Comment vulgariser les données du Covid ?

Par Nicolas BERROD
Journaliste au *Parisien*

Depuis le début de la pandémie de Covid-19, au début de l'année 2020, les données sont indispensables pour suivre l'évolution de l'épidémie et tenter d'anticiper la suite des événements. Taux d'incidence, nombre de patients hospitalisés, part de chaque variant du Sars-CoV-2 parmi les nouveaux cas positifs, couverture vaccinale… Ces indicateurs sont fournis à un rythme quotidien, hebdomadaire ou mensuel par plusieurs organismes officiels. En tant que journalistes, nous avons comme mission de les utiliser et de les vulgariser dans nos articles et commentaires publiés sur les réseaux sociaux. Et lorsque l'on travaille au *Parisien*, journal très grand public, il nous faut aussi parvenir à rendre ces données accessibles à tous. Voici comment nous avons procédé.

Plusieurs sources de données

Les choses seraient simples si toutes les données nécessaires étaient disponibles à un seul endroit. Malheureusement, mais sans surprise, ce n'est pas le cas. Voici les principales sources disponibles.

Santé publique France

L'agence sanitaire nationale publie, quotidiennement, de très nombreux indicateurs liés à la pandémie de Covid-19 : taux d'incidence (c'est-à-dire nombre de cas positifs pour 100 000 habitants sur la semaine écoulée), nombre de patients hospitalisés, nombre de personnes recevant une première dose, une deuxième dose, ou une dose de rappel de vaccin quotidiennement, part de chaque mutation détectée dans les tests positifs passés au criblage, etc. Regroupés en différentes bases de données (SI-DEP pour les tests, SI-VIC pour l'hôpital et VAC-SI pour la vaccination), ils sont généralement disponibles à l'échelle départementale et régionale, sur Data.gouv.fr et sur le site Géodes[1].

Santé publique France est, de très loin, le premier fournisseur de données Covid dans le pays. Néanmoins, à plusieurs reprises, nous avons eu du mal à comprendre que des données publiées dans ses rapports hebdomadaires ne soient pas disponibles en *open data*, nous empêchant ainsi de les exploiter facilement. Je pense, par exemple, à la part de cas positifs symptomatiques et asymptomatiques.

L'Assurance maladie

Fin mai 2021, l'Assurance maladie a mis en ligne ses propres données de vaccination, téléchargeables en fichiers facilement exploitables[2]. Le principal avantage, par rapport aux indicateurs publiés par Santé publique France, est que l'on dispose des couvertures

[1] https://geodes.santepubliquefrance.fr/

[2] https://datavaccin-covid.ameli.fr/pages/home/

vaccinales par type de comorbidités. Ainsi, il est possible d'estimer la part de personnes obèses (particulièrement à risque de forme grave) vaccinées et donc non vaccinées, par exemple.

La Drees

La Direction de la recherche, des études, de l'évaluation et des statistiques (Drees) s'est, notamment, fait connaître pour ses données d'efficacité de la vaccination. À partir de l'été 2021, l'organisme dépendant directement du ministère de la Santé a publié chaque semaine, en *open data*[3], les nombres de patients testés positifs, hospitalisés, admis en soins critiques et décédés en fonction de leur statut vaccinal. En rapportant ces valeurs à la population de chaque groupe (les non-vaccinés, les vaccinés avec une seule dose, les vaccinés avec dose de rappel, etc.), on obtient des taux rapportés à un même nombre d'habitants. Il est possible d'en déduire, par exemple, que le risque d'hospitalisation apparait X fois moins élevé chez une personne vaccinée que chez une autre n'ayant reçu aucune dose de vaccin. Mais ces données comportent de nombreuses limites, sur lesquelles nous reviendrons dans la suite de cet article.

Mais aussi l'Insee, l'Inserm, l'OMS, Our World in Data…

D'autres institutions officielles fournissent des données utiles pour le suivi de la pandémie de Covid-19. L'Insee, par exemple, « produit » le nombre de décès toutes causes confondues recensés chaque jour en France[4]. Les vagues de Covid-19, et notamment les premières d'entre elles, ont évidemment eu un lourd impact. Le Centre d'épidémiologie sur les causes médicales de décès (CépiDc) de l'Inserm, de son côté, fournit le nombre de décès liés au Covid-19, c'est-à-dire pour lesquels le terme « Covid » apparaît sur le certificat de décès[5].

Enfin, des sources de données internationales se sont également révélées très utiles. L'Organisation mondiale de la santé (OMS), par exemple, met régulièrement à jour les nombres de cas positifs et de décès recensés chaque jour dans de nombreux pays dans le monde[6]. Ces informations sont également visibles sur la plateforme Our World in Data[7], qui facilite les comparaisons internationales à partir d'un très grand nombre d'indicateurs.

Comment vulgariser les données ?

De précieuses infographies

L'un des premiers réflexes face à des données chiffrées listées dans un tableau, c'est de se demander sous quelle forme les rendre visuelles et les mettre en images. Nous avons la chance, au *Parisien*, de disposer d'un service infographie très efficace qui peut réaliser de nombreuses figures variées.

Le type de graphique le plus classique, et qui « fonctionne » toujours très bien, est la courbe. Ceci vaut pour l'évolution du taux d'incidence dans plusieurs départements, par

[3] https://data.drees.solidarites-sante.gouv.fr/explore/dataset/covid-19-resultats-par-age-issus-des-appariements-entre-si-vic-si-dep-et-vac-si/export/

[4] https://www.insee.fr/fr/statistiques/6206305

[5] https://opendata.idf.inserm.fr/cepidc/covid-19/

[6] https://covid19.who.int/data

[7] https://ourworldindata.org/explorers/coronavirus-data-explorer

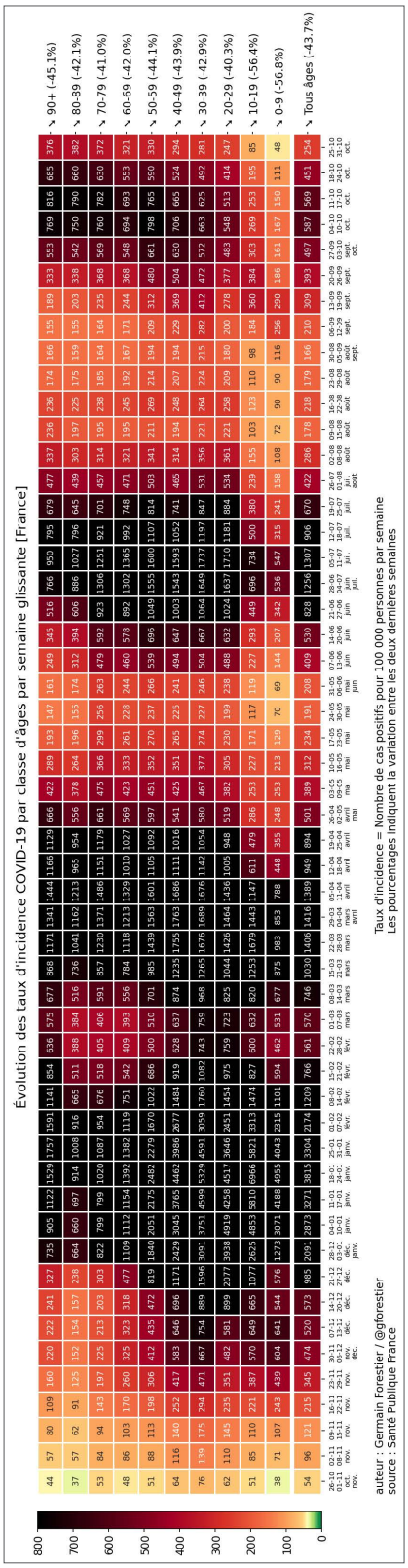

Figure 1. Évolution des taux d'incidence de Covid-19 par classe d'âges par semaine glissante en France (© Germain Forestier, Source : Santé publique France).

exemple, ou bien pour le nombre de patients hospitalisés à deux périodes différentes. La couverture vaccinale peut apparaître sous forme de graphique à barres, en utilisant une échelle de 100 %, ou bien sous forme de carte géographique afin de montrer les différences département par département.

Il est aussi possible d'imaginer des cartes de chaleur, avec un code couleur permettant de visualiser d'emblée les périodes où les valeurs sont les plus élevées (type de figure, notamment, exploitée dès les premiers mois de la pandémie par l'enseignant-chercheur Germain Forestier[8]).

Attention aux termes employés

« Vulgariser » les données passe aussi par bien expliquer ce dont on parle. Ainsi, le grand public n'était évidemment pas habitué, dans un premier temps, au terme « taux d'incidence », par exemple. Il est donc important de bien expliciter les termes employés[9]. Cela vaut de fait pour le « taux d'incidence », mais aussi pour le « facteur de réduction du risque d'hospitalisation » grâce à la vaccination, ou encore pour le nombre d'admissions quotidiennes « pour Covid » et « avec Covid ».

Afin d'être le plus « grand public » possible, des images peuvent être employées. Par exemple, si le taux d'incidence atteint 1 000 dans un département donné, cela signifie qu'un habitant sur 100 y a été testé positif durant la semaine qui vient de s'écouler.

Prendre garde aux interprétations

Il est souvent possible, à partir d'une même donnée, de lui faire dire des choses différentes. Reprenons notre exemple d'un taux d'incidence de 1 000 dans tel département à tel jour. D'un côté, ce nombre est très élevé et il reflète une

[8] https://germain-forestier.info/covid/index.html

[9] « Covid-19 : taux d'incidence, R, hospitalisations… Ces indicateurs à scruter », *Le Parisien*, 17 juillet 2020, https://www.leparisien.fr/societe/covid-19-taux-d-incidence-r-hospitalisations-ces-indicateurs-a-scruter-17-07-2020-8354538.php

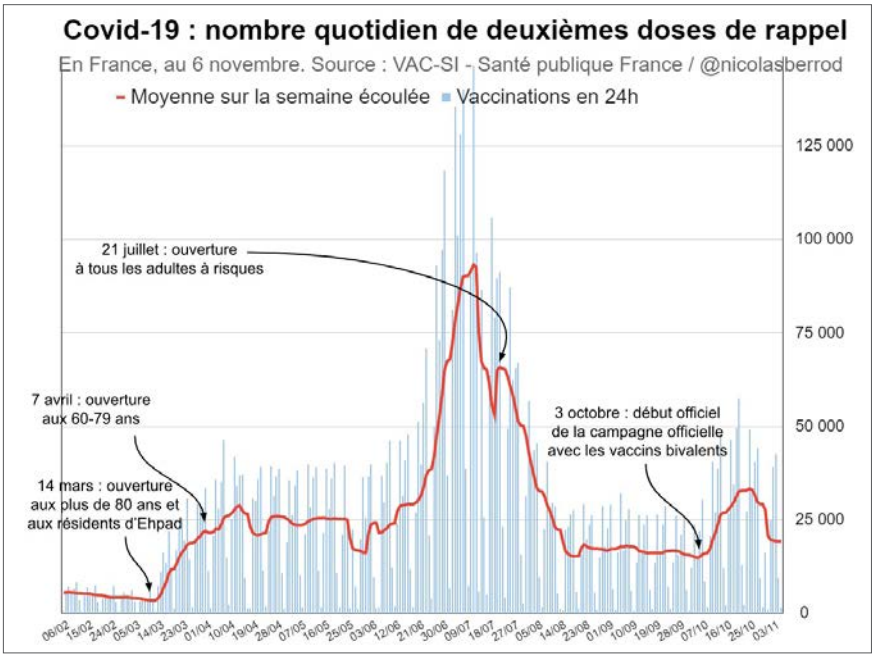

Figure 2. Nombre quotidien de deuxièmes doses de rappel contre le Covid-19 en France au 6 novembre 2021 (© Nicolas Berrod, Source : Santé publique France).

forte circulation du virus. Mais d'un autre côté, si ce taux a fortement diminué durant les jours précédents, ceci constitue l'information la plus importante.

Des vaccins soi-disant inefficaces

Un autre exemple très marquant porte sur les données de la Drees. En novembre 2021, les vaccinés sont devenus plus nombreux que les non-vaccinés parmi les patients diagnostiqués Covid-19 admis chaque jour à l'hôpital. Certaines personnes en ont déduit que la vaccination ne marchait pas, à tort. En effet, plus la couverture vaccinale d'une population augmente, plus la part de patients vaccinés parmi ceux hospitalisés est susceptible d'augmenter elle aussi.

Prenons le cas extrême (et purement théorique) d'une population vaccinée à 100 %. Comme les vaccins ne protègent pas à 100 % contre les formes graves (mais plutôt autour de 90 % durant les premiers mois), certaines personnes tomberaient malades et seraient hospitalisées. Au final, 100 % des patients admis à l'hôpital seraient… vaccinés. Pour bien interpréter ces données, il faut ainsi rapporter le nombre de personnes hospitalisées à la « population » de chaque groupe, celui des vaccinés et celui des non-vaccinés[10].

Gare aux biais d'interprétation

Un autre piège à éviter porte sur les biais, qu'il faut toujours avoir en tête. Prenons les admissions quotidiennes à l'hôpital. Les valeurs pour le samedi et le dimanche sont toujours beaucoup plus basses que les autres jours de la semaine. La raison est simple :

[10] « Covid-19 : pourquoi y a-t-il désormais plus de vaccinés que de non-vaccinés admis à l'hôpital ? », *Le Parisien*, 18 novembre 2021, https://www.leparisien.fr/societe/covid-19-pourquoi-y-a-t-il-desormais-plus-de-vaccines-que-de-non-vaccines-a-lhopital-18-11-2021-4OUPPBEWANBRXMC6EIZWS725CU.php

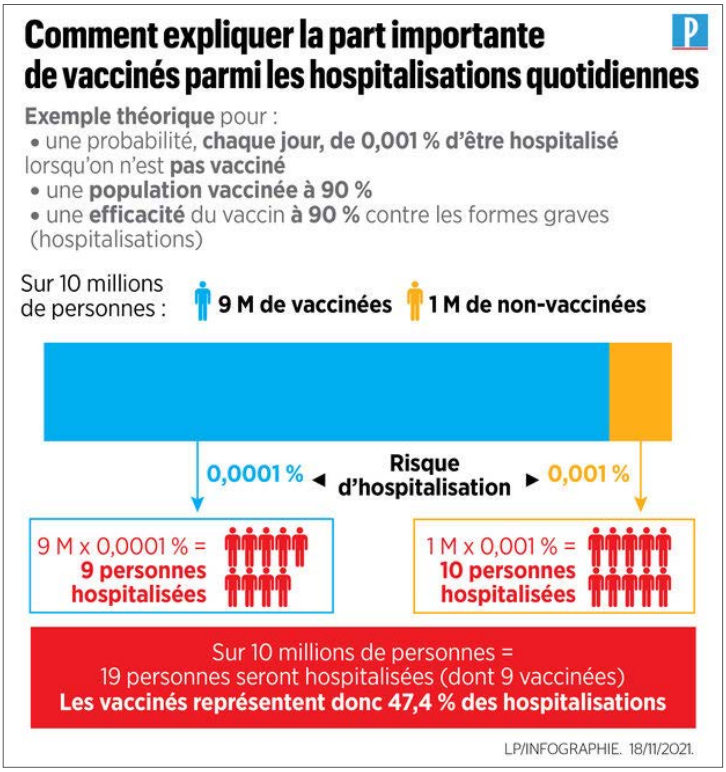

Figure 3. La part des vaccinés parmi les hospitalisations quotidiennes (Source : *Le Parisien*).

le personnel est moins nombreux le week-end, donc il a moins facilement l'opportunité et la possibilité de faire « remonter » les données dans la base SI-VIC[11]. Se contenter du chiffre annoncé le samedi et le comparer à celui de la veille pourrait conduire à dire que les admissions à l'hôpital sont en baisse.

Il faut toujours calculer la moyenne glissante sur sept jours, généralement sur la semaine écoulée, afin de lisser les variations quotidiennes. Mais ceci ne résout pas tout. Lorsqu'un jour férié tombe en semaine, ce qui arrive plusieurs fois chaque année, ces moyennes glissantes vont être touchées puisque le jour férié en question (avec très peu de tests réalisés ou de remontées d'informations) va remplacer un jour « normal ». Là aussi, mentionner ce biais est primordial[12].

Et si des données s'avèrent erronées ?

Lorsque l'on utilise des données fournies par des instances officielles, on part du principe qu'elles sont correctes. Il est cependant arrivé, à plusieurs reprises, qu'elles s'avèrent erronées. Ce qui implique d'y revenir *a posteriori*. Deux exemples permettent de l'illustrer.

[11] « Coronavirus : pourquoi y a-t-il moins de morts annoncés le dimanche ? », *Le Parisien*, 18 avril 2020, https://www.leparisien.fr/societe/coronavirus-pourquoi-y-a-t-il-moins-de-morts-annonces-le-dimanche-18-04-2020-8301816.php

[12] « Covid-19 : pourquoi le taux d'incidence va artificiellement chuter ce jeudi soir ? », *Le Parisien*, 8 avril 2021, https://www.leparisien.fr/societe/covid-19-pourquoi-le-taux-dincidence-va-mecaniquement-chuter-ce-jeudi-soir-08-04-2021-YDM7FWYIM5HW7ALOH3RVK2WSJU.php

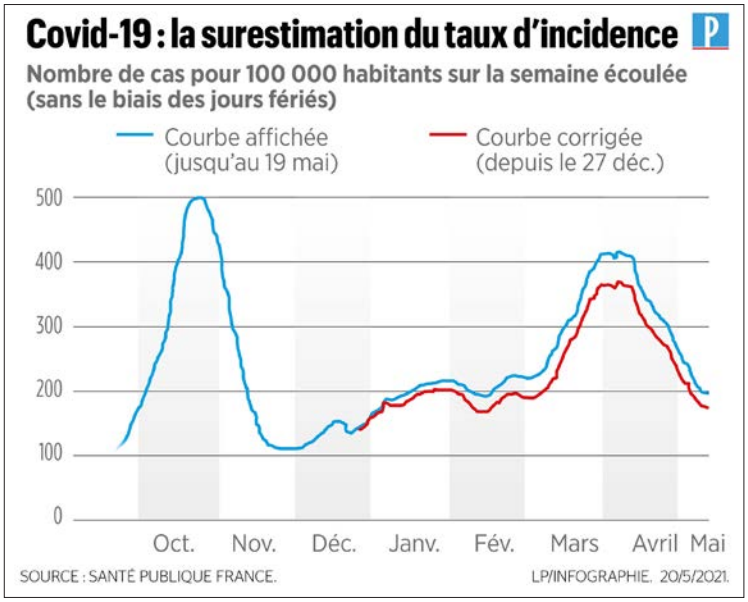

Figure 4. La surestimation du taux d'incidence
(Source : Santé publique France dans *Le Parisien*).

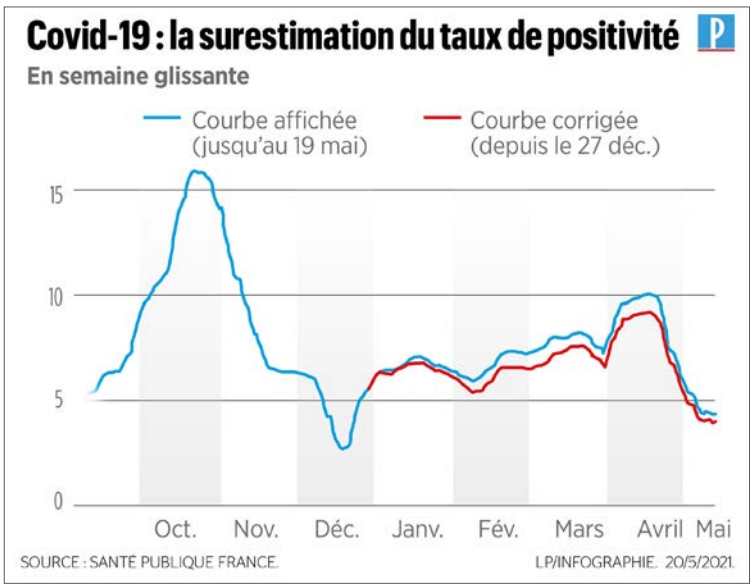

Figure 5. La surestimation du taux de positivité du Covid-19
(Source : Santé publique France dans *Le Parisien*).

Taux d'incidence surestimé

En mars 2021, nous apprenons que le taux d'incidence en France est surestimé « de l'ordre de 10 % ». Lorsqu'une même personne se faisait tester à deux reprises durant un court intervalle de temps, par exemple avec un test antigénique puis un test PCR (afin de

« cribler » le prélèvement antigénique positif), elle pouvait apparaître à deux reprises dans le décompte des cas positifs quotidiens. L'explication est simple : il suffisait d'une faute de tiret ou d'accent lors de la première étape, par exemple, pour que la même personne ne soit pas reconnue à celle d'après[13].

Cet écart a été corrigé deux mois plus tard. On pourrait penser que cela n'a pas eu d'effet sur le quotidien des Français. En réalité, lorsque le taux d'incidence dépassait le seuil d'alerte de 250, le gouvernement pouvait décider de placer le territoire sous surveillance renforcée et de prendre telle ou telle mesure. Certains territoires ont donc pu être concernés… à tort[14].

Quelle part de non-vaccinés ?

L'autre exemple porte sur les couvertures vaccinales, et en particulier les parts d'habitants non vaccinés. Plusieurs sources fournissent des estimations : Santé publique France, la Drees, l'Assurance maladie… Fin 2022, on atteignait généralement environ 7 % de personnes majeures en France n'ayant reçu aucune dose de vaccin.

En réalité, d'après une analyse fouillée de la Drees parue en octobre 2022[15], cette part serait plutôt comprise entre 8,5 et 12 %. Il fallait, tout d'abord, prendre en compte les individus décédés au fil du temps. Ensuite, il était indispensable de sortir du décompte les personnes vaccinées (et donc qui apparaissent dans la base VAC-SI)... mais qui n'habitent pas en France, à commencer par les personnes résidentes à l'étranger et vaccinées lors d'un passage en France. Après plusieurs autres corrections méthodologiques, on obtient des estimations par tranche d'âge sans doute plus conformes à la réalité.

Le syndrome de l'« enfant gâté »

Par rapport à d'autres pays, nous avons la chance, en France, de disposer d'un grand nombre de données Covid mises à jour régulièrement. Cependant, on aimerait toujours en avoir plus. C'est ce que l'on appelle parfois, pour plaisanter entre nous, le syndrome de « l'enfant gâté ». Et on regarde jalousement ce qui peut être fait dans certains pays étrangers.

Se battre pour obtenir des données

Certaines informations figuraient, au début de la pandémie, dans les points épidémiologiques de Santé publique France, mais elles n'étaient pas disponibles dans les données en *open data* et n'étaient donc pas exploitables. On peut de nouveau citer, par exemple, la part de cas positifs symptomatiques et asymptomatiques parmi tous ceux recensés chaque jour. Malgré nos demandes, nous n'avons jamais pu obtenir ces données.

En revanche, Santé publique France a satisfait à plusieurs reprises à nos attentes. On sait depuis le début de l'épidémie qu'une partie des malades diagnostiqués Covid-19 et admis chaque jour à l'hôpital ou en soins critiques sont infectés par le Sars-CoV-2… mais

[13] « Covid-19 : pourquoi le taux d'incidence est surestimé de l'ordre de 10 % », *Le Parisien*, 14 mars 2021, https://www.leparisien.fr/societe/covid-19-pourquoi-le-taux-dincidence-est-surestime-de-lordre-de-10-14-03-2021-OSXEB73FGRH6JMQXVIQYIMFWHI.php?ts=1667746479061

[14] « Cas de Covid surestimés : la fin d'un écart qui durait depuis plusieurs mois », *Le Parisien*, 20 mai 2021, https://www.leparisien.fr/societe/cas-de-covid-19-surestimes-la-fin-dun-bug-qui-a-fausse-les-chiffres-pendant-des-mois-20-05-2021-MMVPEY2IKNGVXNPX7ZS4QS7FBU.php

[15] https://drees.solidarites-sante.gouv.fr/publications-communique-de-presse/drees-methodes/les-taux-de-personnes-vaccinees-et-non-vaccinees

soignés pour un autre motif. C'est ce que l'on appelle les malades « avec Covid »[16]. Pendant près de deux ans, cette distinction n'existait pas en *open data*.

Fin 2021, lorsque le variant Omicron – moins virulent et entraînant énormément d'infections – est arrivé, on s'attendait à ce que la part de malades « avec Covid » augmente. Il nous semblait alors primordial d'obtenir cette information, notamment pour la faire apparaître sur les graphiques. Après de multiples sollicitations, Santé publique France a publié fin janvier 2022 ces données en *open data*, par région et par tranche d'âge.

Les yeux rivés vers l'Angleterre

L'Angleterre nous a toujours fait envie, en raison du grand nombre et de la qualité impressionnante des rapports et des *data* que ses agences officielles produisent sur la pandémie de Covid-19. Le tableau de bord de l'UK Health Security Agency est particulièrement bien fait et il est très simple d'y télécharger les jeux de données que l'on souhaite, par exemple[17].

Par ailleurs, l'Office for National Statistics publie chaque semaine des taux de prévalence et d'incidence réels, c'est-à-dire la part de population porteuse du virus à l'instant T ou bien nouvellement infectée sur une période précise[18]. Ces données sont capitales, car elles permettent de contourner le biais lié au nombre de tests réalisés, qui influence fortement le nombre de cas positifs recensés. En France (et dans beaucoup d'autres pays), aucune enquête de ce type sur un large échantillon d'habitants n'est réalisée.

Conclusion

Exploiter et analyser les données disponibles demande de l'imagination, pour les rendre les plus visuelles et compréhensibles possibles, mais aussi de l'attention, afin de rester conscients des biais et des limites possibles. Cela nécessite, aussi, que les journalistes soient formés aux bases statistiques et scientifiques.

[16] « Un quart de "Covid accessoire" : qui sont ces patients admis à l'hôpital pour un autre motif ? », *Le Parisien*, 22 janvier 2022, https://www.leparisien.fr/societe/sante/un-quart-de-covid-accessoire-qui-sont-ces-patients-admis-a-lhopital-pour-un-autre-motif-22-01-2022-3USJV5TLWRAT7J37DUFEGVIAOY.php

[17] https://coronavirus.data.gov.uk/

[18] https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases