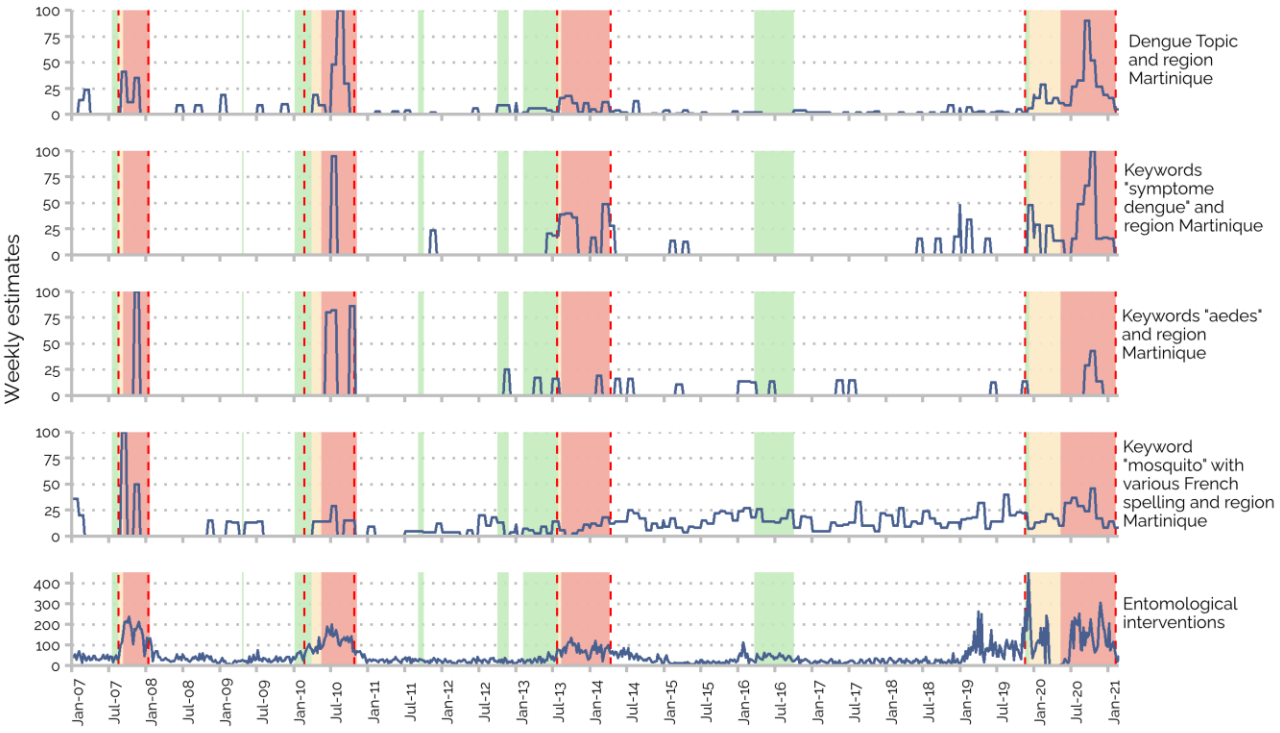



**Figure 2. Weekly estimates for the indicated non-hospital data during the different PSAGE stages from January 2007 to February 2021.**

Blue curves: weekly estimates for the strongest correlated Google keywords and for entomological interventions. Green areas: PSAGE stage 2 (dengue clusters). Yellow areas: PSAGE stage 3 (epidemic risk). Red areas: PSAGE stage 4 (dengue outbreak). Red dashed lines: official dates of the outbreaks that are decided retrospectively by the French Public Health Agency at the end of each outbreak.

# Discussion

## Principal results

This study demonstrates the great potential of real-world data for dengue outbreak monitoring. It indicates that multiple heterogenous data sources, such as clinical data, vector data and novel Big Data streams, should be leveraged simultaneously because they all can play a role in improving traditional dengue surveillance systems. Moreover, some data, such as the weekly hospitalization rates for thrombocytopenia, the weekly number of entomological interventions and Google keywords, were not only significantly correlated with the weekly DENV RT-PCR positive rates, but their increase preceded the increase in RT-PCR positive results by 2 to 4 weeks.

Previous studies have already investigated the role of entomological data[25], inpatient data[26], and Internet data streams [27] in dengue management, but few have assessed all these data sources simultaneously. Here, we found that they should all be considered together, rather than individually. Vector-based data tend to be underused [28], despite their central place in dengue surveillance, although we observed a rather strong correlation between the number of weekly entomological interventions and the increase in DENV RT-PCR positive rates. Therefore, they should be better integrated in the dengue surveillance system to improve its efficiency because both clinical surveillance and vector-based surveillance are essential for optimal dengue management [29]. The role of Internet search engines in dengue surveillance has been frequently addressed in the last few years [30,31], but most studies were carried out in Asia and in bigger American countries, such as Mexico and Brazil [19,20]. Here, we found that even in Martinique, a smaller territory with a smaller population and thus, with a lower data volume from Internet data streams, Google queries were still correlated with the increase of DENV RT-PCR positive rates. This means that they also can be used as part of the surveillance systems across the islands of the Lesser Antilles. However, the methodology framework [23] still needs to be adapted to the size of these territories, and the simplest keywords and Google topics, when available, should be preferred rather than multiple spelling variations. With these small adaptations, we propose a way to offset the limitations related to smaller territories in order to use Internet data streams also in this context because their interest in emerging disease surveillance has been demonstrated in previous studies [32,33]. Finally, dengue

hospitalizations and the symptoms associated with severe dengue cases (thrombocytopenia and liver dysfunction symptoms) should be closely monitored in inpatients, especially in children, because they tend to precede the increase of DENV RT-PCR positive rate by several weeks.

Our study also highlighted homogenous time lags across the different data sources, despite their heterogeneity. This further demonstrates the importance of considering them globally rather than individually, although some of these correlations were moderate. The capacity to identify variables that precede the increase of DENV RT-PCR positive rate is very relevant in dengue management because a rapid and early response can influence the outbreak severity [18].

## Limitations

Our method is promising but has some limitations. First, we did not include climate data in our study because not enough data were available for our time frame. Several studies demonstrated the role of climate data (especially specific variables, such as temperature, humidity, and rainfall) for dengue surveillance, but they were mostly carried out in Asian countries [34,35] or South America [36,37]. The few studies in the Caribbean region showed the role of rainfall and temperature in increasing the risk of dengue outbreaks, but with time lags varying between 7 weeks and 5 months [38,39], which is longer that the time lags we found for the other data sources. Nevertheless, this data source could have been relevant.

Second, our laboratory data did not include private-sector biology laboratories because they did not use RT-PCR techniques before the COVID-19 pandemic in 2020. Before this date, dengue diagnosis in private-sector laboratories was based on NS1 antigen detection and needed sometimes to be confirmed by the more sensitive RT-PCR test at the hospital laboratory. It should be noted that the weekly number of DENV RT-PCR tests increased over time. Therefore, we used the weekly positive rate and not the weekly number of RT-PCR tests. Likewise, the WHO dengue case classification and the guidelines for hospitalization changed during the study period [40], and this may have influenced the results. Nevertheless, the rate of hospitalized patients with a dengue diagnosis was the variable more strongly correlated with the DENV RT-PCR positive rate in our study.

Third, concerning the entomological data, we only studied the correlation between the weekly number of interventions and the increase of DENV RT-PCR positive rate, but we did take into account the number of mosquito clusters (i.e. several clusters can be detected during one intervention). We focused on the simplest variable because vector control programs vary among the countries of this region [4,6], and we wanted to develop a common approach for all Caribbean territories. Furthermore, as entomological interventions tend to increase during outbreaks, we cannot rule out the influence of these practices on our results. Nevertheless, we could show that entomological interventions precede the increase of DENV RT-PCR positive rate by 2 weeks.

Our approach does not intend to replace the traditional monitoring systems based on syndromic surveillance, but to reduce the delays in these systems by leveraging data that are already routinely collected. These new data sources are readily available and can be easily implemented in the already existing surveillance systems with minimal costs and training. However, their ability to predict future dengue outbreaks need to be thoroughly assessed, especially in smaller territories in the Lesser Antilles.

## Conclusions

Our study shows that real-word data are valuable data sources for dengue surveillance in Martinique. Several heterogenous data sources are relevant, from clinical data to vector control data and Google Trends data. Their increase precedes the increase of dengue cases of several weeks, and therefore, they can help to improve traditional surveillance systems in order to provide an early response to dengue outbreaks. By improving the integration of many different sources, we might better respond to dengue outbreaks in endemic regions, and also to other types of vector-borne diseases, such as Zika and chikungunya.

## Conflicts of Interest

None declared

## Abbreviations

AUC: Area Under Curve

CEDRE: Centre de Démoustication et de Recherche en Entomologie, CEDRE

COVID-19: Coronavirus disease 2019

DENV: Dengue virus

ICD-10: International Classification of Diseases, 10th edition

PCC: Pearson's Correlation Coefficient

PSAGE: Programme de surveillance, d'alerte et de gestion des épidémies de dengue

ROC Curve: Receiver Operating Characteristic Curve

RSV: Relative Search Volumes

RT-PCR: Real-Time reverse transcriptase Polymerase Chain Reaction

## Multimedia Appendix 1

ICD-10 codes of the selected diagnoses for inpatient data.

## Multimedia Appendix 2

Correlations between Google Trends keywords and DENV RT-PCR positive rate

## References


1. Bhatt S, Gething PW, Brady OJ, Messina JP, Farlow AW, Moyes CL, Drake JM, Brownstein JS, Hoen AG, Sankoh O, Myers MF, George DB, Jaenisch T, Wint GRW, Simmons CP, Scott TW, Farrar JJ, Hay SI. The global distribution and burden of dengue. Nature 2013 Apr 25;496(7446):504–507. PMID:23563266

2. Stanaway JD, Shepard DS, Undurraga EA, Halasa YA, Coffeng LE, Brady OJ, Hay SI, Bedi N, Bensenor IM, Castañeda-Orjuela CA, Chuang T-W, Gibney KB, Memish ZA, Rafay A, Ukwaja KN, Yonemoto N, Murray CJL. The Global Burden of Dengue: an analysis from the Global Burden of Disease Study 2013. Lancet Infect Dis 2016 Jun;16(6):712–723. PMID:26874619

3. Brady OJ, Gething PW, Bhatt S, Messina JP, Brownstein JS, Hoen AG, Moyes CL, Farlow AW, Scott TW, Hay SI. Refining the global spatial limits of dengue virus transmission by evidence-based consensus. PLoS Negl Trop Dis 2012;6(8):e1760. PMID:22880140

4. Torres JR, Orduna TA, Piña-Pozas M, Vázquez-Vega D, Sarti E. Epidemiological Characteristics of Dengue Disease in Latin America and in the Caribbean: A Systematic Review of the Literature. J Trop Med 2017;2017:8045435. PMID:28392806

5. Dengue - PAHO/WHO | Pan American Health Organization [Internet]. [cited 2022 Feb 7]. Available from: https://www.paho.org/en/topics/dengue

6. Cafferata ML, Bardach A, Rey-Ares L, Alcaraz A, Cormick G, Gibbons L, Romano M, Cesaroni S, Ruvinsky S. Dengue Epidemiology and Burden of Disease in Latin America and the Caribbean: A Systematic Review of the Literature and Meta-Analysis. Value Health Reg Issues 2013 Dec;2(3):347–356. PMID:29702769

7. H G-D, Jr W. Dengue in the Americas: challenges for prevention and control. Cadernos de saude publica [Internet] Cad Saude Publica; 2009 [cited 2022 Jan 25];25 Suppl 1. PMID:19287863

8. Shepard DS, Undurraga EA, Betancourt-Cravioto M, Guzmán MG, Halstead SB, Harris E, Mudin RN, Murray KO, Tapia-Conyer R, Gubler DJ. Approaches to refining estimates of global burden and economics of dengue. PLoS Negl Trop Dis 2014 Nov;8(11):e3306. PMID:25412506

9. Dussart DOF, Gustave10 J, Lagathu G, Koulman11 L, Lordinot ML, Martial J, Nadeau Y, Quatresous12 I, Quenel P, Rosine J. PROGRAMME DE SURVEILLANCE, D'ALERTE ET DE GESTION DES EPIDEMIES DE DENGUE (PSAGE DENGUE) EN MARTINIQUE.

10. Flahault A, Blanchon T, Dorléans Y, Toubiana L, Vibert JF, Valleron AJ. Virtual surveillance of communicable diseases: a 20-year experience in France. Stat Methods Med Res 2006 Oct;15(5):413–421. PMID:17089946

11. Milinovich GJ, Williams GM, Clements ACA, Hu W. Internet-based surveillance systems for monitoring emerging infectious diseases. Lancet Infect Dis 2014 Feb;14(2):160–168. PMID:24290841

12. Jajosky RA, Groseclose SL. Evaluation of reporting timeliness of public health surveillance systems for infectious diseases. BMC Public Health 2004 Jul 26;4:29. PMID:15274746

13. Wilson K, Brownstein JS. Early detection of disease outbreaks using the Internet. CMAJ 2009 Apr 14;180(8):829–831. PMID:19364791

14. Sherman RE, Anderson SA, Dal Pan GJ, Gray GW, Gross T, Hunter NL, LaVange L, Marinac-Dabic D, Marks PW, Robb MA, Shuren J, Temple R, Woodcock J, Yue LQ, Califf RM. Real-World Evidence - What Is It and What Can It Tell Us? N Engl J Med 2016 Dec 8;375(23):2293–2297. PMID:27959688

15. Elliot AJ, Hughes HE, Hughes TC, Locker TE, Shannon T, Heyworth J, Wapling A, Catchpole M, Ibbotson S, McCloskey B, Smith GE. Establishing an emergency department syndromic surveillance system to support the London 2012 Olympic and Paralympic Games. Emerg Med J 2012 Dec;29(12):954–960. PMID:22366039

16. 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. New York City syndromic surveillance systems. MMWR Suppl 2004 Sep 24;53:23–27. PMID:15714622

17. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature 2009 Feb 19;457(7232):1012–1014. PMID:19020500

18. Katzelnick LC, Coloma J, Harris E. Dengue: knowledge gaps, unmet needs, and research priorities. Lancet Infect Dis 2017;17(3):e88–e100. PMID:28185868

19. Gluskin RT, Johansson MA, Santillana M, Brownstein JS. Evaluation of Internet-based dengue query data: Google Dengue Trends. PLoS Negl Trop Dis 2014 Feb;8(2):e2713. PMID:24587465

20. Romero-Alvarez D, Parikh N, Osthus D, Martinez K, Generous N, Del Valle S, Manore CA. Google Health Trends performance reflecting dengue incidence for the Brazilian states. BMC Infect Dis 2020 Mar 26;20(1):252. PMID:32228508

21. Google Trends [Internet]. Google Trends. [cited 2022 Jan 25]. Available from: https://trends.google.com

22. Hswen Y, Zhang A, Ventelou B. Estimation of Asthma Symptom Onset Using Internet Search Queries: Lag-Time Series Analysis. JMIR Public Health and Surveillance 2021 May 10;7(5):e18593. [doi: 10.2196/18593]

23. Mavragani A, Ochoa G. Google Trends in Infodemiology and Infoveillance: Methodology Framework. JMIR Public Health and Surveillance 2019 May 29;5(2):e13439. [doi: 10.2196/13439]

24. R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. [Internet]. Available from: http://www.r-project.org/index.html

25. Marques-Toledo CA, Bendati MM, Codeço CT, Teixeira MM. Probability of dengue transmission and propagation in a non-endemic temperate area: conceptual model and decision risk levels for early alert, prevention and control. Parasit Vectors 2019 Jan 16;12(1):38. PMID:30651125

26. Mello-Román JD, Mello-Román JC, Gómez-Guerrero S, García-Torres M. Predictive Models for the Medical Diagnosis of Dengue: A Case Study in Paraguay. Comput Math Methods Med 2019;2019:7307803. PMID:31485259

27. Klein GH, Neto PG, Tezza R. Big Data and social media: surveillance of networks as management tool. Saude Soc 2017 Mar;26(1):208–217. [doi: 10.1590/S0104-12902017164943]

28. Sylvestre E, Joachim C, Cécilia-Joseph E, Bouzillé G, Campillo-Gimenez B, Cuggia M, Cabié A. Data-driven methods for dengue prediction and surveillance using real-world and Big Data: A systematic review. PLoS Negl Trop Dis 2022 Jan 7;16(1):e0010056. PMID:34995281

29. World Health Organization. Regional Office for South-East Asia. Comprehensive Guideline for Prevention and Control of Dengue and Dengue Haemorrhagic Fever. Revised and expanded edition. WHO Regional Office for South-East Asia; 2011. ISBN:978-92-9022-387-0

30. Husnayain A, Fuad A, Lazuardi L. Correlation between Google Trends on dengue fever and national surveillance report in Indonesia. Glob Health Action [Internet] 2019 Jan 8 [cited 2019 Jun 19];12(1). PMID:31154985

31. Verma M, Kishore K, Kumar M, Sondh AR, Aggarwal G, Kathirvel S. Google Search Trends Predicting Disease Outbreaks: An Analysis from India. Healthc Inform Res 2018 Oct;24(4):300–308. PMID:30443418

32. Cai O, Sousa-Pinto B. United States Influenza Search Patterns Since the Emergence of COVID-19: Infodemiology Study. JMIR Public Health Surveill 2021 Nov 30; PMID:34878996

33. Gianfredi V, Bragazzi NL, Nucci D, Martini M, Rosselli R, Minelli L, Moretti M. Harnessing Big Data for Communicable Tropical and Sub-Tropical Disorders: Implications From a Systematic Review of the Literature. Front Public Health 2018;6:90. PMID:29619364

34. Liu D, Guo S, Zou M, Chen C, Deng F, Xie Z, Hu S, Wu L. A dengue fever predicting model based on Baidu search index data and climate data in South China. PLoS One 2019;14(12):e0226841. PMID:31887118

35. Puengpreeda A, Yhusumrarn S, Sirikulvadhana S. Weekly Forecasting Model for Dengue Hemorrhagic Fever Outbreak in Thailand. Eng J-Thail 2020 May;24(3):71–87. [doi: 10.4186/ej.2020.24.3.71]

36. Flamand C, Fabregue M, Bringay S, Ardillon V, Quénel P, Desenclos J-C, Teisseire M. Mining local climate data to assess spatiotemporal dengue fever epidemic patterns in French Guiana. J Am Med Inform Assoc 2014 Oct;21(e2):e232-240. PMID:24549761

37. Baquero OS, Santana LMR, Chiaravalloti-Neto F. Dengue forecasting in São Paulo city with generalized additive models, artificial neural networks and seasonal autoregressive integrated moving average models. PLoS One 2018;13(4):e0195065. PMID:29608586

38. Gharbi M, Quenel P, Gustave J, Cassadou S, La Ruche G, Girdary L, Marrama L. Time series analysis of dengue incidence in Guadeloupe, French West Indies: forecasting models using climate variables as predictors. BMC Infect Dis 2011 Jun 9;11:166. PMID:21658238

39. Lowe R, Gasparrini A, Van Meerbeeck CJ, Lippi CA, Mahon R, Trotman AR, Rollock L, Hinds AQJ, Ryan SJ, Stewart-Ibarra AM. Nonlinear and delayed impacts

of climate on dengue risk in Barbados: A modelling study. PLoS Med 2018 Jul;15(7):e1002613. PMID:30016319

40. World Health Organization. Dengue Guidelines for Diagnosis, Treatment, Prevention and Control. Special Programme for Research and Training in Tropical Diseases, editor. Geneva: World Health Organization; 2009. ISBN:978-92-4-154787-1

**Multimedia Appendix 1. ICD-10[a] codes of the selected diagnoses for inpatient data.**

| ICD-10 Code | ICD-10 Title |
|---|---|
| **Dengue** | |
| A90[b] | Dengue fever [classical dengue] |
| A91 [b] | Dengue haemorrhagic fever |
| A97.0 | Dengue without warning signs |
| A97.2 | Dengue with warning signs |
| A97.9 | Severe Dengue |
| **Coding errors** | |
| A99 | Unspecified viral haemorrhagic fever |
| R50.9 | Fever, unspecified |
| **Severity symptoms** | |
| E86 | Volume depletion |
| R57.0 | Cardiogenic shock |
| R57.1 | Hypovolaemic shock |
| R57.2 | Septic shock |
| R57.8 | Other shock |
| R57.9 | Shock, unspecified |
| R58 | Haemorrhage, not elsewhere classified |
| **Thrombocytopenia** | |
| D69.4 | Other primary thrombocytopenia |
| D69.5 | Secondary thrombocytopenia |
| D69.6 | Thrombocytopenia, unspecified |
| **Hepatic symptoms** | |
| B17.8 | Other specified acute viral hepatitis |
| B17.9 | Acute viral hepatitis, unspecified |
| B19.0 | Unspecified viral hepatitis with hepatic coma |
| B19.9 | Unspecified viral hepatitis without hepatic coma |
| K72.0 | Acute and subacute hepatic failure |
| K72.9 | Hepatic failure, unspecified |
| K75.9 | Inflammatory liver disease, unspecified |
| K77.0 | Liver disorders in infectious and parasitic diseases classified elsewhere |
| R16.0 | Hepatomegaly, not elsewhere classified |
| R16.1 | Splenomegaly, not elsewhere classified |
| R16.2 | Hepatomegaly with splenomegaly, not elsewhere classified |
| R74.0 | Elevation of levels of transaminase and lactic acid dehydrogenase [LDH] |
| **Neurologic symptoms** | |
| G05.1 | Encephalitis, myelitis and encephalomyelitis in viral diseases classified elsewhere |
| G93.4 | Encephalopathy, unspecified |
| G94.3 | Encephalopathy in diseases classified elsewhere |

[a] International Classification of Diseases, 10th Revision; [b] Former ICD-10 codes for dengue diagnosis

**Multimedia Appendix 2. Correlations between Google Trends keywords and DENV RT-PCR positive rate**

| Keywords | Correlation | P-value | Confidence interval | Max cross-correlation [a] | Lag [b] |
|---|---|---|---|---|---|
| **Dengue** | | | | | |
| Keywords "dengue + dingue" and region: Martinique | 0.597 | **$6.75 \times 10^{-73}$** | [0.548;0.641] | 0.598 | - 1 week |
| Keywords "dengue + dingue" and "martinique" | NA [c] | NA [c] | NA [c] | NA [c] | NA [c] |
| Keywords "dengue" and "martinique" | 0.534 | **$5.21 \times 10^{-56}$** | [0.480;0.583] | 0.611 | -6 weeks |
| Dengue Topic and region Martinique | 0.637 | **$1.37 \times 10^{-85}$** | [0.591;0.677] | 0.643 | - 3 weeks |
| Keyword "dengue" | -0.016 | 0.654 | [-0.088;0.055] | -0.046 | - 7 weeks |
| **Dengue symptoms** | | | | | |
| Keyword "symptome dengue" and region Martinique | 0.412 | **$6.65 \times 10^{-32}$** | [0.351;0.47] | 0.435 | -3 weeks |
| Keyword "symptome dengue" with various French spellings and region Martinique | 0.238 | **$4.61 \times 10^{-11}$** | [0.169;0.305] | 0.249 | +1 week |
| Keywords "symptome dengue" with various French spellings and "martinique" | NA [c] | NA [c] | NA [c] | NA [c] | NA [c] |
| Keywords "symptome dengue" with various French spellings | 0.222 | **$9.66 \times 10^{-10}$** | [0.152;0.289] | 0.246 | -5 weeks |
| Keyword "symptome dengue" with various spellings and languages, region Martinique | 0.209 | **$8.24 \times 10^{-9}$** | [0.140;0.277] | 0.209 | 0 weeks |
| Keywords "symptome dengue" with various spellings and languages and "martinique" | NA [c] | NA [c] | NA [c] | NA [c] | NA [c] |
| Keywords "symptome dengue" with various spellings and languages | 0.215 | **$3.31 \times 10^{-9}$** | [0.145;0.282] | 0.221 | -3 weeks |
| **Mosquito** | | | | | |

| | | | | | |
|---|---|---|---|---|---|
| Keyword mosquito and region: Martinique | -0.086 | **1.88 x $10^{-2}$** | [-0.157;-0.014] | -0.106 | -6 weeks |
| Keyword mosquito | -0.049 | **1.79 x $10^{-1}$** | [-0.121;0.0223] | -0.101 | -10 weeks |
| Keyword mosquito with various spellings and languages and region: Martinique | 0.028 | **4.48 x $10^{-1}$** | [-0.044;0.100] | 0.033 | +1 week |
| Keyword mosquito with various French spellings and region: Martinique | 0.200 | **3.58 x $10^{-8}$** | [0.130;0.268] | 0.200 | 0 weeks |
| Mosquito Topic and region: Martinique | 0.068 | **6.36 x $10^{-2}$** | [-0.004;0.139] | 0.098 | +8 weeks |
| Keywords “mosquito” and “martinique” | -0.088 | **1.68 x $10^{-2}$** | [-0.158;-0.016] | -0.142 | +3 weeks |
| **Aedes aegypti** | | | | | |
| Keyword “aedes aegypti” and region Martinique | 0.109 | **3.06 x $10^{-3}$** | [0.037;0.179] | 0.112 | -1 week |
| Keywords “aedes aegypti” and “martinique” | NA [c] | NA [c] | NA [c] | NA [c] | NA [c] |
| Keyword “aedes aegypti” | -0.098 | **7.66 x $10^{-3}$** | [-0.169;-0.026] | -0.104 | -4 weeks |
| Aedes aegypti Topic and region Martinique | 0.115 | **1.66 x $10^{-3}$** | [0.044;0.185] | 0.119 | -2 weeks |
| **Aedes** | | | | | |
| Keywords aedes and region Martinique | 0.339 | **1.99 x $10^{-21}$** | [0.273;0.401] | 0.369 | -3 weeks |
| Keywords “aedes” and “martinique” | NA [c] | NA [c] | NA [c] | NA [c] | NA [c] |
| Keyword “aedes” | -0.092 | **1.22 x $10^{-2}$** | [-0.163;-0.02] | -0.100 | -7 weeks |
| Aedes Topic and region Martinique | 0.214 | **3.71 x $10^{-9}$** | [0.591;0.677] | 0.304 | -7 weeks |

[a] Maximum cross-correlation
[b] Time lag that results in the maximum cross-correlation
[c] Not enough data available

# Discussion et perspectives

La surveillance syndromique des maladies infectieuses émergentes est devenue un enjeu majeur de santé publique depuis la fin du 20e siècle. Face à leur recrudescence et à l'augmentation croissance de leur vitesse de diffusion, des systèmes de surveillance sanitaires performants et efficaces sont indispensables pour être capable de réagir rapidement dès les premiers signes d'alerte. La dématérialisation des données de santé a permis l'apparition de nouvelles sources de données jusqu'ici inaccessibles. L'utilisation de ces données de vie réelle en complément des stratégies de surveillance syndromique traditionnelles ouvre de nouvelles perspectives pour l'amélioration des systèmes existants. Cependant, pour utiliser ces données de façon efficace et pertinente, il faut tenir compte de leurs particularités, non seulement liées à leur nature, mais aussi aux méthodes d'exploitation.

L'objectif de cette thèse était d'évaluer la place des données de vie réelle dans la surveillance syndromique de la dengue dans la Caraïbe.

## I. Les données de vie réelle sont un atout dans la surveillance syndromique, mais impliquent des contraintes et prérequis

Le premier verrou exploré par ces travaux est celui lié à la nature des données de vie réelle multi-échelles et hétérogènes. Les problématiques d'intégration de ces données font l'objet de nombreuses publications scientifiques, centrées sur le déploiement d'EDS (67,69) à partir desquels des plateformes multicentriques voient désormais le jour (72,88). Cependant, ces déploiements sont coûteux et complexes et ne peuvent pas se faire en situation de crise. Le premier apport de cette thèse a été de concevoir une architecture d'intégration de données simplifiée permettant l'exploitation des données de vie réelle, en situation de crise, pour une surveillance syndromique quasi en temps réel. En ce qui concerne les processus d'interopérabilité sémantique, là encore, ils doivent être faits en amont. Par ailleurs, bien que des processus automatisés existent, une grande partie des alignements terminologiques sont toujours faits manuellement. Ces alignements sont fastidieux, coûteux et sujets à des risques d'erreurs (108). Les alignements manuels sont d'autant plus fréquents quand il s'agit d'interopérabilité sémantique multilingue, car la plupart des ressources telles que l'Unified

Medical Language System (UMLS) (111) ou Bioportal (112) sont majoritairement anglophones (113). Les travaux de cette thèse ont permis de mettre en place un système semi-automatisé d'alignement terminologiques multilingue, afin de permettre le partage d'information entre pays et territoires de la Caraïbe. Ainsi, ils permettent d'échanger des données issues de régions utilisant des langues différentes, ce qui facilite une surveillance commune des arboviroses.

Ces deux premiers apports ont permis de répondre en partie aux priorités de la *Stratégie commune de prévention et de lutte contre les arboviroses* adoptée par les membres de la PAHO en 2016. Par ailleurs, le choix d'outils simples et/ou open source dans ces travaux garantit leur reproductibilité dans d'autres territoires ou pays caribéens à moindre coût et tient compte de la fracture numérique extrêmement présente dans la région (114).

La deuxième grande problématique explorée par cette thèse concerne l'exploitation des données de vie réelle à travers un cas d'usage : la dengue en Martinique. L'apport des données de vie réelle dans la surveillance syndromique des maladies infectieuses est régulièrement étudiée (40), en particulier depuis 2020 avec la pandémie COVID-19 (95), mais les sources de données et les méthodes dépendent des pathologies étudiées. De plus, la dengue est une maladie vectorielle, ce qui implique à la fois une surveillance syndromique des cas (comme dans la grippe ou le COVID-19), mais aussi une lutte antivectorielle basée sur la surveillance entomologique (23). Cette thèse a permis d'identifier et d'utiliser les méthodes et sources pertinentes à ce cas d'usage, à travers un travail de revue systématisée ainsi que les sources sous-utilisées, en particulier les données de surveillance vectorielles (29). Enfin, la plupart des études dans la littérature sont localisées en Asie, et, pour les études américaines, en Amérique du Sud (notamment le Brésil) et du Nord (Mexique, Etats-Unis), des pays aux surfaces bien plus importantes que les îles des Antilles. Or, les méthodologies d'exploitation des données de vie réelle (notamment celles issues de *Google*) supposent un volume minimal de données (115) qui n'est pas toujours garanti dans la région, non seulement à cause de la taille des territoires et pays qui la constitue, mais aussi à cause des fortes inégalités sociaux-économiques qui se traduisent également par une inégalité des accès aux outils numériques (114). La dernière contribution de cette thèse a donc été d'évaluer la pertinence de sources de données de vie réelle (y compris les données de surveillance vectorielle) dans ce contexte particulier.

Ces travaux présentent malgré tout certaines limites. En ce qui concerne l'intégration des données, bien que les travaux se soient appuyés sur un système d'information ne disposant pas d'EDS, ils s'appuient sur des données déjà fortement dématérialisées. Or, cette dématérialisation et l'accès aux données de santé dans la région Caraïbe sont très variables d'un

pays à l'autre (116), ce qui limite la reproductibilité de ces travaux dans la région. Cependant, la pandémie de COVID-19 et son impact sanitaire et économique a révélé cette fracture numérique (117) et incité les décideurs à combler ce retard rapidement, ce qui devrait permettre d'appliquer ces méthodes dans tous les pays et territoires concernés, y compris ceux qui, pour l'instant ne peuvent pas encore le faire.

En ce qui concerne les approches méthodologiques, les données météorologiques qui font partie des sources de données pertinentes identifiées dans la littérature n'ont pas pu être incluses dans ces travaux car elles n'étaient pas assez exhaustives pour être exploitables sur la période étudiée. D'une façon plus générale, ces travaux ne sont reproductibles que si le volume de données utilisé est suffisant (en particulier pour les systèmes de surveillance vectorielle dont l'implémentation varie d'un pays à l'autre (32)). Pour limiter l'impact de ces variations d'exhaustivité, cette thèse s'est limitée aux sources les plus accessibles et aux variables les plus simples disponibles dans la grande majorité des territoires.

Enfin, le contexte sanitaire a rajouté des contraintes supplémentaires, avec les différentes vagues COVID successives, qui n'avaient pas été identifiées en amont. En particulier, la deuxième vague de COVID-19 en Martinique est apparue au moment du pic de la plus longue épidémie de dengue jamais enregistrée depuis la mise en place de la surveillance de la dengue aux Antilles françaises. Dans ce contexte d'urgence, les prérequis techniques et organisationnels à l'intégration et l'exploitation des données de vie réelle ne pouvaient pas être mis en place dans des conditions habituelles et il a fallu adapter leur implémentation à une situation de crise entraînant une désorganisation du système de soins, une priorisation de certaines sources de données et des processus de contrôle qualité diminués.

Malgré tout, les premiers résultats sur l'exploitation des données pour la surveillance de la dengue sont prometteurs et ce travail doit se poursuivre par l'évaluation et la comparaison de différents modèles d'apprentissage machine pour la prédiction des futures épidémies.

## II. Les données de vie réelle et le *preparedness* épidémiologique

L'intégration des données de vie réelle en amont et à plusieurs échelles (établissement, national, supranational) est une première étape indispensable si l'on souhaite être capable de les utiliser

immédiatement en situation de crise. Ainsi, lors de la pandémie de COVID-19, les hôpitaux qui étaient déjà dotés d'EDS ont pu les utiliser quasi en temps réel pour la production de tableaux de bord de surveillance hospitalière (95), ce qui leur a permis une réorganisation des soins au plus près des besoins quotidiens. Cette crise sanitaire a montré que lorsque les données de vie réelle ont déjà été intégrées en amont, leur exploitation est un atout pour le pilotage en situation d'urgence et la diffusion rapide de l'information. Pourtant la véritable anticipation de ces situations est quasi impossible, puisqu'elles sont par nature imprévisibles.

Pour tenter de prévenir et de répondre aux situations d'épidémies, des modèles de prédiction de plus en plus complexes sont régulièrement étudiés (118). Mais ces modèles sont à prendre avec précaution et leur capacité à véritablement prédire des situations reste limitée lorsqu'on essaie de dépasser la preuve de concept (119). De plus, ces études sont principalement réalisées dans les pays à revenu élevé, alors que les pays à faibles et moyens revenus sont les premiers impactés par les maladies émergentes. Compte tenu de la grande variabilité des données d'une source à l'autre, il est indispensable de développer ces études dans les pays les plus concernés par les émergences si l'on souhaite pouvoir véritablement appliquer ces modèles dans des situations d'urgence.

Ces systèmes d'aide à la décision basés sur des modèles mathématiques sont malgré tout utiles, mais il est surtout nécessaire d'avoir des données disponibles et décloisonnées qui peuvent être rapidement partagées en situation de crise. Ainsi, Ioannidis, Cripps et Tanner (2022) (119) ont analysé les raisons des contre-performances des modèles mathématiques de prédiction du COVID, et leur première proposition pour les améliorer est de favoriser une collecte de données propres et de bonne qualité avant toute modélisation théorique. Ils proposent également de réajuster régulièrement ces modèles avec de vraies données du terrain pour qu'ils soient au plus proche de la réalité. Dans l'idéal ces données collectées respectent le principe « FAIR » : Findable, Accessible, Interoperable, Reusable (Repérables, Accessibles, Interopérables, Réutilisables) (120), ce qui minimise les processus de pré-traitement des données en urgence, et une gouvernance claire concernant l'accès et le partage des données entre les différents acteurs du système est établie en amont (121). Ainsi, lorsque la situation se présente, les données pourront être immédiatement exploitées en fonction des besoins qui peuvent varier très rapidement en fonction de l'évolution d'une épidémie.

Par ailleurs, cette collecte de données repose également sur les acteurs impliqués à chacune des étapes. En effet, pour que ces données puissent se transformer en information pertinente permettant l'aide à la décision, elles ont besoin d'être contextualisées en s'appuyant sur

l'expertise et les interactions entre les acteurs du terrain (cliniciens, biologistes, spécialistes de santé publique…) (122). Si cette organisation multidisciplinaire n'existe pas en amont, le partage rapide des données et leur intégration à la prise de décision deviennent extrêmement difficiles. L'absence d'organisation claire sur le rôle de chacun des acteurs impliqués dans l'intégralité de la chaîne de données a d'ailleurs été identifiée dès 2017 par l'OMS comme un frein au partage efficace de données lors des crises sanitaires (123).

Ainsi, en France, pour faire face à la pandémie de COVID-19, l'Agence Technique de l'Information sur l'Hospitalisation (ATIH) a mis en place la remontée simplifiée et accélérée des données du PMSI en MCO (Médecin Chirurgie Obstétrique) (124). Ce *Fast Track PMSI* a privilégié la remontée simplifiée des séjours MCO des patients hospitalisés (pour COVID-19 dans un premier temps, puis hors COVID-19 ensuite) à un rythme bimensuel au lieu du rythme mensuel habituel d'envoi des données du PMSI. Ces données ont ensuite permis d'alimenter le HDH pour permettre notamment à la communauté scientifique de construire des modèles de propagation de l'épidémie. Mais cette organisation de crise a été possible grâce à la présence de Départements d'Information Médicale (DIM) dans l'ensemble des hôpitaux français en amont de la crise. Ces professionnels de l'information médicale ont pu être réactifs et s'adapter grâce à leur expertise des données. De même la présence d'EDS, mais surtout de Centres de Données Cliniques dans les établissements permettent de consolider l'expertise de terrain et de favoriser les approches multidisciplinaires.

Quels que soient les outils numériques à disposition, c'est avant tout l'expertise des acteurs qui va permettre une vraie résilience des systèmes et une capacité à répondre rapidement aux changements apportés par les épidémies.

Enfin, cette disponibilité de données ne doit pas se limiter uniquement aux données cliniques si l'on souhaite être capable de détecter des signes d'alerte au plus tôt (125). Les émergences de maladies infectieuses ont des causes multifactorielles qui incluent notamment des facteurs sociaux et environnementaux tels que le changement climatique, la modification des écosystèmes (3) ou encore la mobilité humaine (126). Des systèmes de surveillance efficaces doivent donc inclure et prétraiter ces données et être capables de les croiser aux données de surveillance clinique lorsque la situation le demande. Cette approche multidisciplinaire a été expérimentée pendant la crise du COVID-19, avec l'utilisation notamment des smartphones et des réseaux sociaux pour anticiper les risques de transmission (127,128). Elle soulève toutefois de nouvelles questions éthiques en termes d'accessibilité aux données collectées et de

transparence de leur utilisation et doit s'accompagner de véritables mesures garantissant la sécurité et la confidentialité des données (129).

Cette approche multidimensionnelle ne se limitant pas à des systèmes d'information performants, mais préférant les systèmes restant au plus près du terrain, et capables d'évoluer très rapidement en fonction des situations (130) permettrait de transformer les systèmes actuels à la réactivité parfois limitée en véritable systèmes numériques agiles et adaptables aux différentes situations de crise.

# Références


1. WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020 [Internet]. [cité 29 janv 2022]. Disponible sur: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020

2. Morse SS. Factors in the emergence of infectious diseases. Emerg Infect Dis. mars 1995;1(1):7-15.

3. World Health Organization. Regional Office for South-East Asia. A brief guide to emerging infectious diseases and zoonoses [Internet]. WHO Regional Office for South-East Asia; 2014 [cité 29 janv 2022]. Disponible sur: https://apps.who.int/iris/handle/10665/204722

4. Morens DM, Fauci AS. Emerging Infectious Diseases: Threats to Human Health and Global Stability. PLOS Pathogens. 4 juill 2013;9(7):e1003467.

5. World Health Organization, Regional Office for the Western Pacific. Asia Pacific strategy for emerging diseases and public health emergencies: advancing implementation of the International Health Regulations (2005) : working together towards health security. [Internet]. 2017 [cité 16 févr 2022]. Disponible sur: https://apps.who.int/iris/handle/10665/259094

6. Jones KE, Patel NG, Levy MA, Storeygard A, Balk D, Gittleman JL, et al. Global trends in emerging infectious diseases. Nature. févr 2008;451(7181):990-3.

7. Spernovasilis N, Tsiodras S, Poulakou G. Emerging and Re-Emerging Infectious Diseases: Humankind's Companions and Competitors. Microorganisms. 4 janv 2022;10(1):98.

8. Bhutta ZA, Sommerfeld J, Lassi ZS, Salam RA, Das JK. Global burden, distribution, and interventions for infectious diseases of poverty. Infect Dis Poverty. 31 juill 2014;3:21.

9. Bhatt S, Gething PW, Brady OJ, Messina JP, Farlow AW, Moyes CL, et al. The global distribution and burden of dengue. Nature. 25 avr 2013;496(7446):504-7.

10. Weaver SC, Vasilakis N. Molecular evolution of dengue viruses: contributions of phylogenetics to understanding the history and epidemiology of the preeminent arboviral disease. Infect Genet Evol. juill 2009;9(4):523-40.

11. Guo C, Zhou Z, Wen Z, Liu Y, Zeng C, Xiao D, et al. Global Epidemiology of Dengue Outbreaks in 1990-2015: A Systematic Review and Meta-Analysis. Front Cell Infect Microbiol. 2017;7:317.

12. Pang T, Mak TK, Gubler DJ. Prevention and control of dengue-the light at the end of the tunnel. Lancet Infect Dis. mars 2017;17(3):e79-87.

13. Katzelnick LC, Coloma J, Harris E. Dengue: knowledge gaps, unmet needs, and research priorities. Lancet Infect Dis. mars 2017;17(3):e88-100.

14. Stanaway JD, Shepard DS, Undurraga EA, Halasa YA, Coffeng LE, Brady OJ, et al. The Global Burden of Dengue: an analysis from the Global Burden of Disease Study 2013. Lancet Infect Dis. juin 2016;16(6):712-23.

15. Torres JR, Orduna TA, Piña-Pozas M, Vázquez-Vega D, Sarti E. Epidemiological Characteristics of Dengue Disease in Latin America and in the Caribbean: A Systematic Review of the Literature. J Trop Med. 2017;2017:8045435.

16. Dengue - PAHO/WHO | Pan American Health Organization [Internet]. [cité 7 févr 2022]. Disponible sur: https://www.paho.org/en/topics/dengue

17. Integrated Management Strategy for Dengue Prevention and Control in the Caribbean Subregion; 2010 (Sólo en inglés) | OPS/OMS | Organisation panaméricaine de la santé [Internet]. [cité 17 févr 2022]. Disponible sur: https://www.paho.org/fr/node/56792

18. Corriveau R, Philippon B, Yébakima A, éditeurs. La dengue dans les départements français d'Amérique : Comment optimiser la lutte contre cette maladie ? [Internet]. La dengue dans les départements français d'Amérique : Comment optimiser la lutte contre cette maladie ? Marseille: IRD Éditions; 2013 [cité 16 févr 2022]. 204 + 173 (CD-Rom). (Expertise collégiale). Disponible sur: http://books.openedition.org/irdeditions/2668

19. Santé Publique France. Surveillance de la dengue Guadeloupe, Martinique, Saint-Martin, Saint-Barthélemy Point épidémiologique N°05/202. 2021.

20. SPF. Bilan de l'épidémie de dengue en Martinique, 2010 [Internet]. [cité 17 févr 2022]. Disponible sur: https://www.santepubliquefrance.fr/antilles/bilan-de-l-epidemie-de-dengue-en-martinique-2010

21. SPF. Bulletin de veille sanitaire Antilles-Guyane. n°2-3 - Mars 2015. [Internet]. [cité 17 févr 2022]. Disponible sur: https://www.santepubliquefrance.fr/guyane2/bulletin-de-veille-sanitaire-antilles-guyane.-n-2-3-mars-2015

22. Brady OJ, Gething PW, Bhatt S, Messina JP, Brownstein JS, Hoen AG, et al. Refining the global spatial limits of dengue virus transmission by evidence-based consensus. PLoS Negl Trop Dis. 2012;6(8):e1760.

23. World Health Organization. Dengue Guidelines for Diagnosis, Treatment, Prevention and Control. Special Programme for Research and Training in Tropical Diseases, éditeur. Geneva: World Health Organization; 2009. 147 p.

24. World Health Organization. Dengue haemorrhagic fever : diagnosis, treatment, prevention and control [Internet]. World Health Organization; 1997 [cité 16 févr 2022]. Disponible sur: https://apps.who.int/iris/handle/10665/41988

25. Bandyopadhyay S, Lum LCS, Kroeger A. Classifying dengue: a review of the difficulties in using the WHO case classification for dengue haemorrhagic fever. Trop Med Int Health. août 2006;11(8):1238-55.

26. Cafferata ML, Bardach A, Rey-Ares L, Alcaraz A, Cormick G, Gibbons L, et al. Dengue Epidemiology and Burden of Disease in Latin America and the Caribbean: A Systematic Review of the Literature and Meta-Analysis. Value Health Reg Issues. déc 2013;2(3):347-56.

27. Qui sommes nous ? [Internet]. [cité 16 févr 2022]. Disponible sur: https://www.santepubliquefrance.fr/a-propos/sante-publique-france-qui-sommes-nous

28. Dussart DOF, Gustave10 J, Lagathu G, Koulman11 L, Lordinot ML, Martial J, et al. PROGRAMME DE SURVEILLANCE, D'ALERTE ET DE GESTION DES EPIDEMIES DE DENGUE (PSAGE DENGUE) EN MARTINIQUE.

29. CD55/16 - Strategy for Arboviral Disease Prevention and Control - PAHO/WHO | Pan American Health Organization [Internet]. [cité 30 janv 2022]. Disponible sur: https://www.paho.org/en/documents/cd5516-strategy-arboviral-disease-prevention-and-control

30. Milinovich GJ, Williams GM, Clements ACA, Hu W. Internet-based surveillance systems for monitoring emerging infectious diseases. Lancet Infect Dis. févr 2014;14(2):160-8.

31. Madoff LC, Fisman DN, Kass-Hout T. A new approach to monitoring dengue activity. PLoS Negl Trop Dis. mai 2011;5(5):e1215.

32. H G-D, Jr W. Dengue in the Americas: challenges for prevention and control. Cadernos de saude publica [Internet]. 2009 [cité 25 janv 2022];25 Suppl 1. Disponible sur: https://pubmed.ncbi.nlm.nih.gov/19287863/

33. Shepard DS, Undurraga EA, Betancourt-Cravioto M, Guzmán MG, Halstead SB, Harris E, et al. Approaches to refining estimates of global burden and economics of dengue. PLoS Negl Trop Dis. nov 2014;8(11):e3306.

34. Carneiro HA, Mylonakis E. Google trends: a web-based tool for real-time surveillance of disease outbreaks. Clin Infect Dis. 15 nov 2009;49(10):1557-64.

35. Eysenbach G. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. J Med Internet Res. 27 mars 2009;11(1):e11.

36. Lu FS, Hattab MW, Clemente CL, Biggerstaff M, Santillana M. Improved state-level influenza nowcasting in the United States leveraging Internet-based data and network approaches. Nat Commun. 11 janv 2019;10(1):147.

37. Clemente L, Lu F, Santillana M. Improved Real-Time Influenza Surveillance: Using Internet Search Data in Eight Latin American Countries. JMIR Public Health Surveill. 4 avr 2019;5(2):e12214.

38. Samaras L, Sicilia M-A, García-Barriocanal E. Predicting epidemics using search engine data: a comparative study on measles in the largest countries of Europe. BMC Public Health. 21 janv 2021;21(1):100.

39. Wilson K, Brownstein JS. Early detection of disease outbreaks using the Internet. CMAJ. 14 avr 2009;180(8):829-31.

40. Gianfredi V, Bragazzi NL, Nucci D, Martini M, Rosselli R, Minelli L, et al. Harnessing Big Data for Communicable Tropical and Sub-Tropical Disorders: Implications From a Systematic Review of the Literature. Front Public Health [Internet]. 21 mars 2018

[cité 28 nov 2018];6. Disponible sur: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5871696/

41. Gluskin RT, Johansson MA, Santillana M, Brownstein JS. Evaluation of Internet-based dengue query data: Google Dengue Trends. PLoS Negl Trop Dis. févr 2014;8(2):e2713.

42. Mavragani A. Infodemiology and Infoveillance: Scoping Review. J Med Internet Res. 28 avr 2020;22(4):e16206.

43. Adoption Model for Analytics Maturity (AMAM) | HIMSS [Internet]. 2021 [cité 16 févr 2022]. Disponible sur: https://www.himss.org/what-we-do-solutions/digital-health-transformation/maturity-models/adoption-model-analytics-maturity-amam

44. Sherman RE, Anderson SA, Dal Pan GJ, Gray GW, Gross T, Hunter NL, et al. Real-World Evidence - What Is It and What Can It Tell Us? N Engl J Med. 8 déc 2016;375(23):2293-7.

45. Bégaud B, Poltron D, Von Lennep F. Les données de vie réelle, un enjeu majeur pour la qualité des soins et la régulation du système de santé - L'exemple du médicament [Internet]. [cité 17 févr 2022]. Disponible sur: https://www.vie-publique.fr/rapport/37068-les-donnees-de-vie-reelle-enjeu-majeur-pour-la-qualite-des-soins

46. Weber GM, Mandl KD, Kohane IS. Finding the Missing Link for Big Biomedical Data. JAMA [Internet]. 22 mai 2014 [cité 5 févr 2022]; Disponible sur: http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.2014.4228

47. Flynn R, Plueschke K, Quinten C, Strassmann V, Duijnhoven RG, Gordillo-Marañon M, et al. Marketing Authorization Applications Made to the European Medicines Agency in 2018-2019: What was the Contribution of Real-World Evidence? Clin Pharmacol Ther. janv 2022;111(1):90-7.

48. PERSONNEL GDPDM, GROUPEMENT D'INTERET PUBLIC DOSSIER MEDICAL PERSONNEL. Dossier médical personnel : la mémoire santé. ADSP : ACTUALITE ET DOSSIER EN SANTE PUBLIQUE. 2007;(58):18-21.

49. Mon espace santé, un nouveau service numérique personnel et sécurisé [Internet]. [cité 16 févr 2022]. Disponible sur: https://www.ameli.fr/assure/actualites/mon-espace-sante-un-nouveau-service-numerique-personnel-et-securise

50. Herrett E, Gallagher AM, Bhaskaran K, Forbes H, Mathur R, van Staa T, et al. Data Resource Profile: Clinical Practice Research Datalink (CPRD). Int J Epidemiol. juin 2015;44(3):827-36.

51. Tuppin P, Rudant J, Constantinou P, Gastaldi-Ménager C, Rachas A, de Roquefeuil L, et al. Value of a national administrative database to guide public decisions: From the système national d'information interrégimes de l'Assurance Maladie (SNIIRAM) to the système national des données de santé (SNDS) in France. Rev Epidemiol Sante Publique. oct 2017;65 Suppl 4:S149-67.

52. Al-Alusi MA, Ding E, McManus DD, Lubitz SA. Wearing Your Heart on Your Sleeve: the Future of Cardiac Rhythm Monitoring. Curr Cardiol Rep. 25 nov 2019;21(12):158.

53. IoT Platform for COVID-19 Prevention and Control: A Survey. IEEE Access. 23 mars 2021;9:49929-41.

54. Beier JI, Arteel GE. Environmental exposure as a risk-modifying factor in liver diseases: Knowns and unknowns. Acta Pharm Sin B. déc 2021;11(12):3768-78.

55. Yeh H-Y, Chen K-H, Chen K-T. Environmental Determinants of Infectious Disease Transmission: A Focus on One Health Concept. Int J Environ Res Public Health. juin 2018;15(6):1183.

56. Boslaugh S. Secondary Data Sources for Public Health: A Practical Guide [Internet]. Cambridge: Cambridge University Press; 2007 [cité 5 févr 2022]. (Practical Guides to Biostatistics and Epidemiology). Disponible sur: https://www.cambridge.org/core/books/secondary-data-sources-for-public-health/C08BD3F16D011E880FDD6BFEDCB8B463

57. The Importance of « Big Data »: A Definition [Internet]. Gartner. [cité 17 févr 2022]. Disponible sur: https://www.gartner.com/en/documents/2057415/the-importance-of-big-data-a-definition

58. Taleb I, Serhani MA, Bouhaddioui C, Dssouli R. Big data quality framework: a holistic approach to continuous quality management. Journal of Big Data. 29 mai 2021;8(1):76.

59. Ehsani-Moghaddam B, Martin K, Queenan JA. Data quality in healthcare: A report of practical experience with the Canadian Primary Care Sentinel Surveillance Network data. HIM J. janv 2021;50(1-2):88-92.

60. Pastorino R, De Vito C, Migliara G, Glocker K, Binenbaum I, Ricciardi W, et al. Benefits and challenges of Big Data in healthcare: an overview of the European initiatives. European Journal of Public Health. 1 oct 2019;29(Supplement_3):23-7.

61. L'anonymisation de données personnelles | CNIL [Internet]. [cité 16 févr 2022]. Disponible sur: https://www.cnil.fr/fr/lanonymisation-de-donnees-personnelles

62. Dash S, Shakyawar SK, Sharma M, Kaushik S. Big data in healthcare: management, analysis and future prospects. Journal of Big Data. 19 juin 2019;6(1):54.

63. Agrawal R, Prabakaran S. Big data in digital healthcare: lessons learnt and recommendations for general practice. Heredity. avr 2020;124(4):525-34.

64. Règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, et abrogeant la directive 95/46/CE (règlement général sur la protection des données) (Texte présentant de l'intérêt pour l'EEE) [Internet]. OJ L avr 27, 2016. Disponible sur: http://data.europa.eu/eli/reg/2016/679/oj/fra

65. Edemekong PF, Annamaraju P, Haydel MJ. Health Insurance Portability and Accountability Act. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2022 [cité 17 févr 2022]. Disponible sur: http://www.ncbi.nlm.nih.gov/books/NBK500019/

66. National Research Council (US) Committee on A Framework for Developing a New Taxonomy of Disease. Toward Precision Medicine: Building a Knowledge Network for

Biomedical Research and a New Taxonomy of Disease [Internet]. Washington (DC): National Academies Press (US); 2011 [cité 7 févr 2022]. (The National Academies Collection: Reports funded by National Institutes of Health). Disponible sur: http://www.ncbi.nlm.nih.gov/books/NBK91503/

67. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc. avr 2010;17(2):124-30.

68. Kimball. The data warehousing toolkit New York. 1997.

69. Madec J, Bouzille G, Riou C, Van Hille P, Merour C, Artigny ML, et al. eHOP Clinical Data Warehouse: From a Prototype to the Creation of an Inter-Regional Clinical Data Centers Network. In: OhnoMachado L, Seroussi B, éditeurs. Medinfo 2019: Health and Wellbeing E-Networks for All. Amsterdam: Ios Press; 2019. p. 1536-7. (Studies in Health Technology and Informatics; vol. 264).

70. Cuggia M, Garcelon N, Campillo-Gimenez B, Bernicot T, Laurent J-F, Garin E, et al. Roogle: an information retrieval engine for clinical data warehouse. Stud Health Technol Inform. 2011;169:584-8.

71. McMurry AJ, Murphy SN, MacFadden D, Weber G, Simons WW, Orechia J, et al. SHRINE: Enabling Nationally Scalable Multi-Site Disease Studies. PLoS ONE. 7 mars 2013;8(3):e55811.

72. Weber GM, Murphy SN, McMurry AJ, MacFadden D, Nigrin DJ, Churchill S, et al. The Shared Health Research Information Network (SHRINE): A Prototype Federated Query Tool for Clinical Data Repositories. Journal of the American Medical Informatics Association. 1 sept 2009;16(5):624-30.

73. Kohane IS, McMurry A, Weber G, MacFadden D, Rappaport L, Kunkel L, et al. The Co-Morbidity Burden of Children and Young Adults with Autism Spectrum Disorders. PLoS ONE. 12 avr 2012;7(4):e33224.

74. Anderson N, Abend A, Mandel A, Geraghty E, Gabriel D, Wynden R, et al. Implementation of a deidentified federated data network for population-based cohort discovery. J Am Med Inform Assoc. juin 2012;19(e1):e60-67.

75. Ganslandt T, Mate S, Helbing K, Sax U, Prokosch HU. Unlocking Data for Clinical Research - The German i2b2 Experience. Appl Clin Inform. 2011;2(1):116-27.

76. Natter MD, Quan J, Ortiz DM, Bousvaros A, Ilowite NT, Inman CJ, et al. An i2b2-based, generalizable, open source, self-scaling chronic disease registry. J Am Med Inform Assoc. 1 janv 2013;20(1):172-9.

77. Soto-Rey I, Trinczek B, Girardeau Y, Zapletal E, Ammour N, Doods J, et al. Efficiency and effectiveness evaluation of an automated multi-country patient count cohort system. BMC Med Res Methodol. 1 mai 2015;15:44.

78. De Moor G, Sundgren M, Kalra D, Schmidt A, Dugas M, Claerhout B, et al. Using electronic health records for clinical research: the case of the EHR4CR project. J Biomed Inform. févr 2015;53:162-73.

79. Corley DA, Feigelson HS, Lieu TA, McGlynn EA. Building Data Infrastructure to Evaluate and Improve Quality: PCORnet. J Oncol Pract. mai 2015;11(3):204-6.

80. Fleurence RL, Curtis LH, Califf RM, Platt R, Selby JV, Brown JS. Launching PCORnet, a national patient-centered clinical research network. J Am Med Inform Assoc. août 2014;21(4):578-82.

81. Hollin IL, Dimmock AE, Bridges JF, Danoff SK, Bascom R. Collecting patient preference information using a Clinical Data Research Network: demonstrating feasibility with idiopathic pulmonary fibrosis. Patient Prefer Adherence. 2019;13:795-804.

82. Gruß I, McCreary GM, Ivlev I, Houlihan ME, Yawn BP, Pasquale C, et al. Developing a patient-driven chronic obstructive pulmonary disease (COPD) research agenda in the U.S. J Patient Rep Outcomes. 4 déc 2021;5(1):126.

83. Marino P, Bannier M, Moulin J-F, Gravis G. [The role and use of Patient Reported Outcomes in the management of cancer patients]. Bull Cancer. juin 2018;105(6):603-9.

84. Common Data Model (CDM) Specification, Version 6.0. :193.

85. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform. 2015;216:574-8.

86. Klann JG, Joss MAH, Embree K, Murphy SN. Data model harmonization for the All Of Us Research Program: Transforming i2b2 data into the OMOP common data model. PLOS ONE. 19 févr 2019;14(2):e0212463.

87. Health Data Hub (HDH) [Internet]. G_NIUS. 2020 [cité 16 févr 2022]. Disponible sur: https://gnius.esante.gouv.fr/fr/acteurs/fiches-acteur/health-data-hub-hdh

88. Bouzillé G, Westerlynck R, Defossez G, Bouslimi D, Bayat S, Riou C, et al. Sharing Health Big Data for Research - A Design by Use Cases: The INSHARE Platform Approach. Stud Health Technol Inform. 2017;245:303-7.

89. Pladys A, Defossez G, Lemordant P, Lassalle M, Ingrand P, Jacquelinet C, et al. Cancer risk in dialyzed patients with and without diabetes. Cancer Epidemiol. avr 2020;65:6.

90. Bannay A, Bories M, Le Corre P, Riou C, Lemordant P, Van Hille P, et al. Leveraging National Claims and Hospital Big Data: Cohort Study on a Statin-Drug Interaction Use Case. JMIR Med Inform. 13 déc 2021;9(12):e29286.

91. BIG DATA : « Ouest Datahub », première plateforme européenne de données hospitalières, lancée par le GCS HUGO [Internet]. Accueil GCS HUGO. 2020 [cité 18 févr 2022]. Disponible sur: https://www.chu-hugo.fr/accueil/2020/12/18/ouest-datahub-premiere-plateforme-europeenne-de-donnees-hospitalieres-lancee-par-le-gcs-hugo/

92. Bouzillé G, Poirier C, Campillo-Gimenez B, Aubert M-L, Chabot M, Chazard E, et al. Leveraging hospital big data to monitor flu epidemics. Computer Methods and Programs in Biomedicine. 1 févr 2018;154:153-60.

93. Poirier C, Lavenu A, Bertaud V, Campillo-Gimenez B, Chazard E, Cuggia M, et al. Real Time Influenza Monitoring Using Hospital Big Data in Combination with Machine

Learning Methods: Comparison Study. JMIR Public Health and Surveillance. 21 déc 2018;4(4):e11361.

94. Fartoukh M, Voiriot G, Guérin L, Ricard JD, Combes A, Faure M, et al. Seasonal burden of severe influenza virus infection in the critically ill patients, using the Assistance Publique-Hôpitaux de Paris clinical data warehouse: a pilot study. Ann Intensive Care. 29 juill 2021;11(1):117.

95. Budd J, Miller BS, Manning EM, Lampos V, Zhuang M, Edelstein M, et al. Digital technologies in the public-health response to COVID-19. Nat Med. août 2020;26(8):1183-92.

96. Brat GA, Weber GM, Gehlenborg N, Avillach P, Palmer NP, Chiovato L, et al. International electronic health record-derived COVID-19 clinical course profiles: the 4CE consortium. NPJ Digit Med. 2020;3:109.

97. Le TT, Gutiérrez-Sacristán A, Son J, Hong C, South AM, Beaulieu-Jones BK, et al. Multinational characterization of neurological phenotypes in patients hospitalized with COVID-19. Sci Rep. 12 oct 2021;11(1):20238.

98. Neuraz A, Lerner I, Digan W, Paris N, Tsopra R, Rogier A, et al. Natural Language Processing for Rapid Response to Emergent Diseases: Case Study of Calcium Channel Blockers and Hypertension in the COVID-19 Pandemic. J Med Internet Res. 14 août 2020;22(8):e20773.

99. Mascitti H, Jourdain P, Bleibtreu A, Jaulmes L, Dechartres A, Lescure X, et al. Prognosis of rash and chilblain-like lesions among outpatients with COVID-19: a large cohort study. Eur J Clin Microbiol Infect Dis. oct 2021;40(10):2243-8.

100. Gangloff C, Rafi S, Bouzillé G, Soulat L, Cuggia M. Machine learning is the key to diagnose COVID-19: a proof-of-concept study. Sci Rep. 30 mars 2021;11(1):7166.

101. Mehra MR, Desai SS, Kuy S, Henry TD, Patel AN. Cardiovascular Disease, Drug Therapy, and Mortality in Covid-19. N Engl J Med. 18 juin 2020;382(25):e102.

102. Mehra MR, Desai SS, Ruschitzka F, Patel AN. RETRACTED: Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis. Lancet. 22 mai 2020;S0140-6736(20)31180-6.

103. Kohane IS, Aronow BJ, Avillach P, Beaulieu-Jones BK, Bellazzi R, Bradford RL, et al. What Every Reader Should Know About Studies Using Electronic Health Record Data but May Be Afraid to Ask. J Med Internet Res. 2 mars 2021;23(3):e22219.

104. Abd-Alrazaq A, Hassan A, Abuelezz I, Ahmed A, Alzubaidi MS, Shah U, et al. Overview of Technologies Implemented During the First Wave of the COVID-19 Pandemic: Scoping Review. J Med Internet Res. 14 sept 2021;23(9):e29136.

105. Bastias-Butler Elizabeth;, Ulrich, Andrea. Digital Transformation of the Health Sector in Latin America and the Caribbean | Publications. avr 2019 [cité 6 févr 2022]; Disponible sur: https://publications.iadb.org/publications/english/document/Digital_Transformation_of_the_Health_Sector_in_Latin_America_and_the_Caribbean_en_en.pdf

106. Zou Y, Lu W. Learning Cross-lingual Distributed Logical Representations for Semantic Parsing. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) [Internet]. Melbourne, Australia: Association for Computational Linguistics; 2018 [cité 16 févr 2022]. p. 673-9. Disponible sur: https://aclanthology.org/P18-2107

107. Liang P. Learning executable semantic parsers for natural language understanding. Commun ACM. 24 août 2016;59(9):68-76.

108. Dos Santos CT, Quaresma P, Vieira R. An API for multilingual ontology matching. In: Proc 7th conference on Language Resources and Evaluation Conference (LREC) [Internet]. No commercial editor.; 2010 [cité 30 mars 2016]. p. 3830-5. Disponible sur: https://hal.inria.fr/hal-00793285/

109. International Classification of Diseases (ICD) [Internet]. [cité 16 févr 2022]. Disponible sur: https://www.who.int/standards/classifications/classification-of-diseases

110. Mistichelli J. Diagnosis Related Groups (DRGs). juin 1984 [cité 17 févr 2022]; Disponible sur: https://repository.library.georgetown.edu/handle/10822/556896

111. Fung KW, Bodenreider O. Utilizing the UMLS for Semantic Mapping between Terminologies. AMIA Annu Symp Proc. 2005;2005:266-70.

112. Whetzel PL, Noy NF, Shah NH, Alexander PR, Nyulas C, Tudorache T, et al. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic Acids Res. juill 2011;39(Web Server issue):W541-545.

113. Doan A, Madhavan J, Domingos P, Halevy A. Ontology matching: A machine learning approach. In: Handbook on ontologies [Internet]. Springer; 2004 [cité 30 mars 2016]. p. 385-403. Disponible sur: http://link.springer.com/chapter/10.1007/978-3-540-24750-0_19

114. Pick J, Sarkar A, Parrish E. The Latin American and Caribbean digital divide: a geospatial and multivariate analysis. Information Technology for Development. 3 avr 2021;27(2):235-62.

115. Mavragani A, Ochoa G. Google Trends in Infodemiology and Infoveillance: Methodology Framework. JMIR Public Health and Surveillance. 29 mai 2019;5(2):e13439.

116. Robinson L, Schulz J, Dodel M, Correa T, Villanueva-Mansilla E, Leal S, et al. Digital Inclusion Across the Americas and Caribbean. Social Inclusion. 14 mai 2020;8:244.

117. Closing the digital gap to end poverty in Latin America and the Caribbean [Internet]. [cité 17 févr 2022]. Disponible sur: https://blogs.worldbank.org/latinamerica/closing-digital-gap-end-poverty-latin-america-and-caribbean

118. Sweileh WM. Global research activity on mathematical modeling of transmission and control of 23 selected infectious disease outbreak. Globalization and Health. 21 janv 2022;18(1):4.

119. Ioannidis JPA, Cripps S, Tanner MA. Forecasting for COVID-19 has failed. Int J Forecast. 2022;38(2):423-38.

120. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 15 mars 2016;3:160018.

121. Group G-RDSW. Principles for Data Sharing in Public Health Emergencies. 30 mars 2017 [cité 17 févr 2022]; Disponible sur: https://wellcome.figshare.com/articles/journal_contribution/Principles_for_Data_Sharing_in_Public_Health_Emergencies/4733590/2

122. Pillai P. How Do Data Bolster Pandemic Preparedness and Response? How Do We Improve Data and Systems to Be Better Prepared? Patterns. 8 janv 2021;2(1):100190.

123. Littler K, Boon W-M, Carson G, Depoortere E, Mathewson S, Mietchen D, et al. Progress in promoting data sharing in public health emergencies. Bull World Health Organ. 1 avr 2017;95(4):243.

124. Remontée simplifiée des données PMSI | Publication ATIH [Internet]. [cité 17 févr 2022]. Disponible sur: https://www.atih.sante.fr/remontee-simplifiee-des-donnees-pmsi

125. Ziemann A, Fouillet A, Brand H, Krafft T. Success Factors of European Syndromic Surveillance Systems: A Worked Example of Applying Qualitative Comparative Analysis. PLOS ONE. 16 mai 2016;11(5):e0155535.

126. Stoddard ST, Morrison AC, Vazquez-Prokopec GM, Paz Soldan V, Kochel TJ, Kitron U, et al. The role of human movement in the transmission of vector-borne pathogens. PLoS Negl Trop Dis. 21 juill 2009;3(7):e481.

127. Bisanzio D, Kraemer MUG, Bogoch II, Brewer T, Brownstein JS, Reithinger R. Use of Twitter social media activity as a proxy for human mobility to predict the spatiotemporal spread of COVID-19 at global scale. Geospat Health. 15 juin 2020;15(1).

128. Yechezkel M, Weiss A, Rejwan I, Shahmoon E, Ben-Gal S, Yamin D. Human mobility and poverty as key drivers of COVID-19 transmission and control. BMC Public Health. 25 mars 2021;21(1):596.

129. Boersma K, Büscher M, Fonio C. Crisis management, surveillance, and digital ethics in the COVID-19 era. Journal of Contingencies and Crisis Management [Internet]. [cité 17 févr 2022];n/a(n/a). Disponible sur: https://onlinelibrary.wiley.com/doi/abs/10.1111/1468-5973.12398

130. Morand S, Figuié M. Émergence de maladies infectieuses - Risques et enjeux de société. :139.

# Annexes

Les annexes ci-dessous correspondent aux annexes de l'article « Data-driven methods for dengue prediction and surveillance using real-world and Big Data: A systematic review » présenté dans le Chapitre 1 de la Troisième partie.

# PRISMA 2020 Checklist

| Section and Topic | Item # | Checklist item | Location where item is reported |
|---|---|---|---|
| **TITLE** | | | |
| Title | 1 | Identify the report as a systematic review. | 1 |
| **ABSTRACT** | | | |
| Abstract | 2 | See the PRISMA 2020 for Abstracts checklist. | |
| **INTRODUCTION** | | | |
| Rationale | 3 | Describe the rationale for the review in the context of existing knowledge. | 5,6 |
| Objectives | 4 | Provide an explicit statement of the objective(s) or question(s) the review addresses. | 6 |
| **METHODS** | | | |
| Eligibility criteria | 5 | Specify the inclusion and exclusion criteria for the review and how studies were grouped for the syntheses. | 7 |
| Information sources | 6 | Specify all databases, registers, websites, organisations, reference lists and other sources searched or consulted to identify studies. Specify the date when each source was last searched or consulted. | 8 |
| Search strategy | 7 | Present the full search strategies for all databases, registers and websites, including any filters and limits used. | 8 |
| Selection process | 8 | Specify the methods used to decide whether a study met the inclusion criteria of the review, including how many reviewers screened each record and each report retrieved, whether they worked independently, and if applicable, details of automation tools used in the process. | 8,9 |
| Data collection process | 9 | Specify the methods used to collect data from reports, including how many reviewers collected data from each report, whether they worked independently, any processes for obtaining or confirming data from study investigators, and if applicable, details of automation tools used in the process. | 9 |
| Data items | 10a | List and define all outcomes for which data were sought. Specify whether all results that were compatible with each outcome domain in each study were sought (e.g. for all measures, time points, analyses), and if not, the methods used to decide which results to collect. | 9,S1 Text |
| | 10b | List and define all other variables for which data were sought (e.g. participant and intervention characteristics, funding sources). Describe any assumptions made about any missing or unclear information. | 9,S1 Text |
| Study risk of bias assessment | 11 | Specify the methods used to assess risk of bias in the included studies, including details of the tool(s) used, how many reviewers assessed each study and whether they worked independently, and if applicable, details of automation tools used in the process. | 9, S2 Table |
| Effect measures | 12 | Specify for each outcome the effect measure(s) (e.g. risk ratio, mean difference) used in the synthesis or presentation of results. | 9 |
| Synthesis methods | 13a | Describe the processes used to decide which studies were eligible for each synthesis (e.g. tabulating the study intervention characteristics and comparing against the planned groups for each synthesis (item #5)). | 9, S2 Table |
| | 13b | Describe any methods required to prepare the data for presentation or synthesis, such as handling of missing summary statistics, or data conversions. | 9 |
| | 13c | Describe any methods used to tabulate or visually display results of individual studies and syntheses. | 9 |
| | 13d | Describe any methods used to synthesize results and provide a rationale for the choice(s). If meta-analysis was performed, describe the model(s), method(s) to identify the presence and extent of statistical heterogeneity, and software package(s) used. | NA |
| | 13e | Describe any methods used to explore possible causes of heterogeneity among study results (e.g. subgroup analysis, meta-regression). | NA |
| | 13f | Describe any sensitivity analyses conducted to assess robustness of the synthesized results. | NA |
| Reporting bias assessment | 14 | Describe any methods used to assess risk of bias due to missing results in a synthesis (arising from reporting biases). | NA |
| Certainty assessment | 15 | Describe any methods used to assess certainty (or confidence) in the body of evidence for an outcome. | NA |

# PRISMA 2020 Checklist

| Section and Topic | Item # | Checklist item | Location where item is reported |
|---|---|---|---|
| **RESULTS** | | | |
| Study selection | 16a | Describe the results of the search and selection process, from the number of records identified in the search to the number of studies included in the review, ideally using a flow diagram. | 9 |
| | 16b | Cite studies that might appear to meet the inclusion criteria, but which were excluded, and explain why they were excluded. | 9, Fig. 1 |
| Study characteristics | 17 | Cite each included study and present its characteristics. | 9, 10 S4 Table |
| Risk of bias in studies | 18 | Present assessments of risk of bias for each included study. | 15,16 |
| Results of individual studies | 19 | For all outcomes, present, for each study: (a) summary statistics for each group (where appropriate) and (b) an effect estimate and its precision (e.g. confidence/credible interval), ideally using structured tables or plots. | S3 Table, S6 Table |
| Results of syntheses | 20a | For each synthesis, briefly summarise the characteristics and risk of bias among contributing studies. | 13-17 |
| | 20b | Present results of all statistical syntheses conducted. If meta-analysis was done, present for each the summary estimate and its precision (e.g. confidence/credible interval) and measures of statistical heterogeneity. If comparing groups, describe the direction of the effect. | NA |
| | 20c | Present results of all investigations of possible causes of heterogeneity among study results. | NA |
| | 20d | Present results of all sensitivity analyses conducted to assess the robustness of the synthesized results. | NA |
| Reporting biases | 21 | Present assessments of risk of bias due to missing results (arising from reporting biases) for each synthesis assessed. | NA |
| Certainty of evidence | 22 | Present assessments of certainty (or confidence) in the body of evidence for each outcome assessed. | NA |
| **DISCUSSION** | | | |
| Discussion | 23a | Provide a general interpretation of the results in the context of other evidence. | 18-23 |
| | 23b | Discuss any limitations of the evidence included in the review. | 22 |
| | 23c | Discuss any limitations of the review processes used. | 22 |
| | 23d | Discuss implications of the results for practice, policy, and future research. | 22-23 |
| **OTHER INFORMATION** | | | |
| Registration and protocol | 24a | Provide registration information for the review, including register name and registration number, or state that the review was not registered. | 6, S2 Text |
| | 24b | Indicate where the review protocol can be accessed, or state that a protocol was not prepared. | 6, S2 Text |
| | 24c | Describe and explain any amendments to information provided at registration or in the protocol. | S2 Text |
| Support | 25 | Describe sources of financial or non-financial support for the review, and the role of the funders or sponsors in the review. | NA |
| Competing interests | 26 | Declare any competing interests of review authors. | NA |
| Availability of data, code and other materials | 27 | Report which of the following are publicly available and where they can be found: template data collection forms; data extracted from included studies; data used for all analyses; analytic code; any other materials used in the review. | NA |

*From:* Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71. doi: 10.1136/bmj.n71

**S1 Table. Quality assessment criteria**

| Section | Checklist item | Page |
|---|---|---|
| **Title and abstract** | | |
| Title | For prediction: Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted. | |
| Abstract | Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions. | |
| **Objective and background** | | |
| Background | For prediction: Explain the context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models. | |
| Objectives | 1. Specify the objectives, including (for prediction) whether the study developed and/or or validated the model. | |
| | 2. Specify the geographic area of the study. | |
| **Methods** | | |
| Source of data | 1. Describe the source of data for dengue infection. | |
| | 2. Describe the other sources of data. | |
| | 3. Specify the study period. | |
| Participants (when applicable) | 1. Specify key elements of the study setting (population, age, gender…). | |
| | 2. Describe the eligibility criteria for participants. | |
| Outcome | Clearly define the study outcome. | |
| Statistical methods | Specify method or model type, all model-building procedures (including feature selection), and method for internal validation. | |
| | Assess whether the statistical methods are appropriate for the study aim. | |
| Predictors (when applicable) | Clearly define all predictors used for developing or validating the model, including how and when they were measured. | |
| Evaluation methods | 1. Specify all measures used to assess statistic and model performance and, if relevant, to compare multiple models. | |
| | 2. If applicable, specify all measures to validate the methods and the use of a training and validation set. | |

**S1 Table.** *(continued)*

| Section | Checklist item | Page |
|---|---|---|
| Results | | |
| Participants (when applicable) | Describe the number of participants and their characteristics including the number of participants with missing data for predictors and outcome. | |
| Statistical methods | Describe completely and clearly all results. | |
| | Explain how to the use the prediction model (when applicable). | |
| Discussion | | |
| Limitations | Discuss the study limitations. | |
| Interpretation | Give an overall interpretation of the results, considering objectives, limitations, and results from similar studies, and other relevant evidence. | |
| Implications | Discuss the potential use of the approach and implications for future research. | |
| Other information | | |
| Funding | Give the source of funding and the role of the funders in the present study. | |
| Conflict of interest | Give a conflict of interest statement if necessary. | |
| Supplementary information | Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets. | |

**S1 Text. Data extraction sheet**

| | |
|---|---|
| **Study characteristics** | |
| Reviewer's initials | |
| Study ID | |
| PMID | |
| DOI | |
| Study title | |
| Authors (all of them) | |
| First author | Last name and First Initial |
| Last author | Last name and First Initial |
| Journal | |
| Year | |
| Type of publication | Article or Conference Paper |
| **Study location** | |
| Continent(s) where the study was performed (select all that apply) | Asia<br>Americas<br>Australia<br>Europe<br>Australia<br>Worldwide |
| Region(s) where the study was performed (select all that apply) | Australia<br>Caribbean<br>East Asia<br>Europe<br>North America<br>South America<br>South Asia<br>South-East Asia<br>Worldwide |
| Country(ies) where the study was performed (specify the country) | |
| **Data sources** | |
| Year(s) of data collection | |
| Used data sources (select all that apply) | Epidemiological and demographic data<br>Clinical and biological data<br>Genomic sequencing data<br>Climate, environmental and geographic data<br>Vector data<br>Internet search engine data (specify)<br>Social media data (specify)<br>Other source (specify) |
| Data origin (select all that apply) | Government agency<br>Hospital<br>Internet search engine<br>Public dataset<br>Social network |

| | |
|---|---|
| | World Health Organization<br>Other source (specify) |
| **Methods** | |
| Did the study evaluate a data source for monitoring? | Yes<br>No<br>Not Specified<br>Unsure |
| Did the study predict a dengue-related outcome? | Yes<br>No<br>Not Specified<br>Unsure |
| What were the dengue-related outcomes? (specify) | |
| What algorithms and/or statistical models were used in the study? | |
| Did the study use a machine learning algorithm? | Yes<br>No<br>Not Specified<br>Unsure |
| If the study used a machine learning algorithm, did it use supervised learning? | Yes (specify)<br>No<br>Not Specified<br>Unsure |
| If the study used a machine learning algorithm, did it use unsupervised learning? | Yes (specify)<br>No<br>Not Specified<br>Unsure |
| If the study did not use a machine learning algorithm, did it use another type of model? | Yes (specify)<br>No<br>Not Specified<br>Unsure |
| Did the study use a natural language processing method? | Yes (specify)<br>No<br>Not Specified<br>Unsure |
| **Evaluation** | |
| Did the study use separate sets to train and test the algorithms? | Yes<br>No<br>Not Specified<br>Unsure<br>Not applicable |
| What were the evaluation metrics? (specify) | |

## PRISMA 2020 for Abstracts Checklist

| Section and Topic | Item # | Checklist item | Reported (Yes/No) |
|---|---|---|---|
| **TITLE** | | | |
| Title | 1 | Identify the report as a systematic review. | 1 |
| **BACKGROUND** | | | |
| Objectives | 2 | Provide an explicit statement of the main objective(s) or question(s) the review addresses. | 2 |
| **METHODS** | | | |
| Eligibility criteria | 3 | Specify the inclusion and exclusion criteria for the review. | 2 |
| Information sources | 4 | Specify the information sources (e.g. databases, registers) used to identify studies and the date when each was last searched. | 2 |
| Risk of bias | 5 | Specify the methods used to assess risk of bias in the included studies. | 2 |
| Synthesis of results | 6 | Specify the methods used to present and synthesise results. | 2 |
| **RESULTS** | | | |
| Included studies | 7 | Give the total number of included studies and participants and summarise relevant characteristics of studies. | 2 |
| Synthesis of results | 8 | Present results for main outcomes, preferably indicating the number of included studies and participants for each. If meta-analysis was done, report the summary estimate and confidence/credible interval. If comparing groups, indicate the direction of the effect (i.e. which group is favoured). | 2-3 |
| **DISCUSSION** | | | |
| Limitations of evidence | 9 | Provide a brief summary of the limitations of the evidence included in the review (e.g. study risk of bias, inconsistency and imprecision). | 3 |
| Interpretation | 10 | Provide a general interpretation of the results and important implications. | 3 |
| **OTHER** | | | |
| Funding | 11 | Specify the primary source of funding for the review. | NA |
| Registration | 12 | Provide the register name and registration number. | 2 |

*From:* Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71. doi: 10.1136/bmj.n71

For more information, visit: http://www.prisma-statement.org/

**S2 Table. Themes* associated with the included studies**

| Study main theme n (%) | Article n=77 | Conference paper n=42 |
|---|---|---|
| **Information Technology & Science** | **23 (30)** | **39 (93)** |
| Computer Science | 12 (16) | 30 (71) |
| Engineering | 5 (6) | 5 (12) |
| Science & Technology – Other Topics | 6 (8) | 4 (10) |
| | | |
| **Medicine** | **28 (36)** | **0 (0)** |
| Infectious Diseases & Tropical Medicine | 20 (26) | 0 (0) |
| Medicine – Other Topics | 8 (10) | 0 (0) |
| | | |
| **Health Informatics, Public Health & Biology** | **26 (34)** | **3 (7)** |
| Biology | 7 (9) | 0 (0) |
| Medical Informatics | 13 (17) | 3 (7) |
| Public Health | 6 (8) | 0 (0) |

*The themes are based on the "Research areas" classification from Web of Science or "Subject areas" from Scopus. Similar themes from both classifications have been aggregated.

**S2 Text. PROSPERO Protocol**

## Amendments to information provided at registration or in the protocol.

**Searches:**

Inclusion criteria: Publication date changed from 2019 to August 31 2020

**Intervention(s), exposure(s)**

We expanded the scope of data and included studies with real-world data in general (*Big Data* is a form of real-world data).

**Risk of bias (quality) assessment**

We expanded our method:

Each reviewer will independently assess the statistical models used in the study. In case of disagreement a third reviewer will help choose the right classification. Regarding prediction, when possible, risk of bias will be assessed using the Prediction model Risk Of Bias AssessmenT (PROBAST). If not (especially for machine learning methods) each reviewer will assess at least each of the following elements: predictor, outcome evaluation metrics.

PROSPERO
International prospective register of systematic reviews

# Harnessing Big Data and machine learning methods for dengue surveillance and prediction: a systematic review

*Emmanuelle Sylvestre, Marc Cuggia, André Cabié, Clarisse Joachim*



## Citation

Emmanuelle Sylvestre, Marc Cuggia, André Cabié, Clarisse Joachim. Harnessing Big Data and machine learning methods for dengue surveillance and prediction: a systematic review. PROSPERO 2020 CRD42020172472 Available from: https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42020172472

## Review question

1. What are the different Big data and machine learning methods used for dengue prediction?
2. What are the data sources used?
3. Are these methods reproducible?

## Searches

We will search articles in PubMed, Web of Science, Scopus and LILACS.
We will also search studies from grey literature, including the World Health Organization (WHO) Dengue Bulletin.
There will be no language restriction.
We will look for articles published between 2000 and 2019.

## Types of study to be included

We will include all epidemiological studies that use Big Data methods to predict: dengue outbreaks, dengue outcomes, dengue severity.
Studies with no original data (reviews, editorials, guidelines, perspective pieces), randomized controlled trials, case series and case reports will not be included.
Descriptive epidemiological studies without any prediction model will not be included.
Studies focusing on other types of arboviruses will not be included.
We will not include studies focusing exclusively on mosquitoes or in vitro studies.

## Condition or domain being studied

We will focus on Dengue virus (DENV) which is one of the most important vector-borne diseases in the world.
The majority of DENV infections are asymptomatic or are characterized by intense flu-like symptoms lasting up to 10 days afterward but they can evolve into the severe forms of dengue hemorrhagic fever/dengue shock syndrome (DHF/DSS) which can lead to death. However, mortality due to dengue can be greatly reduced by early diagnosis, which will influence appropriate clinical management.

Most dengue-endemic regions (South-East Asia, the Americas and the Pacific for the most seriously affected) rely on traditional surveillance, based on hospital syndromic monitoring and laboratory confirmation of a subset of cases reported to a central health agency. While this method is generally very accurate; it can be very slow and expensive due to the time needed to aggregate data, with substantial delays between an event and notifications
On the other hand, numerous studies have successfully used mobile, digital and Internet based systems to crowd-source data from the community. These new sources of data have been already used in pilot studies to improve monitoring and clinical management and predict dengue outbreaks.

**Participants/population**

We will include all people with dengue, regardless of age, gender or severity of the disease.

**Intervention(s), exposure(s)**

We are interested in studies using Big Data methods. According to the MeSH definition of Big Data, this means all methods applied on “extremely large amounts of data which require rapid and often complex computational analyses to reveal patterns, trends, and associations, relating to various facets of human and non-human entities”.
Regarding machine-learning methods, it can be defined as any computer-derived mathematical algorithm using learning to classify data. It includes:
- Supervised machine learning
- Unsupervised machine learning
- Deep learning

**Comparator(s)/control**

Not applicable

**Context**

The diagnostic of dengue in the included papers should be established using any of the standard WHO definition and classifications (1997 or 2009 WHO classification).
Research in low and middle-income countries will also be included.

**Main outcome(s)**

- Number and type of Big Data methods and/or machine learning models used to predict or forecast a dengue outbreak and their performance (Recall, Precision, F-measure)
- Number and type of Big Data methods and/or machine learning models used to predict or forecast a severe dengue outbreak and their performance (Recall, Precision, F-measure)

***Measures of effect***

Not applicable

**Additional outcome(s)**

None

***Measures of effect***

Not applicable

**Data extraction (selection and coding)**

Two authors from the review team will independently extract outcome data from each study using a Microsoft Excel collection form. In case of disagreement, a third reviewer will help to reach a consensus.

The data collection and extraction will proceed as follows.
1. Study screening based on title/abstract for each source
2. Removing duplicates
3. Eligibility based on full text review
4. Data extraction with collection form

The collection form will be based on relevant items from the Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies (CHARMS) checklist

Based on the inclusion criteria and the CHARMS cheklist, the data extracted will include:
1. Study characteristics (title, authors, year of publication, journal...)
2. Participants characteristics (age, gender...)
3. Data sources used for the models
4. Outcomes predicted
5. Machine learning models employed and role in the study
6. Models performance and evaluation

We will use Zotero for managing the search and writing the review.
The extracted data will be stored in Microsoft Excel individually.

**Risk of bias (quality) assessment**

Each reviewer will independently assess risk of bias using the Prediction model Risk Of Bias ASsessment Tool (PROBAST).

**Strategy for data synthesis**

We will summarise the results using descriptive statistics and a narrative synthesis.
For the narrative synthesis, we will identify common patterns and compile the results into sub categories:
- Most frequent methods used (and machine learning category)
- Performance and evaluation of each model
- Most frequent data sources used
- Contribution of non-clinical data versus clinical data
- Influence of study participants on the model performance (scientific outcomes?)

No meta-analysis will be conducted for this review.

**Analysis of subgroups or subsets**

If possible, synthesis will be stratified by:
- country income level (high vs low and middle-income)
- age (adults vs children)
- data sources (traditional data sources vs non traditional data sources)

**Contact details for further information**

Emmanuelle Sylvestre
emmanuelle.sylvestre@chu-martinique.fr

**Organisational affiliation of the review**

Martinique University Hospital

**Review team members and their organisational affiliations**

Dr Emmanuelle Sylvestre. Martinique University Hospital
Professor Marc Cuggia. Rennes University Hospital
Professor André Cabié. Martinique University Hospital
Assistant/Associate Professor Clarisse Joachim. Martinique University Hospital

**Collaborators**

Professor Raymond Césaire. Martinique University Hospital

**Type and method of review**

Narrative synthesis, Systematic review

**Anticipated or actual start date**

09 March 2020

**Anticipated completion date**

31 August 2020

**Funding sources/sponsors**

This work is financed through Martinique Univeristy Hospital and Rennes University

**Conflicts of interest**

**Language**

English

**Country**

France

**Stage of review**

Review Ongoing

**Subject index terms status**

Subject indexing assigned by CRD

**Subject index terms**

MeSH headings have not been applied to this record

**Date of registration in PROSPERO**

28 April 2020

**Date of first submission**

05 March 2020

**Stage of review at time of this submission**

The review has not started

| Stage | Started | Completed |
|---|---|---|
| Preliminary searches | No | No |

| Stage | Started | Completed |
|---|---|---|
| Piloting of the study selection process | No | No |
| Formal screening of search results against eligibility criteria | No | No |
| Data extraction | No | No |
| Risk of bias (quality) assessment | No | No |
| Data analysis | No | No |

*The record owner confirms that the information they have supplied for this submission is accurate and complete and they understand that deliberate provision of inaccurate information or omission of data may be construed as scientific misconduct.*

*The record owner confirms that they will update the status of the review when it is completed and will add publication details in due course.*

**Versions**

28 April 2020

**S3 Table. Characteristics of studies included in the systematic review**

| Reference | Study Type | Geographic region | Data sources* | Research period | Origin of data sources** |
|---|---|---|---|---|---|
| Polwiang et al. (2020)[1] | Article | Thailand | Epidemiological, Climate | 2003-2017 | Government agency |
| Xu et al. (2020)[2] | Article | China | Epidemiological, Climate | 2005-2018 | Government agency |
| Rangarajan et al. (2019)[3] | Article | Singapore, Taiwan, Thailand, Brazil, Mexico, USA | Epidemiological, Clinical, Google | 2001-2015 | Public dataset |
| Anno et al. (2019)[4] | Article | Taiwan | Epidemiological, Climate | 1998-2015 | Government agency |
| Romero et al. (2019)[5] | Article | | Epidemiological, Climate, Vector | 1960-2013/ 1986-2014 | Public dataset |
| Mello-Román et al. (2019)[6] | Article | Paraguay | Epidemiological, Clinical | 2012-2016 | Government agency |
| Stolerman et al. (2019)[7] | Article | Brazil | Epidemiological | 2002-2017 | Government agency |
| Macedo et al. (2019)[8] | Article | Brazil | Clinical | 2007-2013 | Hospital |
| Husnayain et al. (2019)[9] | Article | Indonesia | Epidemiological, Google | 2012-2016 | Government agency, ISE |
| Souza et al. (2019)[10] | Article | Brazil | Epidemiological, Twitter | 2015 | Government agency, SM |
| Ramadona et al. (2019)[11] | Article | Indonesia | Epidemiological, Twitter | 2016-2018 | Government agency, SM |
| Davi et al. (2019)[12] | Article | Brazil | Clinical, Genomic | | Hospital |
| Guo et al. (2019)[13] | Article | China | Epidemiological, Climate, Baidu, Weibo | 2011-2016 | Government agency, SM |
| Koh et al. (2018)[14] | Article | Singapore | Epidemiological, Climate | 2016 | Government agency |
| Carvajal et al. (2018)[15] | Article | Philippines | Epidemiological, Climate | 2009-2013 | Government agency |
| Baquero et al. (2018)[16] | Article | Brazil | Epidemiological, Climate | 2000-2016 | Government agency |
| Chen et al. (2018)[17] | Article | Thailand, Taiwan, Singapore | Epidemiological, Climate | 2003-2014 | Government agency |

*Epidemiological: Epidemiological and demographics data; Climate: Climate, environmental and geographical data; Clinical: Clinical and biological data; SM: Social Media data; Vector: Vector data; Genomic: Genomic data; Cellphone: Cellphone data

**ISE: Internet Search Engine; SM: Social Media; WHO: World Health Organization; IoT: Internet of Things

**S3 Table** *(Continued)*

| Reference | Study Type | Geographic region | Data sources* | Research period | Origin of data sources** |
|---|---|---|---|---|---|
| Villanes et al. (2018)[18] | Article | India | LeXisNexis | 2014 | News database |
| Guo et al. (2017)[19] | Article | China | Epidemiological, Climate, Baidu | 2011-2014 | Government agency |
| Chatterjee et al. (2018)[20] | Article | | Clinical, Genomic | | Public dataset |
| Guo et al. (2017)[21] | Article | China | Epidemiological, Baidu | 2011-2014 | Government agency, ISE |
| Yang et al. (2017)[22] | Article | Mexico, Taiwan, Thailand, Brazil, Singapore | Epidemiological, Google | 2001-2015 | Government agency |
| Marques-Toledo et al. (2017)[23] | Article | Brazil | Epidemiological, Twitter, Wikipedia | 2012-2016 | Government agency, SM |
| Premaratne et al. (2017)[24] | Article | Sri Lanka | Clinical | 2016 | Published dataset |
| Jayasundara et al. (2018)[25] | Article | Sri Lanka | Clinical | 2016 | Published dataset |
| Li et al. (2017)[26] | Article | China | Epidemiological, Climate, Baidu | 2011-2014 | Government agency, ISE |
| Kesorn et al. (2015)[27] | Article | Thailand | Epidemiological, Climate, Vector | 2007-2013 | Government agency |
| Dayama et al. (2014)[28] | Article | Singapore | Epidemiological | 2000-2014 | |
| Sampath et al. (2014)[29] | Article | Singapore | Epidemiological | | Government agency |
| Gluskin et al. (2014)[30] | Article | Mexico | Epidemiological, Climate, Google | 2003-2011 | Government agency, ISE |
| Flamand et al. (2014)[31] | Article | French Guiana | Epidemiological, Climate | 2006-2011 | Government agency |
| Torres et al. (2014)[32] | Article | Colombia | Epidemiological | 1995-2011/ 1997-2011 | Government agency |
| Buczak et al. (2012)[33] | Article | Peru | Epidemiological, Climate, Political stability | 2001-2009 | Government agency |

*Epidemiological: Epidemiological and demographics data; Climate: Climate, environmental and geographical data; Clinical: Clinical and biological data; SM: Social Media data; Vector: Vector data; Genomic: Genomic data; Cellphone: Cellphone data

**ISE: Internet Search Engine; SM: Social Media; WHO: World Health Organization; IoT: Internet of Things

**S3 Table** *(Continued)*

| Reference | Study Type | Geographic region | Data sources* | Research period | Origin of data sources** |
|---|---|---|---|---|---|
| Hoen et al. (2012)[34] | Article | | Epidemiological, Climate, HealthMap | 2009-2011 | Government agency, News database |
| Althouse et al. (2011)[35] | Article | Singapore, Thailand | Epidemiological, Google | 2004-2011 | Government agency |
| Chan et al. (2011)[36] | Article | Bolivia, India, Indonesia, Brazil, Singapore | Epidemiological, Google | 2003-2010 | WHO, ISE |
| Faisal et al. (2012)[37] | Article | Malaysia | Clinical, Bioelectrical impedance analysis | 2010 | Hospital |
| Ibrahim et al. (2010)[38] | Article | Malaysia | Clinical, Bioelectrical impedance analysis | 2010 | Hospital |
| Syamsuddin et al. (2020)[39] | Article | Indonesia | Epidemiological, Google | 2012-2017 | Hospital, ISE |
| Romero-Alvarez et al. (2020)[40] | Article | Brazil | Epidemiological, Google | 2011-2016 | ISE, Government agency |
| Liu et al. (2019)[41] | Article | China | Epidemiological, Climate, Baidu | 2011-2015 | Government agency, ISE |
| Musa et al. (2019)[42] | Article | Taiwan | Epidemiological | 2014–2016 | |
| Messina et al. (2019)[43] | Article | Worldwide | HealthMap | 1960-2015 | HealthMap |
| Titus et al. (2018)[44] | Article | Bangladesh | Epidemiological, Climate | 2000-2009 | Government agency |
| Marques-Toledo et al. (2019)[45] | Article | Brazil | Epidemiological, Climate, Vector, Twitter | 2015 | Government agency, SM |
| Verma et al. (2018)[46] | Article | India | Epidemiological, Google | 2016 | Government agency, ISE |
| Ho et al. (2018)[47] | Article | Philippines | Epidemiological, Google | 2009-2014 | Government agency, ISE |
| Phakhounthong et al. (2018)[48] | Article | Cambodia | Clinical | 2009 | Hospital |
| Strauss et al. (2017)[49] | Article | Venezuela | Epidemiological, Google | 2004-2014 | Government agency |
| Nsoesie et al. (2016)[50] | Article | Brazil | Epidemiological, Twitter | 2012-2014 | Government agency, SM |

*Epidemiological: Epidemiological and demographics data; Climate: Climate, environmental and geographical data; Clinical: Clinical and biological data; SM: Social Media data; Vector: Vector data; Genomic: Genomic data; Cellphone: Cellphone data

**ISE: Internet Search Engine; SM: Social Media; WHO: World Health Organization; IoT: Internet of Things

**S3 Table** *(Continued)*

| Reference | Study Type | Geographic region | Data sources* | Research period | Origin of data sources** |
|---|---|---|---|---|---|
| Liu et al. (2016)[51] | Article | China | Epidemiological, Baidu | 2010-2014 | Government agency, ISE |
| Ximenes et al. (2016)[52] | Article | Brazil | Epidemiological | 2000-2015 | Government agency |
| Mohamad et al. (2014)[53] | Article | Malaysia | Epidemiological, Climate | 2003-2009 | Government agency, Hospital |
| Puengpreeda, et al. (2020)[54] | Article | Thailand | Epidemiological, Climate, Google | 2007-2018 | Government agency, ISE |
| Amin, et al. (2020)[55] | Article | Worldwide | Twitter | 2017-2019 | SM |
| Manogaran, et al. (2018)[56] | Article | India | Epidemiological, Climate | 1979-2016 | Government agency<br>IoT Weather sensor device |
| Agarwal, et al. (2018)[57] | Article | India | Epidemiological, Climate | 2006-2015 | Government agency |
| Manogaran, et al. (2018)[58] | Article | India | Epidemiological, Climate | 1998-2016 | Government agency, Public dataset |
| Jahangir, et al. (2018)[59] | Conference Paper | Pakistan | Clinical | | Hospital |
| Husin, et al. (2018)[60] | Conference Paper | Malaysia | Clinical | 2015 | Hospital, Experts |
| Anggraeni, et al. (2018)[61] | Conference Paper | Indonesia | Epidemiological, Google | 2010-2015 | Hospital, ISE |
| Livelo, et al. (2018)[62] | Conference Paper | Philippines | Epidemiological, Twitter | 2017 | Government agency, SM |
| Wiratmadja, et al. (2018)[63] | Article | Indonesia | Clinical | | Hospital |
| Arafiyah, et al. (2018)[64] | Conference Paper | Indonesia | Clinical | 2017 | Hospital |
| Abuhamad, et al. (2017)[65] | Article | Malaysia | Epidemiological, Clinical, Climate | 2003-2010 | Government agency |
| Manivannan, et al. (2017)[66] | Conference Paper | Vietnam | Clinical | 2010-2013 | Hospital |

*Epidemiological: Epidemiological and demographics data; Climate: Climate, environmental and geographical data; Clinical: Clinical and biological data; SM: Social Media data; Vector: Vector data; Genomic: Genomic data; Cellphone: Cellphone data

**ISE: Internet Search Engine; SM: Social Media; WHO: World Health Organization; IoT: Internet of Things

**S3 Table** *(Continued)*

| Reference | Study Type | Geographic region | Data sources* | Research period | Origin of data sources** |
|---|---|---|---|---|---|
| Dharmawardana, et al. (2017)[67] | Conference Paper | Sri Lanka | Epidemiological, Climate, Cellphone | 2012-2014 | Government agency, Cell towers |
| Espina, et al. (2017)[68] | Conference Paper | Philippines | Epidemiological, Twitter | 2016 | Government agency, SM |
| Rahim, et al. (2017)[69] | Article | Malaysia | Epidemiological, Climate | 2010-2015 | Government agency |
| Klein, et al. (2017)[70] | Article | Brazil | Epidemiological, Facebook, Twitter, Instagram, Flickr, YouTube | 2016 | Government agency, SM |
| Kerdprasop, et al. (2016)[71] | Conference Paper | Thailand | Epidemiological, Climate | 2006-2015 | Government agency |
| Anggraeni, et al. (2016)[72] | Conference Paper | Indonesia | Epidemiological, Google | 2010-2015 | Hospital, ISE |
| Mathulamuthu, et al. (2016)[73] | Conference Paper | Malaysia | Epidemiological, Climate | 2009-2013 | Government agency |
| Rahmawati, et al. (2016)[74] | Conference Paper | Taiwan | Epidemiological, Climate | 2014-2015 | Government agency |
| Missier, et al. (2016)[75] | Conference Paper | Brazil | Epidemiological, Twitter | 2015 | Government agency, SM |
| Abeyrathna, et al. (2016)[76] | Conference Paper | Sri Lanka | Epidemiological, Cellphone | 2013 | Government agency, Cell towers |
| Fathima, et al. (2015)[77] | Article | India | Clinical | 2009-2011 | Hospital |
| Tazkia, et al. (2015)[78] | Conference Paper | Indonesia | Epidemiological, Climate | 2012-2013 | Government agency |
| Wu, et al. (2008)[79] | Conference Paper | Singapore | Epidemiological, Climate | 2001-2006 | Government agency |
| Salam, et al. (2019)[80] | Article | India | Epidemiological, Google | 2004-2017 | Government agency, ISE |
| Saire, et al. (2019)[81] | Conference Paper | Brazil | Twitter | 2009-2017 | SM |

*Epidemiological: Epidemiological and demographics data; Climate: Climate, environmental and geographical data; Clinical: Clinical and biological data; SM: Social Media data; Vector: Vector data; Genomic: Genomic data; Cellphone: Cellphone data

**ISE: Internet Search Engine; SM: Social Media; WHO: World Health Organization; IoT: Internet of Things

**S3 Table** *(Continued)*

| Reference | Study Type | Geographic region | Data sources* | Research period | Origin of data sources** |
|---|---|---|---|---|---|
| Swain, et al. (2017)[82] | Conference Paper | India | Epidemiological, Twitter | 2016 | SM |
| Saravanan, et al. (2017)[83] | Conference Paper | India | Epidemiological, Clinical | | Hospital |
| Carlos, et al. (2017)[84] | Conference Paper | Brazil | Epidemiological, Twitter | 2015-2016 | Government agency, SM |
| Ye, et al. (2016)[85] | Article | China | Epidemiological, Twitter, Weibo | 2014 | Government agency, SM |
| Li, et al. (2016)[86] | Conference Paper | China | Epidemiological, Climate | 2015 | Government agency |
| Fathima, et al. (2011)[87] | Conference Paper | India | Clinical | | Hospital |
| Srilekha et al. (2020)[88] | Article | Sri Lanka | Epidemiological, Climate | 18 years | |
| Ganthimathi et al. (2020)[89] | Article | India | Clinical | | Hospital |
| Kumar et al. (2020)[90] | Article | India | Clinical | 2017 | Hospital |
| Guiyab et al. (2019)[91] | Article | Philippines | Clinical | 2013-2017 | Hospital |
| Chovatiya et al. (2019)[92] | Conference Paper | India | Epidemiological, Climate | | Government agency |
| Kerdprasop et al. (2019)[93] | Conference Paper | Thailand | Epidemiological, Climate | 2003-2017 | Government agency |
| Link et al. (2019)[94] | Conference Paper | Peru, Puerto Rico | Epidemiological, Climate | 2002-2009 | Public dataset |
| Arafiyah et al. (2018)[95] | Conference Paper | Indonesia | Clinical | 2017 | Hospital |
| Mishra et al. (2018)[96] | Article | India | Clinical | | Hospital |

*Epidemiological: Epidemiological and demographics data; Climate: Climate, environmental and geographical data; Clinical: Clinical and biological data; SM: Social Media data; Vector: Vector data; Genomic: Genomic data; Cellphone: Cellphone data

**ISE: Internet Search Engine; SM: Social Media; WHO: World Health Organization; IoT: Internet of Things

**S3 Table** *(Continued)*

| Reference | Study Type | Geographic region | Data sources | Research period | Origin of data sources** |
|---|---|---|---|---|---|
| Wu et al. (2017)[97] | Conference Paper | Taiwan | Epidemiological, Climate, Google | 2006-2016 | Government agency WHO, ISE |
| Albinati et al. (2017)[98] | Conference Paper | Brazil | Epidemiological, Twitter | 2011-2016 | Government agency, SM |
| Zhu et al. (2017)[99] | Conference Paper | Hong-Kong | Epidemiological, Climate | 2004-2015 | Government agency |
| Zainudin et al. (2016)[100] | Article | Malaysia | Epidemiological | 2010-2015 | Government agency |
| Milinovich et al. (2014)[101] | Article | Australia | Epidemiological, Google | 2004-2013 | Government agency, ISE |
| Ongruk et al. (2014)[102] | Conference Paper | Thailand | Epidemiological, Climate, Vector | 2010-2012 | Government agency |
| Balasundaram et al. (2013)[103] | Conference Paper | India | Clinical | | Hospital |
| Wu et al. (2009)[104] | Conference Paper | Singapore | Epidemiological, Climate | 2001-2007 | Government agency |
| Zhang et al. (2020)[105] | Article | India | Epidemiological, LeXisNexis | 2003-2016 | WHO, News database |
| Souza et al. (2018)[106] | Conference Paper | Brazil | SM | 2015 | SM |
| Coberly et al. (2014)[107] | Article | Philippines | Epidemiological, Twitter | 2011 | Government agency, SM |
| Gomide et al. (2011)[108] | Conference Paper | Brazil | Epidemiological, Twitter | 2007-2010 | Government agency, SM |
| Fang et al. (2010)[109] | Conference Paper | Taiwan | Epidemiological, Google | 2004-2008 | Government agency |
| Souza et al. (2020)[110] | Conference Paper | Puerto Rico | Epidemiological, Climate, Google | 2004-2008 | Public dataset |
| Yogapriya et al. (2019)[111] | Article | India | Clinical | | Hospital |

*Epidemiological: Epidemiological and demographics data; Climate: Climate, environmental and geographical data; Clinical: Clinical and biological data; SM: Social Media data; Vector: Vector data; Genomic: Genomic data; Cellphone: Cellphone data

**ISE: Internet Search Engine; SM: Social Media; WHO: World Health Organization; IoT: Internet of Things

**S3 Table** *(Continued)*

| Reference | Study Type | Geographic region | Data sources* | Research period | Origin of data sources** |
|---|---|---|---|---|---|
| Adias et al. (2019)[112] | Conference Paper | Indonesia | Clinical | 2010-2015 | Hospital |
| Jongmuenwai et al. (2019)[113] | Conference Paper | Thailand | Epidemiological, Clinical, Climate | 2007-2016 | Government agency, Hospital |
| Balasaravanan et al. (2018)[114] | Article | India | Clinical | | Government agency |
| Acosta et al. (2016)[115] | Article | Cuba | Clinical | 2014 | Hospital |
| Soonthornphisaj et al. (2016)[116] | Conference Paper | Thailand | Clinical | | Hospital |
| Fathima et al. (2012)[117] | Conference Paper | India | Clinical | | Hospital |
| Fathima et al. (2011)[118] | Conference Paper | India | Clinical | | Hospital |
| Long et al. (2010)[119] | Conference Paper | Malaysia | Epidemiological | | Government agency |

*Epidemiological: Epidemiological and demographics data; Climate: Climate, environmental and geographical data; Clinical: Clinical and biological data; SM: Social Media data; Vector: Vector data; Genomic: Genomic data; Cellphone: Cellphone data

**ISE: Internet Search Engine; SM: Social Media; WHO: World Health Organization; IoT: Internet of Things

## References


1. Polwiang S. The time series seasonal patterns of dengue fever and associated weather variables in Bangkok (2003-2017). BMC Infect Dis. 2020;20: 208. doi:10.1186/s12879-020-4902-6
2. Xu J, Xu K, Li Z, Meng F, Tu T, Xu L, et al. Forecast of Dengue Cases in 20 Chinese Cities Based on the Deep Learning Method. Int J Environ Res Public Health. 2020;17. doi:10.3390/ijerph17020453
3. Rangarajan P, Mody SK, Marathe M. Forecasting dengue and influenza incidences using a sparse representation of Google trends, electronic health records, and time series data. PLoS Comput Biol. 2019;15: e1007518. doi:10.1371/journal.pcbi.1007518
4. Anno S, Hara T, Kai H, Lee M-A, Chang Y, Oyoshi K, et al. Spatiotemporal dengue fever hotspots associated with climatic factors in Taiwan including outbreak predictions based on machine-learning. Geospat Health. 2019;14. doi:10.4081/gh.2019.771

5. Romero D, Olivero J, Real R, Guerrero JC. Applying fuzzy logic to assess the biogeographical risk of dengue in South America. Parasit Vectors. 2019;12: 428. doi:10.1186/s13071-019-3691-5

6. Mello-Román JD, Mello-Román JC, Gómez-Guerrero S, García-Torres M. Predictive Models for the Medical Diagnosis of Dengue: A Case Study in Paraguay. Comput Math Methods Med. 2019;2019: 7307803. doi:10.1155/2019/7307803

7. Stolerman LM, Maia PD, Kutz JN. Forecasting dengue fever in Brazil: An assessment of climate conditions. PLoS One. 2019;14: e0220106. doi:10.1371/journal.pone.0220106

8. Macedo Hair G, Fonseca Nobre F, Brasil P. Characterization of clinical patterns of dengue patients using an unsupervised machine learning approach. BMC Infect Dis. 2019;19: 649. doi:10.1186/s12879-019-4282-y

9. Husnayain A, Fuad A, Lazuardi L. Correlation between Google Trends on dengue fever and national surveillance report in Indonesia. Glob Health Action. 2019;12: 1552652. doi:10.1080/16549716.2018.1552652

10. Souza RCSNP, Assunção RM, Oliveira DM, Neill DB, Meira W. Where did I get dengue? Detecting spatial clusters of infection risk with social network data. Spat Spatiotemporal Epidemiol. 2019;29: 163–175. doi:10.1016/j.sste.2018.11.005

11. Ramadona AL, Tozan Y, Lazuardi L, Rocklöv J. A combination of incidence data and mobility proxies from social media predicts the intra-urban spread of dengue in Yogyakarta, Indonesia. PLoS Negl Trop Dis. 2019;13: e0007298. doi:10.1371/journal.pntd.0007298

12. Davi C, Pastor A, Oliveira T, Neto FB de L, Braga-Neto U, Bigham AW, et al. Severe Dengue Prognosis Using Human Genome Data and Machine Learning. IEEE Trans Biomed Eng. 2019;66: 2861–2868. doi:10.1109/TBME.2019.2897285

13. Guo P, Zhang Q, Chen Y, Xiao J, He J, Zhang Y, et al. An ensemble forecast model of dengue in Guangzhou, China using climate and social media surveillance data. Sci Total Environ. 2019;647: 752–762. doi:10.1016/j.scitotenv.2018.08.044

14. Koh Y-M, Spindler R, Sandgren M, Jiang J. A model comparison algorithm for increased forecast accuracy of dengue fever incidence in Singapore and the auxiliary role of total precipitation information. Int J Environ Health Res. 2018;28: 535–552. doi:10.1080/09603123.2018.1496234

15. Carvajal TM, Viacrusis KM, Hernandez LFT, Ho HT, Amalin DM, Watanabe K. Machine learning methods reveal the temporal pattern of dengue incidence using meteorological factors in metropolitan Manila, Philippines. BMC Infect Dis. 2018;18: 183. doi:10.1186/s12879-018-3066-0

16. Baquero OS, Santana LMR, Chiaravalloti-Neto F. Dengue forecasting in São Paulo city with generalized additive models, artificial neural networks and seasonal autoregressive integrated moving average models. PLoS One. 2018;13: e0195065. doi:10.1371/journal.pone.0195065

17. Chen Y, Chu CW, Chen MIC, Cook AR. The utility of LASSO-based models for real time forecasts of endemic infectious diseases: A cross country comparison. J Biomed Inform. 2018;81: 16–30. doi:10.1016/j.jbi.2018.02.014

18. Villanes A, Griffiths E, Rappa M, Healey CG. Dengue Fever Surveillance in India Using Text Mining in Public Media. Am J Trop Med Hyg. 2018;98: 181–191. doi:10.4269/ajtmh.17-0253

19. Guo P, Liu T, Zhang Q, Wang L, Xiao J, Zhang Q, et al. Developing a dengue forecast model using machine learning: A case study in China. PLoS Negl Trop Dis. 2017;11: e0005973. doi:10.1371/journal.pntd.0005973

20. Chatterjee S, Dey N, Shi F, Ashour AS, Fong SJ, Sen S. Clinical application of modified bag-of-features coupled with hybrid neural-based classifier in dengue fever classification using gene expression data. Med Biol Eng Comput. 2018;56: 709–720. doi:10.1007/s11517-017-1722-y

21. Guo P, Wang L, Zhang Y, Luo G, Zhang Y, Deng C, et al. Can internet search queries be used for dengue fever surveillance in China? Int J Infect Dis. 2017;63: 74–76. doi:10.1016/j.ijid.2017.08.001

22. Yang S, Kou SC, Lu F, Brownstein JS, Brooke N, Santillana M. Advances in using Internet searches to track dengue. PLoS Comput Biol. 2017;13: e1005607. doi:10.1371/journal.pcbi.1005607

23. Marques-Toledo C de A, Degener CM, Vinhal L, Coelho G, Meira W, Codeço CT, et al. Dengue prediction by the web: Tweets are a useful tool for estimating and forecasting Dengue at country and city level. PLOS Neglected Tropical Diseases. 2017;11: e0005729. doi:10.1371/journal.pntd.0005729

24. Premaratne MK, Perera SSN, Malavige GN, Jayasinghe S. Mathematical Modelling of Immune Parameters in the Evolution of Severe Dengue. Comput Math Methods Med. 2017;2017: 2187390. doi:10.1155/2017/2187390

25. Jayasundara SDP, Perera SSN, Malavige GN, Jayasinghe S. Mathematical modelling and a systems science approach to describe the role of cytokines in the evolution of severe dengue. BMC Syst Biol. 2017;11: 34. doi:10.1186/s12918-017-0415-3

26. Li Z, Liu T, Zhu G, Lin H, Zhang Y, He J, et al. Dengue Baidu Search Index data can improve the prediction of local dengue epidemic: A case study in Guangzhou, China. PLoS Negl Trop Dis. 2017;11: e0005354. doi:10.1371/journal.pntd.0005354

27. Kesorn K, Ongruk P, Chompoosri J, Phumee A, Thavara U, Tawatsin A, et al. Morbidity Rate Prediction of Dengue Hemorrhagic Fever (DHF) Using the Support Vector Machine and the Aedes aegypti Infection Rate in Similar Climates and Geographical Areas. PLoS One. 2015;10: e0125049. doi:10.1371/journal.pone.0125049

28. Dayama P, Sampath K. Dengue disease outbreak detection. Stud Health Technol Inform. 2014;205: 1105–1109.

29. Sampath K, Dayama P. Predicting the operations alert levels for dengue surveillance and control. Stud Health Technol Inform. 2014;205: 1100–1104.

30. Gluskin RT, Johansson MA, Santillana M, Brownstein JS. Evaluation of Internet-based dengue query data: Google Dengue Trends. PLoS Negl Trop Dis. 2014;8: e2713. doi:10.1371/journal.pntd.0002713

31. Flamand C, Fabregue M, Bringay S, Ardillon V, Quénel P, Desenclos J-C, et al. Mining local climate data to assess spatiotemporal dengue fever epidemic patterns in French Guiana. J Am Med Inform Assoc. 2014;21: e232-240. doi:10.1136/amiajnl-2013-002348

32. Torres C, Barguil S, Melgarejo M, Olarte A. Fuzzy model identification of dengue epidemic in Colombia based on multiresolution analysis. Artif Intell Med. 2014;60: 41–51. doi:10.1016/j.artmed.2013.11.008

33. Buczak AL, Koshute PT, Babin SM, Feighner BH, Lewis SH. A data-driven epidemiological prediction method for dengue outbreaks using local and remote sensing data. BMC Med Inform Decis Mak. 2012;12: 124. doi:10.1186/1472-6947-12-124

34. Hoen AG, Keller M, Verma AD, Buckeridge DL, Brownstein JS. Electronic event-based surveillance for monitoring dengue, Latin America. Emerg Infect Dis. 2012;18: 1147–1150. doi:10.3201/eid1807.120055

35. Althouse BM, Ng YY, Cummings DAT. Prediction of dengue incidence using search query surveillance. PLoS Negl Trop Dis. 2011;5: e1258. doi:10.1371/journal.pntd.0001258

36. Chan EH, Sahai V, Conrad C, Brownstein JS. Using web search query data to monitor dengue epidemics: a new model for neglected tropical disease surveillance. PLoS Negl Trop Dis. 2011;5: e1206. doi:10.1371/journal.pntd.0001206

37. Faisal T, Taib MN, Ibrahim F. Neural network diagnostic system for dengue patients risk classification. J Med Syst. 2012;36: 661–676. doi:10.1007/s10916-010-9532-x

38. Ibrahim F, Faisal T, Salim MIM, Taib MN. Non-invasive diagnosis of risk in dengue patients using bioelectrical impedance analysis and artificial neural network. Med Biol Eng Comput. 2010;48: 1141–1148. doi:10.1007/s11517-010-0669-z

39. Syamsuddin M, Fakhruddin M, Sahetapy-Engel JTM, Soewono E. Causality Analysis of Google Trends and Dengue Incidence in Bandung, Indonesia With Linkage of Digital Data Modeling: Longitudinal Observational Study. J Med Internet Res. 2020;22: e17633. doi:10.2196/17633

40. Romero-Alvarez D, Parikh N, Osthus D, Martinez K, Generous N, Del Valle S, et al. Google Health Trends performance reflecting dengue incidence for the Brazilian states. BMC Infect Dis. 2020;20: 252. doi:10.1186/s12879-020-04957-0

41. Liu D, Guo S, Zou M, Chen C, Deng F, Xie Z, et al. A dengue fever predicting model based on Baidu search index data and climate data in South China. PLoS One. 2019;14: e0226841. doi:10.1371/journal.pone.0226841

42. Musa SS, Zhao S, Chan H-S, Jin Z, He DH. A mathematical model to study the 2014-2015 large-scale dengue epidemics in Kaohsiung and Tainan cities in Taiwan, China. Math Biosci Eng. 2019;16: 3841–3863. doi:10.3934/mbe.2019190

43. Messina JP, Brady OJ, Golding N, Kraemer MUG, Wint GRW, Ray SE, et al. The current and future global distribution and population at risk of dengue. Nat Microbiol. 2019;4: 1508–1515. doi:10.1038/s41564-019-0476-8

44. Titus Muurlink O, Stephenson P, Islam MZ, Taylor-Robinson AW. Long-term predictors of dengue outbreaks in Bangladesh: A data mining approach. Infect Dis Model. 2018;3: 322–330. doi:10.1016/j.idm.2018.11.004

45. Marques-Toledo CA, Bendati MM, Codeço CT, Teixeira MM. Probability of dengue transmission and propagation in a non-endemic temperate area: conceptual model and decision risk levels for early alert, prevention and control. Parasit Vectors. 2019;12: 38. doi:10.1186/s13071-018-3280-z

46. Verma M, Kishore K, Kumar M, Sondh AR, Aggarwal G, Kathirvel S. Google Search Trends Predicting Disease Outbreaks: An Analysis from India. Healthc Inform Res. 2018;24: 300–308. doi:10.4258/hir.2018.24.4.300

47. Ho CC, Ting C-Y, Raja DB. Using Public Open Data to Predict Dengue Epidemic: Assessment of Weather Variability, Population Density, and Land use as Predictor Variables for Dengue Outbreak Prediction using Support Vector Machine. Indian Journal of Science and Technology. 2018;11. doi:10.17485/ijst/2018/v11i4/115405

48. Phakhounthong K, Chaovalit P, Jittamala P, Blacksell SD, Carter MJ, Turner P, et al. Predicting the severity of dengue fever in children on admission based on clinical features and laboratory indicators: application of classification tree analysis. BMC Pediatr. 2018;18: 109. doi:10.1186/s12887-018-1078-y

49. Strauss RA, Castro JS, Reintjes R, Torres JR. Google dengue trends: An indicator of epidemic behavior. The Venezuelan Case. Int J Med Inform. 2017;104: 26–30. doi:10.1016/j.ijmedinf.2017.05.003

50. Nsoesie EO, Flor L, Hawkins J, Maharana A, Skotnes T, Marinho F, et al. Social Media as a Sentinel for Disease Surveillance: What Does Sociodemographic Status Have to Do with It? PLoS Curr. 2016;8. doi:10.1371/currents.outbreaks.cc09a42586e16dc7dd62813b7ee5d6b6

51. Liu K, Wang T, Yang Z, Huang X, Milinovich GJ, Lu Y, et al. Using Baidu Search Index to Predict Dengue Outbreak in China. Sci Rep. 2016;6: 38040. doi:10.1038/srep38040

52. Ximenes R, Amaku M, Lopez LF, Coutinho FAB, Burattini MN, Greenhalgh D, et al. The risk of dengue for non-immune foreign visitors to the 2016 summer olympic games in Rio de Janeiro, Brazil. BMC Infect Dis. 2016;16: 186. doi:10.1186/s12879-016-1517-z

53. Mohamad Mohsin MF, Abu Bakar A, Hamdan AR. Outbreak detection model based on danger theory. Appl Soft Comput. 2014;24: 612–622. doi:10.1016/j.asoc.2014.08.030

54. Puengpreeda A, Yhusumrarn S, Sirikulvadhana S. Weekly Forecasting Model for Dengue Hemorrhagic Fever Outbreak in Thailand. Eng J-Thail. 2020;24: 71–87. doi:10.4186/ej.2020.24.3.71

55. Amin S, Uddin MI, Hassan S, Khan A, Nasser N, Alharbi A, et al. Recurrent Neural Networks With TF-IDF Embedding Technique for Detection and Classification in Tweets of Dengue Disease. IEEE Access. 2020;8: 131522–131533. doi:10.1109/ACCESS.2020.3009058

56. Manogaran G, Lopez D, Chilamkurti N. In-Mapper combiner based MapReduce algorithm for processing of big climate data. Futur Gener Comp Syst. 2018;86: 433–445. doi:10.1016/j.future.2018.02.048

57. Agarwal N, Koti SR, Saran S, Kumar AS. Data mining techniques for predicting dengue outbreak in geospatial domain using weather parameters for New Delhi, India. Curr Sci. 2018;114: 2281–2291. doi:10.18520/cs/v114/i11/2281-2291

58. Manogaran G, Lopez D. A Gaussian process based big data processing framework in cluster computing environment. Cluster Comput. 2018;21: 189–204. doi:10.1007/s10586-017-0982-5

59. Jahangir I, Abdul-Basit, Hannan A, Javed S. Prediction of Dengue Disease Through Data Mining by Using Modified Apriori Algorithm. Proceedings of the 4th Acm International Conference of Computing for Engineering and Sciences (icces'2018). New York: Assoc Computing Machinery; 2018. doi:10.1145/3213187.3287612

60. Husin NA, Alharogi A, Mustapha N, Hamdan H, Husin UA. Early Self-Diagnosis of Dengue Symptoms Using Fuzzy and Data Mining Approach. In: Nifa F a. A, Lin CK, Hussain A, editors. Proceedings of the 3rd International Conference on Applied Science and Technology (icast'18). Melville: Amer Inst Physics; 2018. p. 020048. doi:10.1063/1.5055450

61. Anggraeni W, Pramudita G, Riksakomara E, Radityo PW, Samopa F, Pujiadi, et al. Artificial Neural Network for Health Data Forecasting, Case Study: Number of Dengue Hemorrhagic Fever Cases in Malang Regency, Indonesia. 2018 International Conference on Electrical Engineering and Computer Science (icecos). New York: Ieee; 2018. pp. 207–212.

62. Dennison Livelo E, Cheng C. Intelligent Dengue Infoveillance Using Gated Recurrent Neural Learning and Cross-Label Frequencies. 2018 Ieee International Conference on Agents (ica). New York: Ieee; 2018. pp. 2–7.

63. Wiratmadja II, Salamah SY, Govindaraju R. Healthcare Data Mining: Predicting Hospital Length of Stay of Dengue Patients. J Eng Technol Sci. 2018;50: 110–126. doi:10.5614/j.eng.technol.sci.2018.50.1.8

64. Arafiyah R, Hermin F. Data mining for dengue hemorrhagic fever (DHF) prediction with naive Bayes method. 1st International Conference of Education on Sciences, Technology, Engineering, and Mathematics (ice-Stem). Bristol: Iop Publishing Ltd; 2018. p. 012077. doi:10.1088/1742-6596/948/1/012077

65. Abuhamad HIS, Abu Bakar A, Zainudin S, Sahani M, Ali ZM. Feature Selection Algorithms for Malaysian Dengue Outbreak Detection Model. Sains Malays. 2017;46: 255–265. doi:10.17576/jsm-2017-4602-10

66. Manivannan P, Devi PI. Dengue Fever Prediction Using K-Means Clustering Algorithm. 2017 Ieee International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (incos). New York: Ieee; 2017.

67. Dharmawardana KGS, Lokuge JN, Dassanayake PSB, Sirisena ML, Fernando ML, Perera AS, et al. Predictive Model for the Dengue Incidences in Sri Lanka Using Mobile Network Big Data. 2017 Ieee International Conference on Industrial and Information Systems (iciis). New York: Ieee; 2017. pp. 278–283.

68. Espina K, Estuar MRJE. Infodemiology for Syndromic Surveillance of Dengue and Typhoid Fever in the Philippines. In: CruzCunha MM, Varajao JEQ, Rijo R, Martinho R, Peppard J, SanCristobal JR, et al., editors. Centeris 2017 - International Conference on Enterprise Information Systems / Projman 2017 - International Conference on Project Management / Hcist 2017 - International Conference on Health and Social Care Information Systems and Technologies, Centeri. Amsterdam: Elsevier Science Bv; 2017. pp. 554–561. doi:10.1016/j.procs.2017.11.073

69. Rahim NF, Taib SM, Abidin AIZ. Dengue Fatality Prediction Using Data Mining. J Fundam Appl Sci. 2017;9: 671–683. doi:10.4314/jfas.v9i6s.52

70. Klein GH, Neto PG, Tezza R. Big Data and social media: surveillance of networks as management tool. Saude Soc. 2017;26: 208–217. doi:10.1590/S0104-12902017164943

71. Kerdprasop N, Kerdprasop K. Remote Sensing Based Modeling of Dengue Outbreak with Regression and Binning Classification. 2016 2nd Ieee International Conference on Computer and Communications (iccc). New York: Ieee; 2016. pp. 46–49.

72. Anggraeni W, Aristiani L. Using Google Trend Data in Forecasting Number of Dengue Fever Cases with ARIMAX Method Case Study : Surabaya, Indonesia. Proceedings of 2016 International Conference on Information & Communication Technology and Systems (icts). New York: Ieee; 2016. pp. 114–118.

73. Mathulamuthu SS, Asirvadam VS, Dass SC, Gill BS, Loshini T. Predicting Dengue Incidences Using Cluster Based Regression on Climate Data. 2016 6th Ieee International Conference on Control System, Computing and Engineering (iccsce). New York: Ieee; 2016. pp. 245–250.

74. Rahmawati D, Huang Y-P. Using C-support Vector Classification to Forecast Dengue Fever Epidemics in Taiwan. In: Wang WJ, Lee PJ, Er MJ, Jeng JT, editors. 2016 International Conference on System Science and Engineering (icsse). New York: Ieee; 2016.

75. Missier P, Romanovsky A, Miu T, Pal A, Daniilakis M, Garcia A, et al. Tracking Dengue Epidemics Using Twitter Content Classification and Topic Modelling. In: Casteleyn S, Dolog P, Pautasso C, editors. Current Trends in Web Engineering, Icwe 2016 International Workshops. Cham: Springer International Publishing Ag; 2016. pp. 80–92. doi:10.1007/978-3-319-46963-8_7

76. Abeyrathna MP a. R, Abeygunawrdane DA, Wijesundara R a. a. V, Mudalige VB, Bandara M, Perera S, et al. Dengue Propagation Prediction using Human Mobility. 2nd International Mercon 2016 Moratuwa Engineering Research Conference. New York: Ieee; 2016. pp. 156–161.

77. Fathima AS, Manimeglai D. Analysis of Significant Factors for Dengue Infection Prognosis Using the Random Forest Classifier. Int J Adv Comput Sci Appl. 2015;6: 240–245.

78. Tazkia RAK, Narita V, Nugroho AS. Dengue Outbreak Prediction for GIS based Early Warning System. 2015 International Conference on Science in Information Technology (ICSITech). New York: Ieee; 2015. pp. 121–125.

79. Wu Y, Lee G, Fu X, Hung T. Detect climatic factors contributing to dengue outbreak based on wavelet, support vector machines and genetic algorithm. In: Ao SI, Gelman L, Hukins DWL, Hunter A, Korsunsky AM, editors. World Congress on Engineering 2008, Vols I-Ii. Hong Kong: Int Assoc Engineers-Iaeng; 2008. pp. 303-+.

80. Salam N, Deeba F, Qadir F, Al-Hijli F, Al-Otaibi YN. Analysis of Correlation between Google Search Trends and Dengue Outbreaks from India. J Clin Diagn Res. 2019;13: LC13–LC15. doi:10.7860/JCDR/2019/42611.13304

81. Chire Saire JE. Building Intelligent Indicators to Detect Dengue Epidemics in Brazil using Social Networks. OrjuelaCanon AD, editor. 2019 Ieee Colombian Conference on Applications in Computational Intelligence (colcaci). New York: Ieee; 2019.

82. Swain S, Seeja KR. Analysis of Epidemic Outbreak in Delhi Using Social Media Data. In: Kaushik S, Gupta D, Kharb L, Chahal D, editors. Information, Communication and Computing Technology. Singapore: Springer-Verlag Singapore Pte Ltd; 2017. pp. 25–34.

83. Saravanan N, Gayathri V. Classification of Dengue Dataset Using J48 Algorithm and Ant Colony Based Aj48 Algorithm. New York: Ieee; 2017.

84. Carlos MA, Nogueira M, Machado RJ. Analysis of Dengue Outbreaks Using Big Data Analytics and Social Networks. 2017 4th International Conference on Systems and Informatics (icsai). New York: Ieee; 2017. pp. 1592–1597.

85. Ye X, Li S, Yang X, Qin C. Use of Social Media for the Detection and Analysis of Infectious Diseases in China. ISPRS Int Geo-Inf. 2016;5: 156. doi:10.3390/ijgi5090156

86. Li W, Chen Y. Risk Factor Identification and Spatiotemporal Diffusion Path During the Dengue Outbreak. In: Weng Q, Gamba P, Xian G, Chen JM, Liang S, editors. 2016 4rth International Workshop on Earth Observation and Remote Sensing Applications (EORSA). New York: Ieee; 2016.

87. Fathima S, Hundewale N. Comparison of Classification Techniques-SVM and Naives Bayes to predict the Arboviral Disease-Dengue. In: Chen B, Chen J, Chen X, Chen Y, Cho YR, Cui J, et al., editors. 2011 Ieee International Conference on Bioinformatics and Biomedicine Workshops. Los Alamitos: Ieee Computer Soc; 2011. pp. 538–539.

88. Srilekha G, Assistant Professor CD, Anupama B, Assistant Professor CD. Prediction of Dengue Outbreaks with Big Data using Machine Learning. GEDRAG & ORGANISATIE REVIEW. 33. Available: https://www.academia.edu/42849518/Prediction_of_Dengue_Outbreaks_with_Big_Data_using_Machine_Learning

89. Ganthimathi M, Thangamani M, Mallika C, Prasanna Balaji V. Prediction of dengue fever using intelligent classifier. International Journal of Emerging Trends in Engineering Research. 2020;8: 1338–1341. doi:10.30534/ijeter/2020/65842020

90. Kumar NK, Sikamani KT. Prediction of chronic and infectious diseases using machine learning classifiers-A systematic approach. International Journal of Intelligent Engineering and Systems. 2020;13: 11–20. doi:10.22266/IJIES2020.0831.02

91. Guiyab RB. Development of prediction models for the dengue survivability prediction: An integration of data mining and decision support system. International Journal of Innovative Technology and Exploring Engineering. 2019;8: 2199–2205. doi:10.35940/ijitee.J9411.0881019

92. Chovatiya M, Dhameliya A, Deokar J, Gonsalves J, Mathur A. Prediction of dengue using recurrent neural network. 2019. pp. 926–929. doi:10.1109/icoei.2019.8862581

93. Kerdprasop K, Kerdprasop N, Chansilp K, Chuaybamroong P. The Use of Spaceborne and Oceanic Sensors to Model Dengue Incidence in the Outbreak Surveillance System. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2019;11619 LNCS: 447–460. doi:10.1007/978-3-030-24289-3_33

94. Link H, Richter SN, Leung VJ, Brost RC, Phillips CA, Staid A. Statistical models of dengue fever. Communications in Computer and Information Science. 2019;996: 175–186. doi:10.1007/978-981-13-6661-1_14

95. Arafiyah R, Hermin F, Kartika IR, Alimuddin A, Saraswati I. Classification of Dengue Haemorrhagic Fever (DHF) using SVM, naive bayes and random forest. 2018. doi:10.1088/1757-899X/434/1/012070

96. Mishra S, Tripathy HK, Panda AR. An improved and adaptive attribute selection technique to optimize Dengue fever prediction. International Journal of Engineering and Technology(UAE). 2018;7: 480–486. doi:10.14419/ijet.v7i2.29.13802

97. Wu C-H, Kao S-C, Kan M-H. Knowledge discovery in open data of dengue epidemic. 2017. doi:10.1145/3092090.3092093

98. Albinati J, Meira W Jr, Pappa GL, Teixeira M, Marques-Toledo C. Enhancement of epidemiological models for dengue fever based on twiter data. 2017. pp. 109–118. doi:10.1145/3079452.3079464

99. Zhu G, Hunter J, Jiang Y. Improved Prediction of Dengue Outbreak Using the Delay Permutation Entropy. 2017. pp. 828–832. doi:10.1109/iThings-GreenCom-CPSCom-SmartData.2016.172

100. Zainudin Z, Shamsuddin SM. Predictive analytics in Malaysian dengue data from 2010 until 2015 using BigML. International Journal of Advances in Soft Computing and its Applications. 2016;8: 18–30.

101. Milinovich GJ, Avril SMR, Clements ACA, Brownstein JS, Tong S, Hu W. Using internet search queries for infectious disease surveillance: Screening diseases for suitability. BMC Infectious Diseases. 2014;14. doi:10.1186/s12879-014-0690-1

102. Ongruk P, Siriyasatien P, Kesorn K. New key factors discovery to enhance dengue fever forecasting model. Advanced Materials Research. 2014;931–932: 1457–1461. doi:10.4028/www.scientific.net/AMR.931-932.1457

103. Balasundaram A, Bhuvaneswari PTV. Comparative study on decision tree based data mining algorithm to assess risk of epidemic. 2013. pp. 390–396. doi:10.1049/ic.2013.0344

104. Wu Y, Lee G, Fu X, Soh H, Hung T. Mining weather information in dengue outbreak: Predicting future cases based on wavelet, SVM and GA. Lecture Notes in Electrical Engineering. 2009;39 LNEE: 483–494. doi:10.1007/978-90-481-2311-7_41

105. Zhang Y, Ibaraki M, Schwartz FW. Disease surveillance using online news: Dengue and zika in tropical countries. Journal of Biomedical Informatics. 2020;102. doi:10.1016/j.jbi.2020.103374

106. Souza RCSNP. Detecting spatial clusters of infection risk with geo-located social media data. 2018.

107. Coberly JS, Fink CR, Elbert Y, Yoon I-K, Velasco JM, Tomayao AD, et al. Tweeting Fever: Can Twitter Be Used to Monitor the Incidence of Dengue-Like Illness in the Philippines? JOHNS HOPKINS APL TECHNICAL DIGEST. 2014;32: 12.

108. Gomide J, Veloso A, Meira Jr. W, Almeida V, Benevenuto F, Ferraz F, et al. Dengue surveillance based on a computational model of spatio-temporal locality of Twitter. 2011. doi:10.1145/2527031.2527049

109. Fang Z-H, Tzeng J-S, Chen CC, Chou T-C. A study of machine learning models in epidemic surveillance: Using the query logs of search engines. 2010. pp. 1438–1449.

110. Souza J, Leung CK, Cuzzocrea A. An Innovative Big Data Predictive Analytics Framework over Hybrid Big Data Sources with an Application for Disease Analytics. Advances in Intelligent Systems and Computing. 2020;1151 AISC: 669–680. doi:10.1007/978-3-030-44041-1_59

111. Yogapriya P, Geetha P. Dengue disease detection using K-means, hierarchical, kohonen-SOM clustering. International Journal of Innovative Technology and Exploring Engineering. 2019;8: 904–907. doi:10.35940/ijitee.J9066.0881019

112. Adias Sabara M, Somantri O, Nurcahyo H, Kurnia Achmadi N, Latifah U, Harsono. Diagnosis classification of dengue fever based on Neural Networks and Genetic algorithms. 2019. doi:10.1088/1742-6596/1175/1/012065

113. Jongmuenwai B, Lowanichchai S, Jabjone S. Comparision using data mining algorithm techniques for predicting of dengue fever data in northeastern of Thailand. 2019. pp. 532–535. doi:10.1109/ECTICon.2018.8619953

114. Balasaravanan K, Prakash M. Detection of dengue disease using artificial neural network based classification techniquetion. International Journal of Engineering and Technology(UAE). 2018;7: 13–15. doi:10.14419/ijet.v7i1.3.8978

115. Acosta Torres J, Oller Meneses L, Sokol N, Balado Sardiñas R, Montero Díaz D, Balado Sansón R, et al. Decision tree technique applied to the clinical method in the dengue diagnosis. Revista Cubana de Pediatria. 2016;88: 441–453.

116. Soonthornphisaj N, Thitiprayoonwongse D. Knowledge discovery on dengue patients using data mining techniques. 2016. pp. 371–375.

117. Fathima SA, Hundewale N. Comparitive analysis of machine learning techniques for classification of arbovirus. 2012. pp. 376–379. doi:10.1109/BHI.2012.6211593

118. Fathima S, Hundewale N. Comparison of classification techniques-SVM and naives bayes to predict the Arboviral disease-Dengue. 2011. pp. 538–539. doi:10.1109/BIBMW.2011.6112426

119. Long ZA, Abu Bakar A, Razak Hamdan A, Sahani M. Multiple attribute frequent mining-based for dengue outbreak. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2010;6440 LNAI: 489–496. doi:10.1007/978-3-642-17316-5_46

**S4 Table. Studied outcomes for dengue surveillance and prediction**

| Outcome | n | % |
|---|---|---|
| Dengue incidence rate | 58 | 49 |
| Dengue diagnosis | 20 | 17 |
| Dengue outbreak | 18 | 15 |
| Dengue severity | 8 | 7 |
| Dengue-related tweet | 4 | 3 |
| | | |
| Other outcome | 18 | 15 |
| Dengue risk level | 3 | 2.5 |
| Dengue spatial cluster | 2 | 1.7 |
| Dengue mortality | 2 | 1.7 |
| Dengue transmission risk | 2 | 1.7 |
| Dengue serotype | 1 | 0.8 |
| Dengue maps | 1 | 0.8 |
| Dynamic mobility-weighted incidence index | 1 | 0.8 |
| Length of hospital stay | 1 | 0.8 |
| Peak time of dengue incidence | 1 | 0.8 |
| Peak value | 1 | 0.8 |
| R0 | 1 | 0.8 |
| Trend in dengue articles | 1 | 0.8 |
| Yearly dengue deaths | 1 | 0.8 |

**S5 Table. Detailed study outcomes and results**

| Reference | Study outcome | Study aim: Assesing a data source | Study aim: Prediction | Data source | Association between source and outcome? | Best model | Analysis to find significant predictors among covariates? | Comparison of several models? |
|---|---|---|---|---|---|---|---|---|
| Polwiang et al. (2020) [1] | Dengue incidence rate | yes | yes | Climate data | yes | Regression model | yes | yes |
| Xu et al. (2020) [2] | Dengue incidence rate | no | yes | NA | NA | Artificial Neural Network | no | yes |
| Rangarajan et al. (2019) [3] | Dengue incidence rate, Peak val | no | yes | NA | NA | Time series model | no | yes |
| Anno et al. (2019) [4] | Dengue outbreak | yes | no | Geographical data | yes | NA | NA | NA |
| Romero et al. (2019) [5] | Dengue outbreak | no | yes | NA | NA | Regression model | yes | no |
| Mello-Román et al. (2019) [6] | Dengue diagnosis | no | yes | NA | NA | Artificial Neural Network | no | yes |
| Stolerman et al. (2019) [7] | Dengue incidence rate | no | yes | NA | NA | Support Vector Machine | yes | no |
| Macedo et al. (2019) [8] | Dengue severity | no | yes | NA | NA | Artificial Neural Network | yes | no |
| Husnayain et al. (2019) [9] | Dengue incidence rate | yes | no | Google | yes | NA | NA | NA |
| Souza et al. (2019) [10] | Dengue spatial cluster | yes | no | Twitter | yes | NA | NA | NA |
| Ramadona et al. (2019) [11] | Dengue incidence rate, Dynamic | yes | yes | Twitter | yes | Bayesian | yes | no |
| Davi et al. (2019) [12] | Dengue severity | no | yes | NA | NA | Artificial Neural Network | yes | no |
| Guo et al. (2019) [13] | Dengue incidence rate | yes | yes | Baidu, Twitter | yes | Regression model | no | yes |
| Koh et al. (2018) [14] | Dengue incidence rate | no | yes | NA | NA | Artificial Neural Network | yes | yes |
| Carvajal et al. (2018) [15] | Dengue incidence rate | yes | yes | Climate data | yes | Decision Tree | yes | yes |
| Baquero et al. (2018) [16] | Dengue incidence rate | no | yes | NA | NA | Regression model | yes | yes |
| Chen et al. (2018) [17] | Dengue incidence rate | no | yes | NA | NA | Regression model | yes | no |
| Villanes et al. (2018) [18] | Trend in dengue articles | yes | no | LexisNexis | yes | NA | NA | NA |
| Guo et al. (2017) [19] | Dengue incidence rate | yes | yes | Baidu | yes | Support Vector Machine | no | yes |
| Chatterjee et al. (2018) [20] | Dengue severity | no | yes | NA | NA | Artificial Neural Network | no | yes |
| Guo et al. (2017) [21] | Dengue incidence rate | yes | no | Baidu | yes | NA | NA | NA |
| Yang et al. (2017) [22] | Dengue incidence rate | no | yes | NA | NA | Time series model | no | yes |
| Marques-Toledo et al. (2017) [23] | Dengue incidence rate | yes | yes | Twitter, Google, Wikiped | yes | Regression model | yes | no |
| Premaratne et al. (2017) [24] | Dengue severity | no | yes | NA | NA | Other | yes | no |
| Jayasundara et al. (2018) [25] | Dengue severity | no | yes | NA | NA | Other | no | no |
| Li et al. (2017) [26] | Dengue incidence rate | yes | yes | Baidu | yes | Regression model | yes | no |
| Kesorn et al. (2015) [27] | Dengue incidence rate | no | yes | NA | NA | Support Vector Machine | yes | yes |
| Dayama et al. (2014) [28] | Dengue outbreak | no | yes | NA | NA | Time series model | no | yes |
| Sampath et al. (2014) [29] | Dengue outbreak | no | yes | NA | NA | Regression model | no | yes |
| Gluskin et al. (2014) [30] | Dengue incidence rate | yes | yes | Google | yes | Regression model | yes | no |
| Flamand et al. (2014) [31] | Dengue incidence rate | yes | no | Climate data | yes | NA | NA | NA |
| Torres et al. (2014) [32] | Dengue incidence rate | no | yes | NA | NA | Other | no | yes |
| Buczak et al. (2012) [33] | Dengue outbreak | no | yes | NA | NA | Association Rules | no | yes |
| Hoen et al. (2012) [34] | Dengue outbreak | yes | no | HealthMap | yes | NA | NA | NA |
| Althouse et al. (2011) [35] | Dengue incidence rate | yes | yes | Google | yes | Regression model | yes | no |
| Chan et al. (2011) [36] | Dengue incidence rate | yes | no | Google | yes | NA | NA | NA |
| Faisal et al. (2010) [37] | Dengue risk level | no | yes | NA | NA | Artificial Neural Network | yes | yes |
| Ibrahim et al. (2010) [38] | Dengue risk level | no | yes | NA | NA | Artificial Neural Network | no | yes |
| Syamsuddin et al. (2020) [39] | Dengue incidence rate | yes | no | Google | yes | NA | NA | NA |
| Romero-Alvarez et al. (2020) [40] | Dengue incidence rate | yes | no | Google | yes | NA | NA | NA |
| Liu et al. (2019) [41] | Dengue incidence rate | yes | yes | Baidu | yes | Regression model | yes | yes |
| Musa et al. (2019) [42] | R0 | no | yes | NA | NA | Other | no | no |
| Messina et al. (2019) [43] | Dengue maps | no | yes | NA | NA | Regression model | yes | no |
| Titus et al. (2018) [44] | Dengue outbreak, Dengue incid | yes | yes | Climate data | yes | Regression model | yes | no |
| Marques-Toledo et al. (2019) [45] | Dengue transmission risk | no | yes | NA | NA | Regression model | yes | no |
| Verma et al. (2018) [46] | Dengue incidence rate | yes | no | Google | yes | NA | NA | NA |
| Ho et al. (2018) [47] | Dengue incidence rate | yes | no | Google | yes | NA | NA | NA |
| Phakhounthong et al. (2018) [48] | Dengue severity | no | yes | NA | NA | Decision Tree | yes | yes |
| Strauss et al. (2017) [49] | Dengue incidence rate | yes | no | Google | yes | NA | NA | NA |
| Nsoesie et al. (2016) [50] | Dengue incidence rate | yes | yes | Twitter | yes | Bayesian | yes | yes |
| Liu et al. (2016) [51] | Dengue incidence rate | yes | yes | Baidu | yes | Decision Tree | yes | no |
| Ximenes et al. (2016) [52] | Dengue risk level | no | yes | NA | NA | Other | no | no |
| Mohamad et al. (2014) [53] | Dengue outbreak | no | yes | NA | NA | Other | no | yes |
| Puengpreeda, et al. (2020) [54] | Dengue incidence rate | no | yes | NA | NA | Decision Tree | no | yes |
| Amin, et al. (2020) [55] | Dengue diagnosis | no | yes | NA | NA | Artificial Neural Network | no | yes |
| Manogaran, et al. (2018) [56] | Dengue incidence rate | yes | no | Climate data | yes | NA | NA | NA |
| Agarwal, et al. (2018) [57] | Dengue incidence rate | yes | yes | Climate data | yes | Regression model | yes | yes |
| Manogaran, et al. (2018) [58] | Dengue incidence rate | no | yes | NA | NA | Bayesian | no | yes |
| Jahangir, et al. (2018) [59] | Dengue diagnosis | no | yes | NA | NA | Association Rules | yes | no |
| Husin, et al. (2018) [60] | Dengue diagnosis | no | yes | NA | NA | Artificial Neural Network | yes | no |
| Anggraeni, et al. (2018) [61] | Dengue incidence rate | no | yes | NA | NA | Artificial Neural Network | no | yes |
| Livelo, et al. (2018) [62] | Dengue-related tweet | yes | yes | Twitter | yes | NA | NA | NA |
| Wiratmadja, et al. (2018) [63] | Length of hospital stay | no | yes | NA | NA | Decision Tree | yes | no |
| Arafiyah, et al. (2018) [64] | Dengue severity | no | yes | NA | NA | Bayesian | no | no |
| Abuhamad, et al. (2017) [65] | Dengue outbreak | no | yes | NA | NA | Artificial Neural Network | yes | yes |
| Manivannan, et al. (2017) [66] | Dengue diagnosis, Dengue serot | no | yes | NA | NA | Clustering | yes | no |
| Dharmawardana, et al. (2017) [67] | Dengue incidence rate | yes | yes | Mobile network big data | yes | Artificial Neural Network | no | yes |
| Espina, et al. (2017) [68] | Dengue incidence rate | yes | no | Twitter | yes | NA | NA | NA |
| Rahim, et al. (2017) [69] | Dengue outbreak, Dengue morta | no | yes | NA | NA | Decision Tree | no | yes |
| Klein, et al. (2017) [70] | Dengue incidence rate | yes | no | Twitter | yes | NA | NA | NA |
| Kerdprasop, et al. (2016) [71] | Dengue incidence rate | yes | yes | Climate data | yes | Decision Tree | yes | yes |
| Anggraeni, et al. (2016) [72] | Dengue incidence rate | yes | yes | Google | yes | Time series model | yes | yes |
| Mathulamuthu, et al. (2016) [73] | Dengue incidence rate | no | yes | NA | NA | Clustering | yes | no |
| Rahmawati, et al. (2016) [74] | Dengue outbreak | no | yes | NA | NA | Support Vector Machine | yes | no |
| Missier, et al. (2016) [75] | Dengue-related tweet | yes | no | Twitter | yes | NA | NA | NA |
| Abeyrathna, et al. (2016) [76] | Dengue incidence rate | no | yes | NA | NA | Other | yes | no |
| Fathima, et al. (2015) [77] | Dengue diagnosis | no | yes | NA | NA | Decision Tree | yes | no |
| Tazkia, et al. (2015) [78] | Dengue outbreak | no | yes | NA | NA | Bayesian | yes | no |
| Wu, et al. (2008) [79] | Dengue incidence rate | yes | yes | Climate data | yes | Support Vector Machine | yes | yes |
| Salam, et al. (2019) [80] | Dengue incidence rate, Yearly d | yes | no | Google | yes | NA | NA | NA |
| Saire, et al. (2019) [81] | Dengue-related tweet | yes | no | Twitter | yes | NA | NA | NA |
| Swain, et al. (2017) [82] | Dengue-related tweet | yes | no | Twitter | yes | NA | NA | NA |
| Saravanan, et al. (2017) [83] | Dengue diagnosis | no | yes | NA | NA | Decision Tree | no | yes |
| Carlos, et al. (2017) [84] | Dengue incidence rate | yes | no | Twitter | yes | NA | NA | NA |
| Ye, et al. (2016) [85] | Dengue incidence rate | yes | no | Weibo | yes | NA | NA | NA |
| Li, et al. (2016) [86] | Dengue outbreak | yes | no | Vector data, Climate data | yes | NA | NA | NA |
| Fathima, et al. (2011) [87] | Dengue diagnosis | no | yes | NA | NA | Support Vector Machine | no | yes |
| Srilekha et al. (2020) [88] | Dengue outbreak | no | yes | NA | NA | Decision Tree | no | yes |
| Ganthimathi et al. (2020) [89] | Dengue diagnosis | no | yes | NA | NA | Support Vector Machine | no | yes |
| Kumar et al. (2020) [90] | Dengue diagnosis | no | yes | NA | NA | Artificial Neural Network | no | yes |
| Guiyab et al. (2019) [91] | Dengue survivability | no | yes | NA | NA | Decision Tree | yes | yes |
| Chovatiya et al. (2019) [92] | Dengue incidence rate | no | yes | NA | NA | Artificial Neural Network | yes | no |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Kerdprasop et al. (2019) [93] | Dengue incidence rate | no | yes | NA | NA | Decision Tree | yes | yes |
| Link et al. (2019) [94] | Dengue outbreak | no | yes | NA | NA | Time series model | yes | yes |
| Arafiyah et al. (2018) [95] | Dengue diagnosis | no | yes | NA | NA | Decision Tree | no | yes |
| Mishra et al. (2018) [96] | Dengue diagnosis | no | yes | NA | NA | Other | no | yes |
| Wu et al. (2017) [97] | Dengue incidence rate | yes | yes | Google | no | Decision Tree | yes | no |
| Albinati et al. (2017) [98] | Dengue incidence rate | yes | yes | Twitter | yes | Other | yes | no |
| Zhu et al. (2017) [99] | Dengue outbreak | yes | yes | Climate data | yes | Support Vector Machine | yes | no |
| Zainudin et al. (2016) [100] | Dengue transmission risk | no | yes | NA | NA | Decision Tree | no | no |
| Milinovich et al. (2014) [101] | Dengue incidence rate | yes | no | Google | yes | NA | NA | NA |
| Ongruk et al. (2014) [102] | Dengue outbreak | yes | yes | Vector data, Climate data | yes | Regression model | yes | no |
| Balasundaram et al. (2013) [103] | Dengue diagnosis | no | yes | NA | NA | Decision Tree | no | yes |
| Wu et al. (2009) [104] | Dengue outbreak | yes | yes | Climate data | yes | Support Vector Machine | yes | yes |
| Zhang et al. (2020) [105] | Dengue incidence rate | yes | no | LexisNexis | yes | NA | NA | NA |
| Souza et al. (2018) [106] | Dengue spatial cluster | yes | no | Twitter | yes | NA | NA | NA |
| Coberly et al. (2014) [107] | Dengue incidence rate | yes | no | Twitter | yes | NA | NA | NA |
| Gomide et al. (2011) [108] | Dengue incidence rate | yes | no | Twitter | yes | NA | NA | NA |
| Fang et al. (2010) [109] | Dengue incidence rate | yes | no | Google | yes | NA | NA | NA |
| Souza et al. (2020) [110] | Dengue incidence rate | yes | yes | Google, Climate data | yes | Support Vector Machine | yes | yes |
| Yogapriya et al. (2019) [111] | Dengue diagnosis | no | yes | NA | NA | Clustering | no | yes |
| Adias et al. (2019) [112] | Dengue diagnosis | no | yes | NA | NA | Artificial Neural Network | no | yes |
| Jongmuenwai et al. (2019) [113] | Dengue diagnosis | no | yes | NA | NA | Decision Tree | no | yes |
| Balasaravanan et al. (2018) [114] | Dengue diagnosis | no | yes | NA | NA | Artificial Neural Network | yes | yes |
| Acosta et al. (2016) [115] | Dengue diagnosis | no | yes | NA | NA | Decision Tree | yes | no |
| Soonthornphisaj et al. (2016) [116] | Dengue severity | no | yes | NA | NA | Decision Tree | yes | yes |
| Fathima et al. (2012) [117] | Dengue diagnosis | no | yes | NA | NA | Support Vector Machine | no | yes |
| Fathima et al. (2011) [118] | Dengue diagnosis | no | yes | NA | NA | Support Vector Machine | no | yes |
| Long et al. (2010) [119] | Dengue incidence rate | no | yes | NA | NA | Association Rules | NA | yes |

## References


1. Polwiang S. The time series seasonal patterns of dengue fever and associated weather variables in Bangkok (2003-2017). BMC Infect Dis. 2020;20: 208. doi:10.1186/s12879-020-4902-6
2. Xu J, Xu K, Li Z, Meng F, Tu T, Xu L, et al. Forecast of Dengue Cases in 20 Chinese Cities Based on the Deep Learning Method. Int J Environ Res Public Health. 2020;17. doi:10.3390/ijerph17020453
3. Rangarajan P, Mody SK, Marathe M. Forecasting dengue and influenza incidences using a sparse representation of Google trends, electronic health records, and time series data. PLoS Comput Biol. 2019;15: e1007518. doi:10.1371/journal.pcbi.1007518
4. Anno S, Hara T, Kai H, Lee M-A, Chang Y, Oyoshi K, et al. Spatiotemporal dengue fever hotspots associated with climatic factors in Taiwan including outbreak predictions based on machine-learning. Geospat Health. 2019;14. doi:10.4081/gh.2019.771
5. Romero D, Olivero J, Real R, Guerrero JC. Applying fuzzy logic to assess the biogeographical risk of dengue in South America. Parasit Vectors. 2019;12: 428. doi:10.1186/s13071-019-3691-5
6. Mello-Román JD, Mello-Román JC, Gómez-Guerrero S, García-Torres M. Predictive Models for the Medical Diagnosis of Dengue: A Case Study in Paraguay. Comput Math Methods Med. 2019;2019: 7307803. doi:10.1155/2019/7307803
7. Stolerman LM, Maia PD, Kutz JN. Forecasting dengue fever in Brazil: An assessment of climate conditions. PLoS One. 2019;14: e0220106. doi:10.1371/journal.pone.0220106
8. Macedo Hair G, Fonseca Nobre F, Brasil P. Characterization of clinical patterns of dengue patients using an unsupervised machine learning approach. BMC Infect Dis. 2019;19: 649. doi:10.1186/s12879-019-4282-y
9. Husnayain A, Fuad A, Lazuardi L. Correlation between Google Trends on dengue fever and national surveillance report in Indonesia. Glob Health Action. 2019;12: 1552652. doi:10.1080/16549716.2018.1552652
10. Souza RCSNP, Assunção RM, Oliveira DM, Neill DB, Meira W. Where did I get dengue? Detecting spatial clusters of infection risk with social network data. Spat Spatiotemporal Epidemiol. 2019;29: 163–175. doi:10.1016/j.sste.2018.11.005
11. Ramadona AL, Tozan Y, Lazuardi L, Rocklöv J. A combination of incidence data and mobility proxies from social media predicts the intra-urban spread of dengue in Yogyakarta, Indonesia. PLoS Negl Trop Dis. 2019;13: e0007298. doi:10.1371/journal.pntd.0007298
12. Davi C, Pastor A, Oliveira T, Neto FB de L, Braga-Neto U, Bigham AW, et al. Severe Dengue Prognosis Using Human Genome Data and Machine Learning. IEEE Trans Biomed Eng. 2019;66: 2861–2868. doi:10.1109/TBME.2019.2897285
13. Guo P, Zhang Q, Chen Y, Xiao J, He J, Zhang Y, et al. An ensemble forecast model of dengue in Guangzhou, China using climate and social media surveillance data. Sci Total Environ. 2019;647: 752–762. doi:10.1016/j.scitotenv.2018.08.044
14. Koh Y-M, Spindler R, Sandgren M, Jiang J. A model comparison algorithm for increased forecast accuracy of dengue fever incidence in Singapore and the auxiliary role of total precipitation information. Int J Environ Health Res. 2018;28: 535–552. doi:10.1080/09603123.2018.1496234
15. Carvajal TM, Viacrusis KM, Hernandez LFT, Ho HT, Amalin DM, Watanabe K. Machine learning methods reveal the temporal pattern of dengue incidence using meteorological factors in metropolitan Manila, Philippines. BMC Infect Dis. 2018;18: 183. doi:10.1186/s12879-018-3066-0
16. Baquero OS, Santana LMR, Chiaravalloti-Neto F. Dengue forecasting in São Paulo city with generalized additive models, artificial neural networks and seasonal autoregressive integrated moving average models. PLoS One. 2018;13: e0195065. doi:10.1371/journal.pone.0195065
17. Chen Y, Chu CW, Chen MIC, Cook AR. The utility of LASSO-based models for real time forecasts of endemic infectious diseases: A cross country comparison. J Biomed Inform. 2018;81: 16–30. doi:10.1016/j.jbi.2018.02.014
18. Villanes A, Griffiths E, Rappa M, Healey CG. Dengue Fever Surveillance in India Using Text Mining in Public Media. Am J Trop Med Hyg. 2018;98: 181–191. doi:10.4269/ajtmh.17-0253
19. Guo P, Liu T, Zhang Q, Wang L, Xiao J, Zhang Q, et al. Developing a dengue forecast model using machine learning: A case study in China. PLoS Negl Trop Dis. 2017;11: e0005973. doi:10.1371/journal.pntd.0005973
20. Chatterjee S, Dey N, Shi F, Ashour AS, Fong SJ, Sen S. Clinical application of modified bag-of-features coupled with hybrid neural-based classifier in dengue fever classification using gene expression data. Med Biol Eng Comput. 2018;56: 709–720. doi:10.1007/s11517-017-1722-y
21. Guo P, Wang L, Zhang Y, Luo G, Zhang Y, Deng C, et al. Can internet search queries be used for dengue fever surveillance in China? Int J Infect Dis. 2017;63: 74–76. doi:10.1016/j.ijid.2017.08.001
22. Yang S, Kou SC, Lu F, Brownstein JS, Brooke N, Santillana M. Advances in using Internet searches to track dengue. PLoS Comput Biol. 2017;13: e1005607. doi:10.1371/journal.pcbi.1005607
23. Marques-Toledo C de A, Degener CM, Vinhal L, Coelho G, Meira W, Codeço CT, et al. Dengue prediction by the web: Tweets are a useful tool for estimating and forecasting Dengue at country and city level. PLOS Neglected Tropical Diseases. 2017;11: e0005729. doi:10.1371/journal.pntd.0005729
24. Premaratne MK, Perera SSN, Malavige GN, Jayasinghe S. Mathematical Modelling of Immune Parameters in the Evolution of Severe Dengue. Comput Math Methods Med. 2017;2017: 2187390. doi:10.1155/2017/2187390
25. Jayasundara SDP, Perera SSN, Malavige GN, Jayasinghe S. Mathematical modelling and a systems science approach to describe the role of cytokines in the evolution of severe dengue. BMC Syst Biol. 2017;11: 34. doi:10.1186/s12918-017-0415-3
26. Li Z, Liu T, Zhu G, Lin H, Zhang Y, He J, et al. Dengue Baidu Search Index data can improve the prediction of local dengue epidemic: A case study in Guangzhou, China. PLoS Negl Trop Dis. 2017;11: e0005354. doi:10.1371/journal.pntd.0005354
27. Kesorn K, Ongruk P, Chompoosri J, Phumee A, Thavara U, Tawatsin A, et al. Morbidity Rate Prediction of Dengue Hemorrhagic Fever (DHF) Using the Support Vector Machine and the Aedes aegypti Infection Rate in Similar Climates and Geographical Areas. PLoS One. 2015;10: e0125049. doi:10.1371/journal.pone.0125049
28. Dayama P, Sampath K. Dengue disease outbreak detection. Stud Health Technol Inform. 2014;205: 1105–1109.
29. Sampath K, Dayama P. Predicting the operations alert levels for dengue surveillance and control. Stud Health Technol Inform. 2014;205: 1100–1104.
30. Gluskin RT, Johansson MA, Santillana M, Brownstein JS. Evaluation of Internet-based dengue query data: Google Dengue Trends. PLoS Negl Trop Dis. 2014;8: e2713. doi:10.1371/journal.pntd.0002713
31. Flamand C, Fabregue M, Bringay S, Ardillon V, Quénel P, Desenclos J-C, et al. Mining local climate data to assess spatiotemporal dengue fever epidemic patterns in French Guiana. J Am Med Inform Assoc. 2014;21: e232-240. doi:10.1136/amiajnl-2013-002348
32. Torres C, Bargüil S, Melgarejo M, Olarte A. Fuzzy model identification of dengue epidemic in Colombia based on multiresolution analysis. Artif Intell Med. 2014;60: 41–51. doi:10.1016/j.artmed.2013.11.008
33. Buczak AL, Koshute PT, Babin SM, Feighner BH, Lewis SH. A data-driven epidemiological prediction method for dengue outbreaks using local and remote sensing data. BMC Med Inform Decis Mak. 2012;12: 124. doi:10.1186/1472-6947-12-124
34. Hoen AG, Keller M, Verma AD, Buckeridge DL, Brownstein JS. Electronic event-based surveillance for monitoring dengue, Latin America. Emerg Infect Dis. 2012;18: 1147–1150. doi:10.3201/eid1807.120055
35. Althouse BM, Ng YY, Cummings DAT. Prediction of dengue incidence using search query surveillance. PLoS Negl Trop Dis. 2011;5: e1258. doi:10.1371/journal.pntd.0001258
36. Chan EH, Sahai V, Conrad C, Brownstein JS. Using web search query data to monitor dengue epidemics: a new model for neglected tropical disease surveillance. PLoS Negl Trop Dis. 2011;5: e1206. doi:10.1371/journal.pntd.0001206
37. Faisal T, Taib MN, Ibrahim F. Neural network diagnostic system for dengue patients risk classification. J Med Syst. 2012;36: 661–676. doi:10.1007/s10916-010-9532-x
38. Ibrahim F, Faisal T, Salim MIM, Taib MN. Non-invasive diagnosis of risk in dengue patients using bioelectrical impedance analysis and artificial neural network. Med Biol Eng Comput. 2010;48: 1141–1148. doi:10.1007/s11517-010-0669-z
39. Syamsuddin M, Fakhruddin M, Sahetapy-Engel JTM, Soewono E. Causality Analysis of Google Trends and Dengue Incidence in Bandung, Indonesia With Linkage of Digital Data Modeling: Longitudinal Observational Study. J Med Internet Res. 2020;22: e17633. doi:10.2196/17633
40. Romero-Alvarez D, Parikh N, Osthus D, Martinez K, Generous N, Del Valle S, et al. Google Health Trends performance reflecting dengue incidence for the Brazilian states. BMC Infect Dis. 2020;20: 252. doi:10.1186/s12879-020-04957-0
41. Liu D, Guo S, Zou M, Chen C, Deng F, Xie Z, et al. A dengue fever predicting model based on Baidu search index data and climate data in South China. PLoS One. 2019;14: e0226841. doi:10.1371/journal.pone.0226841
42. Musa SS, Zhao S, Chan H-S, Jin Z, He DH. A mathematical model to study the 2014-2015 large-scale dengue epidemics in Kaohsiung and Tainan cities in Taiwan, China. Math Biosci Eng. 2019;16: 3841–3863. doi:10.3934/mbe.2019190
43. Messina JP, Brady OJ, Golding N, Kraemer MUG, Wint GRW, Ray SE, et al. The current and future global distribution and population at risk of dengue. Nat Microbiol. 2019;4: 1508–1515. doi:10.1038/s41564-019-0476-8
44. Titus Muurlink O, Stephenson P, Islam MZ, Taylor-Robinson AW. Long-term predictors of dengue outbreaks in Bangladesh: A data mining approach. Infect Dis Model. 2018;3: 322–330. doi:10.1016/j.idm.2018.11.004
45. Marques-Toledo CA, Bendati MM, Codeço CT, Teixeira MM. Probability of dengue transmission and propagation in a non-endemic temperate area: conceptual model and decision risk levels for early alert, prevention and control. Parasit Vectors. 2019;12: 38. doi:10.1186/s13071-018-3280-z
46. Verma M, Kishore K, Kumar M, Somdi AR, Aggarwal G, Kathirvel S. Google Search Trends Predicting Disease Outbreaks: An Analysis from India. Healthc Inform Res. 2018;24: 300–308. doi:10.4258/hir.2018.24.4.300
47. Ho CC, Ting C-Y, Raja DB. Using Public Open Data to Predict Dengue Epidemic: Assessment of Weather Variability, Population Density, and Land use as Predictor Variables for Dengue Outbreak Prediction using Support Vector Machine. Indian Journal of Science and Technology. 2018;11. doi:10.17485/ijst/2018/v11i4/115405
48. Phakhounthong K, Chaovalit P, Jittamala P, Blacksell SD, Carter MJ, Turner P, et al. Predicting the severity of dengue fever in children on admission based on clinical features and laboratory indicators: application of classification tree analysis. BMC Pediatr. 2018;18: 109. doi:10.1186/s12887-018-1078-y
49. Strauss RA, Castro JS, Reintjes R, Torres JR. Google dengue trends: An indicator of epidemic behavior. The Venezuelan Case. Int J Med Inform. 2017;104: 26–30. doi:10.1016/j.ijmedinf.2017.05.003
50. Nsoesie EO, Flor L, Hawkins J, Maharana A, Skotnes T, Marinho F, et al. Social Media as a Sentinel for Disease Surveillance: What Does Sociodemographic Status Have to Do with It? PLoS Curr. 2016;8. doi:10.1371/currents.outbreaks.cc09a42586e16dc7dd62813b7ee5d6b6
51. Liu K, Wang T, Yang Z, Huang X, Milinovich GJ, Lu Y, et al. Using Baidu Search Index to Predict Dengue Outbreak in China. Sci Rep. 2016;6: 38040. doi:10.1038/srep38040
52. Ximenes R, Amaku M, Lopez LF, Coutinho FAB, Burattini MN, Greenhalgh D, et al. The risk of dengue for non-immune foreign visitors to the 2016 summer olympic games in Rio de Janeiro, Brazil. BMC Infect Dis. 2016;16: 186. doi:10.1186/s12879-016-1517-z
53. Mohamad Mohsin MF, Abu Bakar A, Hamdan AR. Outbreak detection model based on danger theory. Appl Soft Comput. 2014;24: 612–622. doi:10.1016/j.asoc.2014.08.030
54. Puengpreeda A, Yhusumram S, Sirikulvadhana S. Weekly Forecasting Model for Dengue Hemorrhagic Fever Outbreak in Thailand. Eng J-Thail. 2020;24: 71–87. doi:10.4186/ej.2020.24.3.71
55. Amin S, Uddin MI, Hassan S, Khan A, Nasser N, Alharbi A, et al. Recurrent Neural Networks With TF-IDF Embedding Technique for Detection and Classification in Tweets of Dengue Disease. IEEE Access. 2020;8: 131522–131533. doi:10.1109/ACCESS.2020.3009058
56. Manogaran G, Lopez D, Chilamkurti N. In-Mapper combiner based MapReduce algorithm for processing of big climate data. Futur Gener Comp Syst. 2018;86: 433–445. doi:10.1016/j.future.2018.02.048
57. Agarwal N, Koti SR, Saran S, Kumar AS. Data mining techniques for predicting dengue outbreak in geospatial domain using weather parameters for New Delhi, India. Curr Sci. 2018;114: 2281–2291. doi:10.18520/cs/v114/i11/2281-2291
58. Manogaran G, Lopez D. A Gaussian process based big data processing framework in cluster computing environment. Cluster Comput. 2018;21: 189–204. doi:10.1007/s10586-017-0982-5
59. Jahangir I, Abdul-Basit, Hannan A, Javed S. Prediction of Dengue Disease Through Data Mining by Using Modified Apriori Algorithm. Proceedings of the 4th Acm International Conference of Computing for Engineering and Sciences (icces'2018). New York: Assoc Computing Machinery; 2018. doi:10.1145/3213187.3287612
60. Husin NA, Alharogi A, Mustapha N, Hamdan H, Husin UA. Early Self-Diagnosis of Dengue Symptoms Using Fuzzy and Data Mining Approach. In: Nifa F a. A, Lin CK, Hussain A, editors. Proceedings of the 3rd International Conference on Applied Science and Technology (icast'18). Melville: Amer Inst Physics; 2018. p. 020048. doi:10.1063/1.5055450
61. Anggraeni W, Pramudita G, Riksakomara E, Radityo PW, Samopa F, Pujiadi, et al. Artificial Neural Network for Health Data Forecasting, Case Study: Number of Dengue Hemorrhagic Fever Cases in Malang Regency, Indonesia. 2018 International Conference on Electrical Engineering and Computer Science (icecos). New York: Ieee; 2018. pp. 207–212.
62. Dennison Livelo E, Cheng C. Intelligent Dengue Infoveillance Using Gated Recurrent Neural Learning and Cross-Label Frequencies. 2018 Ieee International Conference on Agents (ica). New York: Ieee; 2018. pp. 2–7.
63. Wiratmadja II, Salamah SY, Govindaraju R. Healthcare Data Mining: Predicting Hospital Length of Stay of Dengue Patients. J Eng Technol Sci. 2018;50: 110–126. doi:10.5614/j.eng.technol.sci.2018.50.1.8
64. Arafiyah R, Hermin F. Data mining for dengue hemorrhagic fever (DHF) prediction with naive Bayes method. 1st International Conference of Education on Sciences, Technology, Engineering, and Mathematics (ice-Stem). Bristol: Iop Publishing Ltd; 2018. p. 012077. doi:10.1088/1742-6596/948/1/012077
65. Abuhamad HIS, Abu Bakar A, Zainudin S, Sahani M, Ali ZM. Feature Selection Algorithms for Malaysian Dengue Outbreak Detection Model. Sains Malays. 2017;46: 255–265. doi:10.17576/jsm-2017-4602-10
66. Manivannan P, Devi PI. Dengue Fever Prediction Using K-Means Clustering Algorithm. 2017 Ieee International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (incos). New York: Ieee; 2017.

67. Dharmawardana KGS, Lokuge JN, Dassanayake PSB, Sirisena ML, Fernando ML, Perera AS, et al. Predictive Model for the Dengue Incidences in Sri Lanka Using Mobile Network Big Data. 2017 Ieee International Conference on Industrial and Information Systems (iciis). New York: Ieee; 2017. pp. 278–283.
68. Espina K, Estuar MRJE. Infodemiology for Syndromic Surveillance of Dengue and Typhoid Fever in the Philippines. In: CruzCunha MM, Varajao JEQ, Rijo R, Martinho R, Peppard J, SanCristobal JR, et al., editors. Centeris 2017 - International Conference on Enterprise Information Systems / Projman 2017 - International Conference on Project Management / Hcist 2017 - International Conference on Health and Social Care Information Systems and Technologies, Centeri. Amsterdam: Elsevier Science Bv; 2017. pp. 554–561. doi:10.10
69. Rahim NF, Taib SM, Abidin AIZ. Dengue Fatality Prediction Using Data Mining. J Fundam Appl Sci. 2017;9: 671–683. doi:10.4314/jfas.v9i6s.52
70. Klein GH, Neto PG, Tezza R. Big Data and social media: surveillance of networks as management tool. Saude Soc. 2017;26: 208–217. doi:10.1590/S0104-12902017164943
71. Kerdprasop N, Kerdprasop K. Remote Sensing Based Modeling of Dengue Outbreak with Regression and Binning Classification. 2016 2nd Ieee International Conference on Computer and Communications (iccc). New York: Ieee; 2016. pp. 46–49.
72. Anggraeni W, Aristiani L. Using Google Trend Data in Forecasting Number of Dengue Fever Cases with ARIMAX Method Case Study : Surabaya, Indonesia. Proceedings of 2016 International Conference on Information & Communication Technology and Systems (icts). New York: Ieee; 2016. pp. 114–118.
73. Mathulamuthu SS, Asirvadam VS, Dass SC, Gill BS, Loshini T. Predicting Dengue Incidences Using Cluster Based Regression on Climate Data. 2016 6th Ieee International Conference on Control System, Computing and Engineering (iccsce). New York: Ieee; 2016. pp. 245–250.
74. Rahmawati D, Huang Y-P. Using C-support Vector Classification to Forecast Dengue Fever Epidemics in Taiwan. In: Wang WJ, Lee PJ, Er MJ, Jeng JT, editors. 2016 International Conference on System Science and Engineering (icsse). New York: Ieee; 2016.
75. Missier P, Romanovsky A, Miu T, Pal A, Daniilakis M, Garcia A, et al. Tracking Dengue Epidemics Using Twitter Content Classification and Topic Modelling. In: Casteleyn S, Dolog P, Pautasso C, editors. Current Trends in Web Engineering, Icwe 2016 International Workshops. Cham: Springer International Publishing Ag; 2016. pp. 80–92. doi:10.1007/978-3-319-46963-8_7
76. Abeyrathna MP a. R, Abeygunawrdane DA, Wijesundara R a. a. V, Mudalige VB, Bandara M, Perera S, et al. Dengue Propagation Prediction using Human Mobility. 2nd International Mercon 2016 Moratuwa Engineering Research Conference. New York: Ieee; 2016. pp. 156–161.
77. Fathima AS, Manimeglai D. Analysis of Significant Factors for Dengue Infection Prognosis Using the Random Forest Classifier. Int J Adv Comput Sci Appl. 2015;6: 240–245.
78. Tazkia RAK, Narita V, Nugroho AS. Dengue Outbreak Prediction for GIS based Early Warning System. 2015 International Conference on Science in Information Technology (ICSITech). New York: Ieee; 2015. pp. 121–125.
79. Wu Y, Lee G, Fu X, Hung T. Detect climatic factors contributing to dengue outbreak based on wavelet, support vector machines and genetic algorithm. In: Ao SI, Gelman L, Hukins DWL, Hunter A, Korsunsky AM, editors. World Congress on Engineering 2008, Vols I-Ii. Hong Kong: Int Assoc Engineers-Iaeng; 2008. pp. 303–+.
80. Salam N, Deeba F, Qadir F, Al-Hjili F, Al-Otaibi YN. Analysis of Correlation between Google Search Trends and Dengue Outbreaks from India. J Clin Diagn Res. 2019;13: LC13–LC15. doi:10.7860/JCDR/2019/42611.13304
81. Chire Saire JE. Building Intelligent Indicators to Detect Dengue Epidemics in Brazil using Social Networks. OrjuelaCanon AD, editor. 2019 Ieee Colombian Conference on Applications in Computational Intelligence (colcaci). New York: Ieee; 2019.
82. Swain S, Seeja KR. Analysis of Epidemic Outbreak in Delhi Using Social Media Data. In: Kaushik S, Gupta D, Kharb L, Chahal D, editors. Information, Communication and Computing Technology. Singapore: Springer-Verlag Singapore Pte Ltd; 2017. pp. 25–34.
83. Saravanan N, Gayathri V. Classification of Dengue Dataset Using J48 Algorithm and Ant Colony Based Aj48 Algorithm. New York: Ieee; 2017.
84. Carlos MA, Nogueira M, Machado RJ. Analysis of Dengue Outbreaks Using Big Data Analytics and Social Networks. 2017 4th International Conference on Systems and Informatics (icsai). New York: Ieee; 2017. pp. 1592–1597.
85. Ye X, Li S, Yang X, Qin C. Use of Social Media for the Detection and Analysis of Infectious Diseases in China. ISPRS Int Geo-Inf. 2016;5: 156. doi:10.3390/ijgi5090156
86. Li W, Chen Y. Risk Factor Identification and Spatiotemporal Diffusion Path During the Dengue Outbreak. In: Weng Q, Gamba P, Xian G, Chen JM, Liang S, editors. 2016 4rth International Workshop on Earth Observation and Remote Sensing Applications (EORSA). New York: Ieee; 2016.
87. Fathima S, Hundewale N. Comparison of Classification Techniques-SVM and Naives Bayes to predict the Arboviral Disease-Dengue. In: Chen B, Chen J, Chen X, Chen Y, Cho YR, Cui J, et al., editors. 2011 Ieee International Conference on Bioinformatics and Biomedicine Workshops. Los Alamitos: Ieee Computer Soc; 2011. pp. 538–539.
88. Srilekha G, Assistant Professor CD, Anuparna B, Assistant Professor CD. Prediction of Dengue Outbreaks with Big Data using Machine Learning. GEDRAG & ORGANISATIE REVIEW. 33. Available: https://www.academia.edu/42849518/Prediction_of_Dengue_Outbreaks_with_Big_Data_using_Machine_Learning
89. Ganthimathi M, Thangamani M, Mallika C, Prsanna Balaji V. Prediction of dengue fever using intelligent classifier. International Journal of Emerging Trends in Engineering Research. 2020;8: 1338–1341. doi:10.30534/ijeter/2020/65842020
90. Kumar NK, Sikamani KT. Prediction of chronic and infectious diseases using machine learning classifiers-A systematic approach. International Journal of Intelligent Engineering and Systems. 2020;13: 11–20. doi:10.22266/IJIES2020.0831.02
91. Guiyab RB. Development of prediction models for the dengue survivability prediction: An integration of data mining and decision support system. International Journal of Innovative Technology and Exploring Engineering. 2019;8: 2199–2205. doi:10.35940/ijitee.J9411.0881019
92. Chovatiya M, Dhameliya A, Deokar J, Gonsalves J, Mathur A. Prediction of dengue using recurrent neural network. 2019. pp. 926–929. doi:10.1109/icoei.2019.8862581
93. Kerdprasop K, Kerdprasop N, Chansilp K, Chuaybamroong P. The Use of Spaceborne and Oceanic Sensors to Model Dengue Incidence in the Outbreak Surveillance System. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2019;11619 LNCS: 447–460. doi:10.1007/978-3-030-24289-3_33
94. Link H, Richter SN, Leung VJ, Brost RC, Phillips CA, Staid A. Statistical models of dengue fever. Communications in Computer and Information Science. 2019;996: 175–186. doi:10.1007/978-981-13-6661-1_14
95. Arafiyah R, Hermin F, Kartika IR, Alimuddin A, Saraswati I. Classification of Dengue Haemorrhagic Fever (DHF) using SVM, naive bayes and random forest. 2018. doi:10.1088/1757-899X/434/1/012070
96. Mishra S, Tripathy HK, Panda AR. An improved and adaptive attribute selection technique to optimize Dengue fever prediction. International Journal of Engineering and Technology(UAE). 2018;7: 480–486. doi:10.14419/ijet.v7i2.29.13802
97. Wu C-H, Kao S-C, Kan M-H. Knowledge discovery in open data of dengue epidemic. 2017. doi:10.1145/3092090.3092093
98. Albinati J, Meira W Jr, Pappa GL, Teixeira M, Marques-Toledo C. Enhancement of epidemiological models for dengue fever based on twiter data. 2017. pp. 109–118. doi:10.1145/3079452.3079464
99. Zhu G, Hunter J, Jiang Y. Improved Prediction of Dengue Outbreak Using the Delay Permutation Entropy. 2017. pp. 828–832. doi:10.1109/iThings-GreenCom-CPSCom-SmartData.2016.172
100. Zainudin Z, Shamsuddin SM. Predictive analytics in Malaysian dengue data from 2010 until 2015 using BigML. International Journal of Advances in Soft Computing and its Applications. 2016;8: 18–30.
101. Milinovich GJ, Avril SMR, Clements ACA, Brownstein JS, Tong S, Hu W. Using internet search queries for infectious disease surveillance: Screening diseases for suitability. BMC Infectious Diseases. 2014;14. doi:10.1186/s12879-014-0690-1
102. Ongruk P, Siriyasatien P, Kesorn K. New key factors discovery to enhance dengue fever forecasting model. Advanced Materials Research. 2014;931–932: 1457–1461. doi:10.4028/www.scientific.net/AMR.931-932.1457
103. Balasundaram A, Bhuvaneswari PTV. Comparative study on decision tree based data mining algorithm to assess risk of epidemic. 2013. pp. 390–396. doi:10.1049/ic.2013.0344
104. Wu Y, Lee G, Fu X, Soh H, Hung T. Mining weather information in dengue outbreak: Predicting future cases based on wavelet, SVM and GA. Lecture Notes in Electrical Engineering. 2009;39 LNEE: 483–494. doi:10.1007/978-90-481-2311-7_41
105. Zhang Y, Ibaraki M, Schwartz FW. Disease surveillance using online news: Dengue and zika in tropical countries. Journal of Biomedical Informatics. 2020;102. doi:10.1016/j.jbi.2020.103374
106. Souza RCSNP. Detecting spatial clusters of infection risk with geo-located social media data. 2018.
107. Coberly JS, Fink CR, Elbert Y, Yoon I-K, Velasco JM, Tomayao AD, et al. Tweeting Fever: Can Twitter Be Used to Monitor the Incidence of Dengue-Like Illness in the Philippines? JOHNS HOPKINS APL TECHNICAL DIGEST. 2014;32: 12.
108. Gomide J, Veloso A, Meira Jr. W, Almeida V, Benevenuto F, Ferraz F, et al. Dengue surveillance based on a computational model of spatio-temporal locality of Twitter. 2011. doi:10.1145/2527031.2527049
109. Fang Z-H, Tzeng J-S, Chen CC, Chou T-C. A study of machine learning models in epidemic surveillance: Using the query logs of search engines. 2010. pp. 1438–1449.
110. Souza J, Leung CK, Cuzzocrea A. An Innovative Big Data Predictive Analytics Framework over Hybrid Big Data Sources with an Application for Disease Analytics. Advances in Intelligent Systems and Computing. 2020;1151 AISC: 669–680. doi:10.1007/978-3-030-44041-1_59
111. Yogapriya P, Geetha P. Dengue disease detection using K-means, hierarchical, kohonen-SOM clustering. International Journal of Innovative Technology and Exploring Engineering. 2019;8: 904–907. doi:10.35940/ijitee.J9066.0881019
112. Adias Sabara M, Somantri O, Nurcahyo H, Kurnia Achmadi N, Latifah U, Harsono. Diagnosis classification of dengue fever based on Neural Networks and Genetic algorithms. 2019. doi:10.1088/1742-6596/1175/1/012065
113. Jongmaenwai B, Lowanichchai S, Jabjone S. Comparision using data mining algorithm techniques for predicting of dengue fever data in northeastern of Thailand. 2019. pp. 532–535. doi:10.1109/ECTICon.2018.8619953
114. Balasaravanan K, Prakash M. Detection of dengue disease using artificial neural network based classification techniquetion. International Journal of Engineering and Technology(UAE). 2018;7: 13–15. doi:10.14419/ijet.v7i1.3.8978
115. Acosta Torres J, Oller Meneses L, Sokol N, Balado Sardiñas R, Montero Diaz D, Balado Sansón R, et al. Decision tree technique applied to the clinical method in the dengue diagnosis. Revista Cubana de Pediatria. 2016;88: 441–453.
116. Soonthornphisaj N, Thitiprayoonwongse D. Knowledge discovery on dengue patients using data mining techniques. 2016. pp. 371–375.
117. Fathima SA, Hundewale N. Comparitive analysis of machine learning techniques for classification of arbovirus. 2012. pp. 376–379. doi:10.1109/BHI.2012.6211593
118. Fathima S, Hundewale N. Comparison of classification techniques-SVM and naives bayes to predict the Arboviral disease-Dengue. 2011. pp. 538–539. doi:10.1109/BIBMW.2011.6112426
119. Long ZA, Abu Bakar A, Razak Hamdan A, Sahani M. Multiple attribute frequent mining-based for dengue outbreak. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2010;6440 LNAI: 489–496. doi:10.1007/978-3-642-17316-5_46

UNIVERSITÉ DE RENNES 1

**Titre :** Réutilisation des données de vie réelle dans la gestion des épidémies d'arboviroses en territoire Caribéen. Application à la surveillance de la dengue en Martinique.

**Mots clés :** Données de vie réelle ; Données massives en santé ; Surveillance syndromique ; Caraïbe ; Dengue

**Résumé :** Les maladies infectieuses émergentes sont devenues un enjeu majeur de santé publique ces dernières années et sont à l'origine de nombreuses crises sanitaires touchant en majorité les régions tropicales et subtropicales telles que la Caraïbe. À l'heure actuelle, les systèmes de surveillance syndromique basés sur la collecte et l'analyse de données souffrent de plusieurs limites, notamment les délais entre un événement de santé et sa notification et le coût élevé de ces systèmes, alors que l'accélération de la propagation de ces pathologies demande des systèmes de plus en plus réactifs. Des alternatives basées sur les données de vie réelle, c'est-à-dire l'ensemble des données recueillies en dehors de la recherche clinique, sont étudiées depuis plusieurs années pour tenter de répondre aux enjeux de réactivité et de disponibilité des données.

L'objectif de cette thèse est d'évaluer la place des données de vie réelle dans la surveillance syndromique et plus spécifiquement de la surveillance de la dengue dans la Caraïbe.
Cette thèse sur articles explore trois axes importants : une première partie fait un état des lieux et étudie les enjeux des systèmes de surveillance syndromique dans la Caraïbe. La seconde partie expose les verrous liés aux données de vie réelle. La troisième partie explore les verrous liés à l'exploitation des données à travers le cas d'usage de la surveillance de la dengue en Martinique. Enfin, la discussion s'intéresse aux atouts et contraintes des données de vie réelle dans la surveillance syndromique et leur place dans la préparation (preparedness) et la capacité de réponse des systèmes à répondre aux épidémies.

**Title:** Harnessing real-world data to manage arboviral disease outbreaks in the Caribbean. The dengue use case in Martinique

**Keywords :** Real-world data; Health Big Data; Syndromic surveillance; Caribbean; Dengue

**Abstract :** The incidence of emerging infectious diseases has increased considerably over the last few years, causing several health crises, especially in tropical and subtropical areas such as the Caribbean. Currently, traditional surveillance systems based on data collection are hampered by several limitations, especially delays between a case and its notification, and high costs despite their need to be responsive. Alternatives approaches based on real-world data, ie. data not collected in experimental conditions, have been studied for several years to help improve responsiveness and data availability of these systems. The aim of this thesis is to assess the role of real-world data in syndromic surveillance and more specifically dengue monitoring in the Caribbean.

This thesis explored three major axes: first, we conducted a state of the art on syndromic surveillance systems in the Caribbean and their limitations. Secondly, we identified the challenges related to real-world data. The third section explored the challenges related to data exploitation through the use case of dengue surveillance in Martinique. Finally, the discussion focused on the challenges and opportunities of real-world data in syndromic surveillance and their place in epidemic preparedness and response.