Santé – Enquête

Covid-19 (France) – Données épidémiologiques : la pénurie cachée. Aveugles sur le cheminement du virus

dimanche 12 juillet 2020, par BARNEOUD Lise (Date de rédaction antérieure : 11 juillet 2020).

Les autorités ont eu les yeux rivés sur les hôpitaux mais, faute d'épidémiologie de terrain, nous n'avions pas idée de la partie immergée de l'iceberg. Comme pour les tests ou les masques, les données ont gravement manqué.

Généralement, lorsque des embouteillages apparaissent, on analyse la configuration des rues, on modélise la circulation des voitures, puis on repense l'ensemble du réseau pour fluidifier le trafic. Dans la crise de Covid-19, seuls les lieux d'embouteillages étaient disséqués : les hôpitaux.

Nombre de nouveaux cas confirmés, nombre de personnes en réanimation, nombre de décès, nombre de respirateurs disponibles… Ces chiffres, égrenés jusqu'à deux fois par jour durant le confinement, étaient au cœur des décisions. Grâce à un énorme travail de collecte et de consolidation des données hospitalières, la « data science » a indiscutablement sauvé des vies. Mais cette masse de données s'est avérée fort inutile pour empêcher quiconque de tomber malade.

Et pour cause : nous étions aveugles sur le cheminement du virus au sein de la population, la façon dont il se propageait parmi nous, la configuration de ses routes. *« Nous avons répondu à un afflux de cas, mais nous n'avons pas répondu à l'épidémie »*, résume Emmanuel Baron, médecin épidémiologiste directeur d'Epicentre, le groupe de recherche de Médecins sans frontières.

Pour comprendre, Mediapart a enquêté dans les coulisses de la fabrique des données, en rembobinant jusqu'au début de l'épidémie.

« Nous ne savons pas nous servir d'un stéthoscope, mais nous pouvons peut-être aider ceux qui savent en leur apportant de la data facilement exploitable et de la dataviz. » Lorsque Lior Perez, ingénieur en informatique et en intelligence artificielle à Météo-France, envoie son appel à contributions [1] auprès de la communauté française des datascientists début mars, les seules données disponibles sur l'épidémie sont éparpillées dans les différents communiqués de presse des Agences régionales de santé (ARS).

On y trouve des informations sur le nombre de nouveaux cas confirmés ou le cumul des cas ; certaines ARS précisent également le département, parfois l'âge des patients et les liens entre ces patients. Mais l'OMS a beau avoir déclaré l'état de pandémie le 11 mars, nous en sommes encore, en France, au stade d'informations éparses, sous une forme exclusivement littéraire.

« Il était impossible d'utiliser ces communiqués de presse pour tracer des courbes et suivre l'évolution de l'épidémie, se souvient Lior Perez. *J'ai regardé ce qui s'était fait en Asie et j'ai*

découvert qu'en Corée du Sud, une communauté de datascientists *s'était emparée du problème et avait épluché toutes les données pour les mettre à disposition dans un même fichier informatique.* » Une demi-douzaine de *datascientists* répondent à son appel. « *Nous nous sommes répartis les régions et avons commencé à éplucher chaque communiqué de presse et à intégrer les données à la main dans un tableau, en nous aidant parfois de la presse pour retrouver des informations manquantes.* »

Bientôt, plus d'une centaine de contributeurs collectent, analysent et visualisent les données chiffrées disponibles sur les cas de Covid-19. Dont certains membres d'Etalab, un récent département de la direction interministérielle du numérique qui coordonne la politique d'ouverture et de partage des données publiques. Consécration : le 28 mars, Édouard Philippe utilise leur tableau de bord durant son point presse.

« *Le libre accès aux données relatives à l'épidémie, ce qu'on appelle l'*open data*, assure la confiance des citoyens dans les éléments qui leur sont communiqués, favorise les actions de prévention contre la propagation du virus et facilite la prise de décision* », souligne l'ex-premier ministre à cette occasion.

Durant cette même période, une équipe de *data scientists* de la Direction de la recherche, des études, de l'évaluation et des statistiques du ministère de la santé (Drees) est envoyée en renfort auprès du centre de crise. Objectif : centraliser les données existantes, les nettoyer et les mettre à disposition des décideurs sous la forme de cartes et d'indicateurs. Il faudra attendre fin mai pour que ces données soient également partagées sur une plateforme avec d'autres acteurs, notamment des chercheurs.

Les sources d'information sont multiples, mais elles ne sont pas harmonisées et sont de qualité diverse. Le Système d'information pour le suivi des victimes (SI-VIC), mis en place le 13 mars, renseigne sur le nombre de patients hospitalisés et le nombre de patients en état grave (en réanimation ou en soins intensifs). Il ne permet pas en revanche de savoir d'où viennent les patients ni de connaître leurs comorbidités.

Sont disponibles également les données issues des tests réalisés par un réseau de laboratoires, qui représentent à l'époque un tiers des laboratoires existants (seuls les cas graves étaient testés à cette époque). À cela s'ajoutent les données de recours aux urgences et aux associations SOS Médecins (sans confirmation virologique), les déclarations de décès (avec un délai de deux à trois semaines), les capacités en respirateurs des hôpitaux ou encore les lits disponibles et installés.

Problème : ces données concernent les cas graves, « l'embouteillage » visible de l'épidémie. Il y a évidemment une certaine logique à cela, explique l'épidémiologiste Mathieu Moslonka-Lefebvre : « *La plupart des modèles développés par la communauté scientifique pour éclairer la gestion de crise étaient focalisés sur cet enjeu : éviter la saturation des systèmes de soin.* » Mais ces indicateurs présentent un gros désavantage : ils ont nécessairement un train de retard sur la circulation du virus.

De fait, le Sars-Cov2 incube silencieusement pendant cinq jours en moyenne, au bout de sept jours il déclenche des symptômes graves chez environ 5 % des personnes infectées, puis il faut encore ajouter 10 jours en moyenne entre l'hospitalisation et l'éventuel décès. Une longue période mise à profit par le virus pour circuler au sein de la population générale. Or, cette masse de personnes infectées, la partie immergée de l'iceberg, nous était totalement invisible en début d'épidémie.

En France, plus de 97 % des personnes infectées n'ont pas mis le pied à l'hôpital. Pour autant, un seul indicateur hors hospitalier, issu des médecins de ville, était disponible à partir du 17 mars : le

réseau Sentinelles, créé en 1984 pour surveiller les maladies virales. Un réseau performant pour surveiller la grippe ou la varicelle, *« mais avec moins de 800 médecins généralistes impliqués dans cette surveillance, soit environ 0,8 % des praticiens, on peut facilement passer au travers d'un signal émergent et faible. Il faut inventer autre chose pour cette surveillance-là »*, explique Thierry Blanchon, directeur adjoint du réseau.

D'autant plus qu'il n'était pas possible à l'époque de tester tous les cas suspects : seuls quelques dizaines d'échantillons faisaient l'objet d'une analyse virologique chaque semaine dans le cadre de ce réseau. Pour ce chercheur de l'Institut Pierre-Louis d'épidémiologie et de santé publique, *« la médecine de ville est vraiment l'endroit où nous avons des progrès à faire »* : *« J'ai le sentiment que nous avons rapidement bénéficié de données hospitalières assez exhaustives, mais les épidémies démarrent en milieu communautaire, pas à l'hôpital. »*

Début avril, face à l'absence d'autres indicateurs en médecine de ville, le syndicat MG France lance également une série d'enquêtes auprès de 2 000 médecins généralistes, mais sans possibilité de confirmer la réalité du diagnostic, faute de test disponible. En extrapolant les résultats à l'ensemble des 60 000 médecins généralistes en France, le syndicat évaluait à plus de 1,5 million le nombre de personnes potentiellement atteintes par le Sars-Cov-2 entre le 17 mars et le 3 avril. *« À la lumière de l'expérience acquise depuis que l'on teste nos patients, je pense qu'un grand nombre de ces cas suspects n'étaient pas des Covid*, reconnaît aujourd'hui le président du syndicat, Jacques Battistoni. *Mais notre travail montrait bien la limite de la politique de l'époque, exclusivement centrée sur l'hôpital. »*

D'autres systèmes similaires ont été proposés, sans succès : la Caisse nationale d'assurance maladie ne lancera son téléservice *« contact Covid »* que le 11 mai, au moment du déconfinement, permettant aux médecins généralistes d'enregistrer les informations concernant un patient Covid et ses éventuels contacts.

« Si le peu de tests disponibles en février-mars avait été alloué à la médecine de ville plutôt qu'à l'hôpital, nous aurions pu voir venir l'épidémie beaucoup plus tôt, estime le docteur Claude Leicher, président national des Communautés professionnelles territoriales de santé (Cpts). *En Allemagne, 8 tests sur 10 étaient réalisés en médecine de ville et non à l'hôpital. »*

Mais la France regardait ailleurs, focalisée sur le sommet émergé de l'iceberg : l'hôpital et les décès. Jusqu'à la fin du mois de mars, les malades du Covid-19 n'existaient que s'ils franchissaient les murs de l'hôpital. Même les décès en Ehpad étaient encore sous les radars.

Il faut certes souligner quelques enquêtes épidémiologiques de terrain, notamment dans l'un des premiers foyers détectés, aux Contamines-Montjoie, ou encore dans l'Oise. Mais malgré l'appel à la fameuse réserve sanitaire, Santé publique France tout comme les ARS ont vite été débordées. *« Investiguer chaque cas et leur contact demande énormément de moyens humains. Très vite, ce n'était plus tenable »*, confie-t-on du côté des ARS.

« On travaille avec des outils archaïques, on a beaucoup bricolé à base d'Excel et de copier-coller, on a perdu des journées et des soirées à refaire nos tableaux à la main. Nous sommes bien loin de la start-up nation… », révèle un agent d'une ARS, qui souhaite garder l'anonymat. Ces suivis dits de contacts de cas étaient ainsi réalisés de façon très hétérogène sur le territoire et la plupart n'ont donné lieu à aucune publication.

En théorie, les données récoltées durant ces *« contact tracing »* étaient ensuite remontées *via* l'application Godata de l'OMS. Selon Santé publique France, peu loquace durant toute cette crise, des informations sur quelque 11 142 cas confirmés y étaient ainsi recensées au 17 mars, date de

début du confinement. À ce moment-là, le passage en stade 3 de l'épidémie signait l'arrêt de ces suivis. Une erreur, a récemment estimé l'ancien directeur de la santé William Dab durant son audition devant la commission d'enquête parlementaire sur la gestion de l'épidémie de Covid-19. L'épidémiologiste n'aura cessé durant toute cette crise de demander plus d'enquêtes de terrain pour comprendre comment se contaminent les personnes.

« J'en ai parlé au professeur Chêne, la directrice générale de Santé publique France, elle m'a dit : "Mais nous n'avons pas les moyens" », a-t-il rapporté durant son audition, insistant sur *« la grande faiblesse des forces de santé publique sur le terrain »*. L'épidémiologiste responsable de l'unité des infections respiratoires de Santé publique France, Daniel Levy-Bruhl, l'a également reconnu dans une interview accordée au *Monde* : *« Nous n'avions pas les outils pour identifier les chaînes de transmission individuelles et les sources de contamination. »*

« Des réflexes qui ne servent pas le bien public »

Mais ce n'est pas tout. Plusieurs acteurs soulignent également un manque de partage de ces rares informations descriptives de terrain. Des équipes de chercheurs ont ainsi demandé à accéder à certaines données, notamment la distribution des cas secondaires, qui permet de détecter d'éventuels superpropagateurs, la durée qui sépare deux infections ou encore des données à l'échelle départementale, en vain.

Même Etalab, qui est pourtant une administration publique dont la mission est *« d'utiliser les données pour améliorer l'action publique »*, n'a pu accéder aux données de la base Godata. Seuls Santé publique France, les ARS et les Centres nationaux de référence des virus respiratoire y sont autorisés.

Pour l'heure, aucune exploitation de ces données n'a encore été communiquée. *« Une publication de Santé publique France issue de ces données est en cours de préparation »*, se borne à faire savoir l'agence.

Plusieurs freins au partage semblent ici coexister. En premier lieu, la protection des données personnelles. Il est en effet interdit par la loi de publier des données qui permettraient l'identification d'une personne physique. *« Dans quelle mesure la crise du Covid-19 justifie que l'on bafoue les données privées ? »*, s'interroge le *datascientist* Lior Perez, qui comprend cette réticence au partage.

« Bien sûr qu'il faut prêter une grande attention aux données personnelles, mais, pour moi, ces histoires de risque de réidentification après anonymisation, c'est un alibi : on peut toujours anonymiser les données et les chercheurs ne sont pas du tout intéressés par l'identité des cas », critique Antoine Flahault, directeur de l'Institut de santé globale de l'université de Genève, dont l'application de modélisation est en *open source*.

Lorsqu'il dirigeait le réseau Sentinelles à la fin des années 1990, il a ouvert l'accès aux données issues de ce réseau de médecins. Trente ans plus tard, les chercheurs sont nombreux à les utiliser. *« À l'époque, on m'a dit que j'étais fou, que c'était risqué politiquement, que j'allais tuer le tourisme, etc. Les réticences au partage sont bien souvent motivées par d'autres raisons que le respect de la vie privée… »*

Parmi ces *« autres raisons »*, certains soulignent le risque d'une perte de pouvoir et la crainte d'être décrédibilisé face à des données de mauvaise qualité ou trop éparses. Certains hôpitaux évitent par exemple de partager leurs données car leur codage est trop erratique. Même son de cloche pour la

base Godata, *« assez pauvre et remplie de façon hétérogène »*. Des ARS et certains laboratoires de dépistage apparaîtraient inévitablement sous un mauvais jour. *« Ouvrir ces données reviendrait à montrer qu'ils ont mal fait leur boulot »*, estiment plusieurs sources proches du dossier.

Autre frein identifié : les partenariats exclusifs avec certains chercheurs, dans le but de publier en premier. *« Les données, c'est de l'or, ça génère des réflexes qui ne servent pas le bien public »*, confirme Antoine Flahault. Des données assez basiques comme la durée d'hospitalisation ou le temps entre le début de l'infection et l'entrée en réanimation n'ont ainsi été partagées qu'après la publication dans la revue *Science* de l'équipe de Simon Cauchemez, de l'Institut Pasteur, associée pour l'occasion à Santé publique France, au ministère de la santé ou encore à l'Institut Pierre-Louis d'épidémiologie et de santé publique.

« C'est dommage, car cela a bridé la recherche et au final empêché de comparer les prévisions de différents modèles », souligne Samuel Alizon, spécialiste en modélisation des maladies infectieuses au Cnrs, qui refuse actuellement de commenter l'actualité de l'épidémie de Covid pour alerter sur le projet de loi de programmation pluriannuelle de la recherche *(voir nos articles sur le sujet ici, là ou encore là).*

L'Institut Pasteur, qui a un statut de fondation privée avec une reconnaissance d'utilité publique, est également l'exclusif destinataire d'une autre source précieuse de données, celle du site maladiecoronavirus.fr. Lancé le 18 mars, ce site d'évaluation des symptômes et de préconisations d'orientation recevait jusqu'à 15 000 connexions par seconde durant la deuxième quinzaine de mars ! Quelques 4,5 millions de français ont ainsi répondu au questionnaire en ligne. Des informations potentiellement utiles pour la gestion de l'épidémie, car on consulte bien souvent Google avant son médecin…

*« Ma volonté, c'était de faire de l'*open data*, mais nous avons préféré jouer la sécurité de la donnée, car il y avait un risque de réidentification possible, surtout dans de petites communes avec peu de cas »*, dit Florian Le Goff, de la start-up Kelindi, à l'origine du projet. Les données recueillies sont ainsi séquestrées dans un hub numérique français certifié, chez Docapost, et seul l'Institut Pasteur a accès aux données brutes.

Là encore, aucune exploitation de ces données n'a encore été publiée. L'outil aurait également pu s'avérer intéressant pour détecter en temps réel l'émergence de *clusters* (foyers infectieux). Dans cette optique, Florian Le Goff avait d'ailleurs contacté les ARS pour leur proposer de participer au projet, *« mais on* [lui] *répondait : "On est en train de chercher des masques…" »* : *« Je n'ai pas ressenti d'appétence de leur part pour ces données. »*

« On assiste à une sorte de ruée vers l'or avec ces données liées au coronavirus », estime le chercheur Samuel Alizon, qui souligne même que des informations plutôt bien partagées en début d'épidémie, comme les séquences génétiques du virus, le sont de moins en moins : *« Le partage des informations est en train de se tarir et il faudra probablement attendre plusieurs mois que des articles de recherche soient publiés pour un nouveau partage. »*

L'Institut dirigé par Didier Raoult, l'IHU Méditerranée infection à Marseille, annonce par exemple avoir séquencé 494 génomes complets du Sars-Cov-2, mais ces séquences n'ont toujours pas été partagées, alors qu'il existe une plateforme internationale baptisée Gisaid qui vise précisément à mettre en commun toutes les séquences pour mener des comparaisons phylogénétiques du virus. Fin juin, la France avait publié 394 génomes dans cette base, soit à peine 1 % du nombre total de génomes partagés au niveau international…

« Il faut investir dans la collecte, le nettoyage et le traitement des données », conclut

l'épidémiologiste américain John Ioannidis, directeur du centre de prévention de l'université Stanford pour qui le manque de données disponibles et leur mauvaise qualité ont mis en échec les modélisations du Covid-19. Les prévisions de Neil Fergusson, de l'Imperial College de Londres, qui ont joué un rôle clé dans les décisions de confinement des pays européens, n'échappent d'ailleurs pas à ces critiques (voir notamment l'émission Newsnight de la BBC ou l'interview de Johan Giesecke, ex-épidémiologiste en chef de la Suède, dans le magazine *Unherd*).

John Ioannidis souligne également le manque d'interdisciplinarité dans la collecte des données : *« Presque tous les modèles qui ont joué un rôle de premier plan dans la prise de décision ne se sont concentrés que sur une ou quelques données, comme les décès ou les besoins hospitaliers. »*

« Nous avions un focus trop biologique, alors que nous avons désespérément besoin de données en science sociale, confirme Laurent Hébert-Dufresne, spécialiste de l'« épidémiologie sur réseau » au sein du laboratoire interdisciplinaire du Vermont Complex Systems Center, aux États-Unis. *On oublie trop souvent qu'une épidémie est la résultante de caractéristiques propres au virus, aux individus, à l'environnement, aux interventions mises en place… »*

La structure des interactions sociales, c'est précisément toutes les rues et les ruelles empruntées par le virus, pour reprendre la comparaison de Gianluca Manzo, sociologue au sein du Groupe d'étude des méthodes de l'analyse sociologique de la Sorbonne *(lire son article ici)*. Sans visibilité sur ce réseau, une seule recommandation est envisageable pour contrôler les flux, écrit ce chercheur du Cnrs : *« Tout le monde doit rester chez soi. »*

Là encore, une initiative citoyenne a tenté de combler ce point aveugle : l'association Datacovid, cofondée par Mathieu Moslonka-Lefebvre, ingénieur passé par la recherche en épidémiologie mathématique. Grâce à une levée de fonds privés, l'association, en partenariat avec Ipsos, mène depuis le 7 avril un sondage hebdomadaire sur 5 000 personnes, offrant des données en *open data* sur l'évolution du nombre de contacts proches (à moins de 1 mètre), la durée et les raisons de sortie durant le confinement, l'application des gestes barrières, etc.

Autant d'informations cruciales pour commencer à dessiner plus finement l'infrastructure routière par laquelle voyage le virus. Et sortir de ce focus sur les embouteillages, qui ne nous a manifestement pas permis d'enrayer l'épidémie.

Lise Barnéoud

P.-S.

• « Données épidémiologiques : la pénurie cachée ». MEDIAPART. 11 juillet 2020 : https://www.mediapart.fr/journal/france/110720/donnees-epidemiologiques-la-penurie-cachee?onglet=full

Si vous avez des informations à nous communiquer, vous pouvez nous contacter à l'adresse enquete mediapart.fr. Si vous souhaitez adresser des documents en passant par une plateforme hautement sécurisée, vous pouvez vous connecter au site frenchleaks.fr.

Notes

[1] https://www.linkedin.com/posts/liorperez_opendata-covid19-dataset-activity-6642013973241704448-55L2