24e Congrès des économistes

Approche mathématique d'une épidémie : modélisation, *machine learning* et données de contacts sociaux

Nicolas Franco (UNamur & UHasselt)

La modélisation mathématique des maladies infectieuses est un outil qui a connu un essor important avec la pandémie de COVID-19. Elle permet d'estimer certaines caractéristiques encore inobservées du virus et de projeter les effets potentiels de différents scénarios. Elle fait appel à la fois à des techniques complexes de mathématiques appliquées, à des puissants calculs informatiques ainsi qu'à l'analyse de nombreuses données incluant des données de contacts sociaux. Elle peut être un élément important pour les prises de décision relatives à la santé publique à condition de bien comprendre son utilisation et ses limitations.

Il y a une réelle difficulté pour le grand public et les instances politiques à imaginer que les mathématiques puissent être d'une utilité concernant un domaine a priori médical. Cela vient probablement du fait que l'idée des mathématiques est souvent réduite à une notion de « calcul » telle que vue lors de la scolarité, ou au mieux de statistique. Or les sciences mathématiques permettent de créer des modèles beaucoup plus complexes faisant intervenir simultanément des domaines très variés comme les systèmes dynamiques, l'optimisation, l'inférence statistique, l'analyse de données et la programmation scientifique.

Complexité et hétérogénéité sont la clé

Le modèle mathématique épidémiologique le plus simple est le modèle SIR divisant la population en trois compartiments : les individus susceptibles d'être infectés (S), les individus infectieux (I) et les individus rétablis (R). Des paramètres, mesurant le taux ou la transmission d'un compartiment à un autre, doivent être estimés en fonction des données disponibles. Le modèle permet alors d'extrapoler une projection de l'évolution au-delà des données connues. Ce modèle a été utilisé par de nombreux « apprentisépidémiologistes » sur les réseaux sociaux durant l'épidémie de COVID-19. Cependant, excepté à un stade préliminaire où les données sont trop peu nombreuses, un tel modèle ne peut réellement être utilisé. Il est beaucoup trop simplifié pour correspondre à la réalité du terrain et peut conduire à des résultats très erronés. Trois problèmes majeurs sont présents dans un tel modèle.

Le premier problème est l'estimation du nombre réel d'individus présents dans le compartiment représentant les personnes infectées (I). Les tests PCR positifs annoncés sans cesse dans les médias sont dépendants du nombre de tests effectués et donc complètement biaisés. Un modèle rigoureux doit se baser sur des indicateurs nettement plus fiables tels que le nombre de personnes hospitalisées, le nombre de personnes décédées et les tests sérologiques. Ces indicateurs possèdent cependant leurs propres biais mais qui peuvent être estimés et corrigés : par exemple, le nombre annoncé d'hospitalisations est en moyenne sous-estimé d'environ 17% en raison notamment de patients initialement admis pour une autre raison et finalement transférés en unité COVID-19. Afin de solutionner ce premier problème, il convient de rajouter des compartiments à notre modèle : un compartiment comprenant les personnes hospitalisées (Q) et un autre comprenant les personnes décédées (D). Ces nouveaux compartiments pourront être calibrés sur les données réelles tandis que le

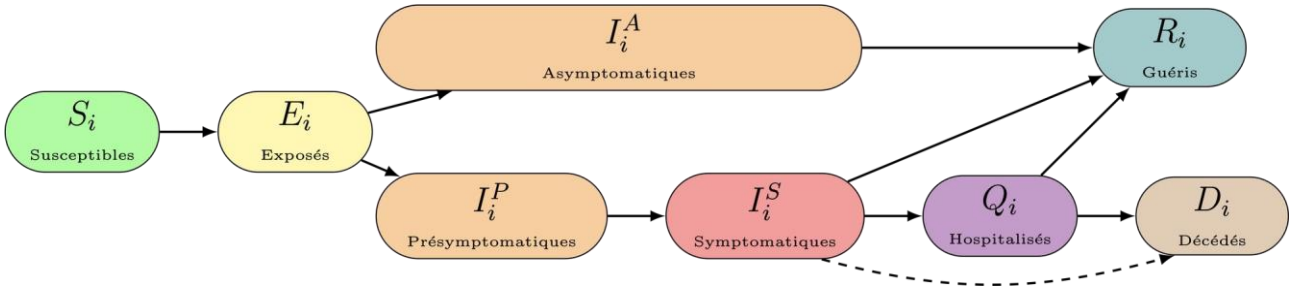
compartiment des infectés (I) restera estimé.

Second problème, une infection au COVID-19 se divise en différentes phases : une période de latence non contagieuse, une période asymptomatique contagieuse suivie d'une période symptomatique avec une contagiosité plus forte. De plus, certaines personnes perdurent asymptomatiques pendant la totalité de leur période infectieuse. Ces particularités de la COVID-19 peuvent être prises en compte en séparant le compartiment I en différents compartiments représentant les périodes d'infection.

Enfin, comme troisième problème, le virus se comporte de façon différente suivant l'âge des individus. Ainsi, la plupart des paramètres (comme les taux de guérison, d'hospitalisation ou de décès, la probabilité de faire une maladie complètement asymptomatique, etc.) doivent avoir des valeurs différentes suivant l'âge. La solution est d'utiliser des modèles structurés en âge, c'est-à-dire où tous les compartiments (et les paramètres associés) sont multipliés par le nombre de groupes d'âge choisis (par exemple par tranche de 10 ans). D'autres améliorations peuvent encore être ajoutées, comme la séparation des hospitalisés simples et des soins intensifs, de considérer certains paramètres variables dans le temps (par exemple l'amélioration des soins hospitaliers), de séparer l'épidémie au sein des maisons de retraites du reste de la population, d'estimer la réimportation du virus par les vacanciers, de prendre en compte les effets des nouveaux variants et de la vaccination, etc.

Un exemple de modèle étendu et hétérogène est présenté à la Figure 1. Il existe en tout trois modèles de ce type en Belgique (Abrams et al. 2021, Franco 2021, Alleman et al. 2020), le premier ayant la particularité supplémentaire d'être stochastique, c'est-à-dire que les passages d'un compartiment à un autre se font suivant un processus aléatoire. Il existe également d'autres types de modèles complexes, comme un modèle métapopulation (Coletti et al. 2021) et un modèle dit individuel ou à base d'agents (Willem et al. 2021). Ce dernier type de modèle intègre une simulation de tous les individus comme des entités individuelles avec des contacts et transmissions aléatoires, et est idéal pour étudier l'impact potentiel de mesures de restrictions individuelles telles que les bulles sociales, les quarantaines ainsi que l'effet potentiel de stratégies de vaccination.

Figure 1 : Schéma d'un modèle compartimental étendu hétérogène. Les indices représentent les différentes classes d'âge



La nécessité des données de contacts sociaux

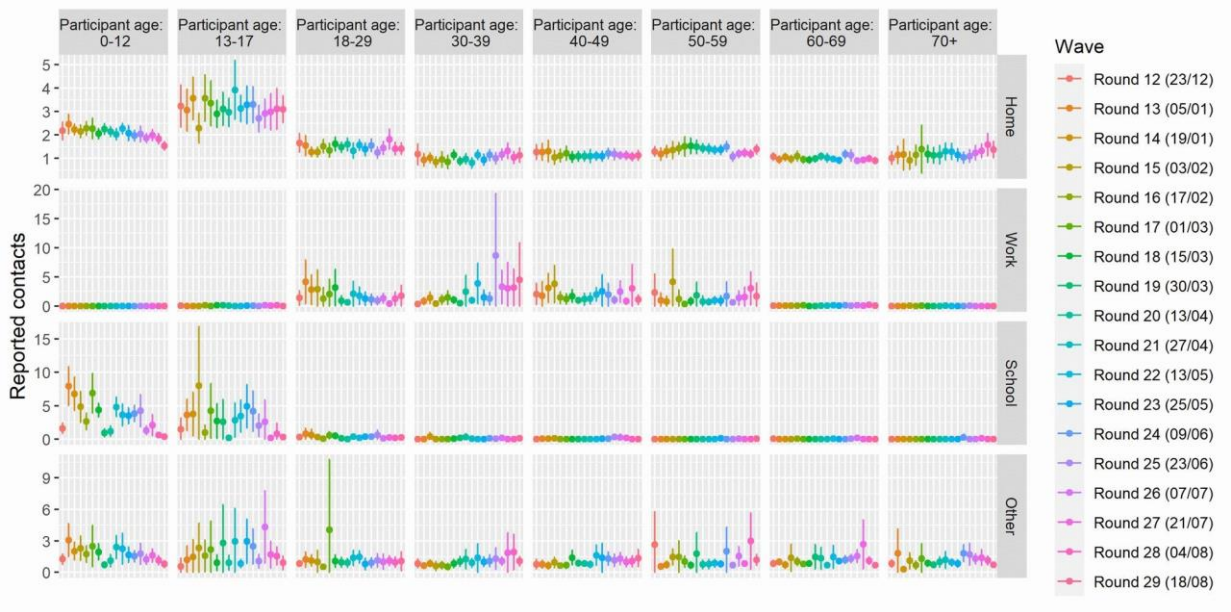
La transmission d'un compartiment d'individus susceptibles vers un compartiment d'individus infectés se fait sous l'hypothèse que chaque contact rapproché peut mener à une infection avec une certaine

probabilité estimée. Plus un individu infectieux rencontrera d'individus susceptibles, plus il générera de nouvelles infections. De même plus un individu susceptible rencontrera un individu infectieux, plus il aura de chance d'être infecté. Il est donc nécessaire d'avoir une idée du nombre moyen de contacts journaliers entre les individus, sachant que ce nombre peut varier suivant la densité de population du pays et ses habitudes sociales propres.

Les principaux modèles utilisés étant structurés en âge, il convient de dissocier les contacts entre les différentes classes d'âge car ceux-ci peuvent varier fortement. Mais si un modèle possède 10 classes d'âges, cela représente 100 paramètres de contacts différents. De plus, en cas de mesures de restriction, certaines classes d'âge peuvent être impactées plus que d'autres, par exemple en cas de fermeture des écoles, d'une obligation de télétravail ou de restriction des activités de loisirs. Il est impossible de pouvoir estimer plusieurs centaines de paramètres de contacts sans qu'ils ne soient informés par des données réelles.

Ces données sont fournies par des « matrices de contacts sociaux » qui sont des matrices obtenues par enquêtes et qui contiennent une estimation des contacts journaliers moyens entre chaque classe d'âge et lors de différentes activités distinctes (travail, école, famille, loisirs, transport, autre). Une première enquête de grande ampleur, intitulée POLYMOD, a eu lieu en 2005-2006 dans huit pays européens (Mossong et al. 2008). Au cours de la pandémie de COVID-19, de nouvelles données ont été régulièrement collectées via l'enquête CoMix (Coletti et al. 2020), initiée au Royaume-Uni, aux Pays-Bas et en Belgique, puis étendue à d'autres pays Européens. Cette enquête est financée en Belgique par le projet européen EpiPose (numéro de subvention H2020 101003688), Sciensano et Johnson and Johnson. Cette enquête toujours en cours nous donne des informations presque en temps réel sur l'évolution des contacts en Belgique tous les 15 jours, comme illustré sur la Figure 2. Ces données sont publiquement disponibles, après un temps de traitement, surle site http://www.socialcontactdata.org.

Figure 2 : Visualisation des données de contacts sociaux issues de l'étude CoMix en Belgique par tranches d'âge et par localisation pour la période 23/12/2020 au 18/08/2021. Intervalles de confiances de 95% (bootstrap)



Source : James Wambua, UHasselt

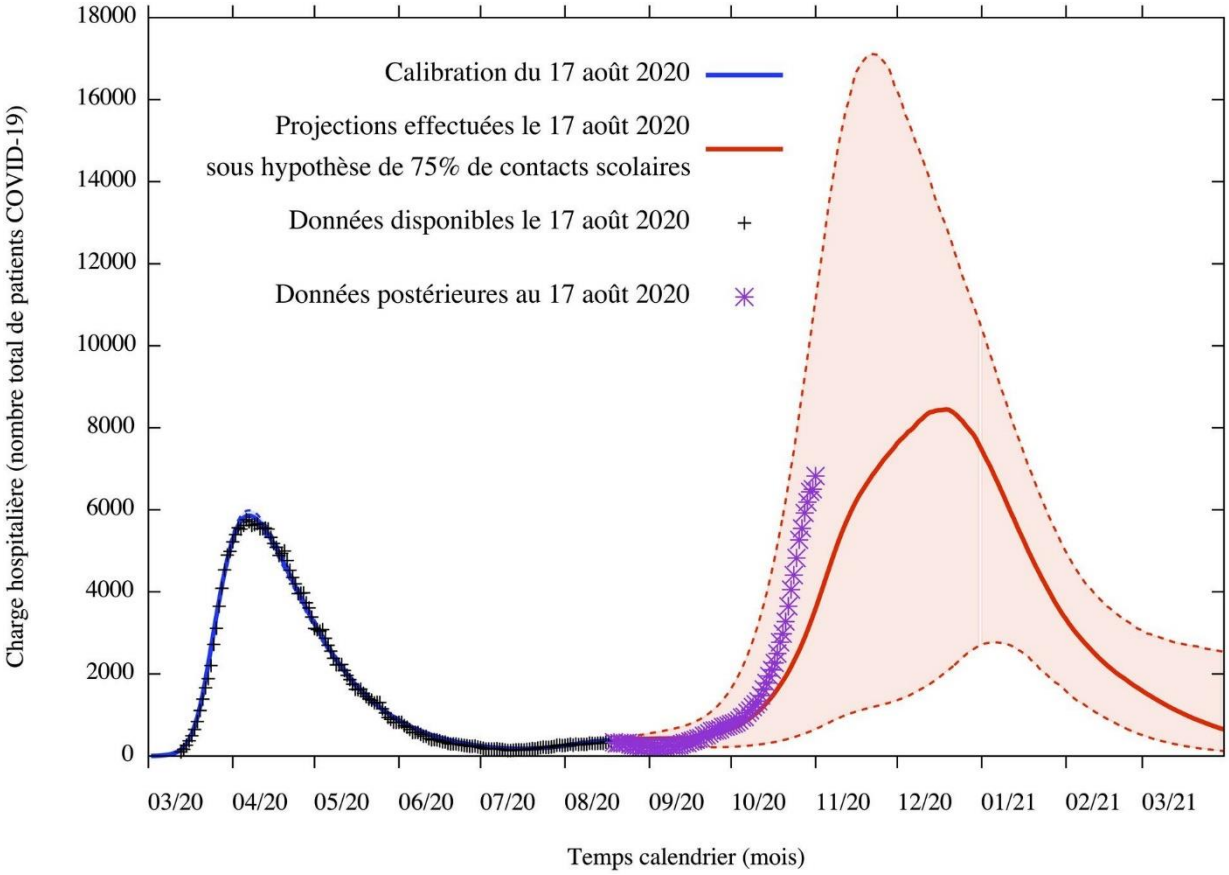
Machine learning, projections et incertitudes

Malgré des données de contacts sociaux disponibles et de certaines estimations pouvant venir de la littérature scientifique, les modèles COVID-19 contiennent encore un très grand nombre de paramètres inconnus, de plusieurs dizaines à plus d'une centaine. La façon de déterminer ces paramètres est d'utiliser des programmes informatiques puissants.

Ces programmes utilisent un processus d'optimisation numérique aléatoire surnommé « boite noire ». L'idée simplifiée de ce processus est la suivante : le programme tire « au hasard » les éléments qu'il ne connait pas, calcule l'évolution théorique à partir de ces éléments et regarde si ce qui est obtenu peut correspondre aux données fiables observées. À chaque fois qu'un résultat est obtenu, les paramètres sont légèrement modifiés et une nouvelle évolution est calculée. Si cette nouvelle évolution correspond davantage aux données réelles, elle est conservée, et le programme poursuit sa recherche afin de se rapprocher progressivement de la réalité cachée qui semble correspondre le mieux aux données de l'épidémie. Le modèle retient alors non pas une solution, mais bien un certain ensemble de solutions suffisamment proches des données observées de façon à tenir compte de l'incertitude sur ces données et sur les paramètres (méthode d'inférence bayésienne). L'estimation de ces paramètres peut être améliorée au fur et à mesure des jours grâce à l'obtention de donnéessupplémentaires, ce qui fait que le modèle s'améliore progressivement d'une façon un peu similaire à un processus de *machine learning*. Ces programmes de grande complexité nécessitent généralement plus de 10000 heures de calcul. Une infrastructure particulière de calcul de haute performance ou « *cluster* » est donc nécessaire permettant une exécution en parallèle sur des centaines de processeurs (comme avec Hercules2de la Plateforme Technologique de calcul Intensif de l'UNamur).

Les modèles permettent par la suite de faire des projections sur base des paramètres estimés mais l'incertitude calculée sur ces paramètres engendre une incertitude sur les projections. Vu qu'il existe différents jeux de paramètres pouvant correspondre à la réalité existante, il existe une multitude de projections possibles et aucune d'entre elles ne peut être considérée comme plus probable que les autres. Pour plus de lisibilité, ces projections sont généralement représentées à l'aide d'un intervalle de confiance et d'une moyenne ou médiane. Cependant, il est erroné de se focaliser sur cette valeur médiane qui ne représente qu'une possibilité parmi les autres. L'ensemble des possibilités représentées doivent être considérées comme potentielles. Ainsi, les différents modèles belges ont projeté l'existence d'une 2^{e} vague en Belgique dès le mois d'août 2020, comme représenté sur la Figure 3. L'incertitude sur les paramètres ne permettait cependant pas encore de trancher entre une vague dépassant ou non les capacités hospitalières et les nouvelles données réelles représentées en couleur sur la figure ont finalement concordé avec les projections les plus pessimistes.

Figure 3 : Projections du 17 août concernant une possible 2e vague à l'automne. La ligne continue représente la médiane et la partie colorée l'intervalle de confiance de 90%

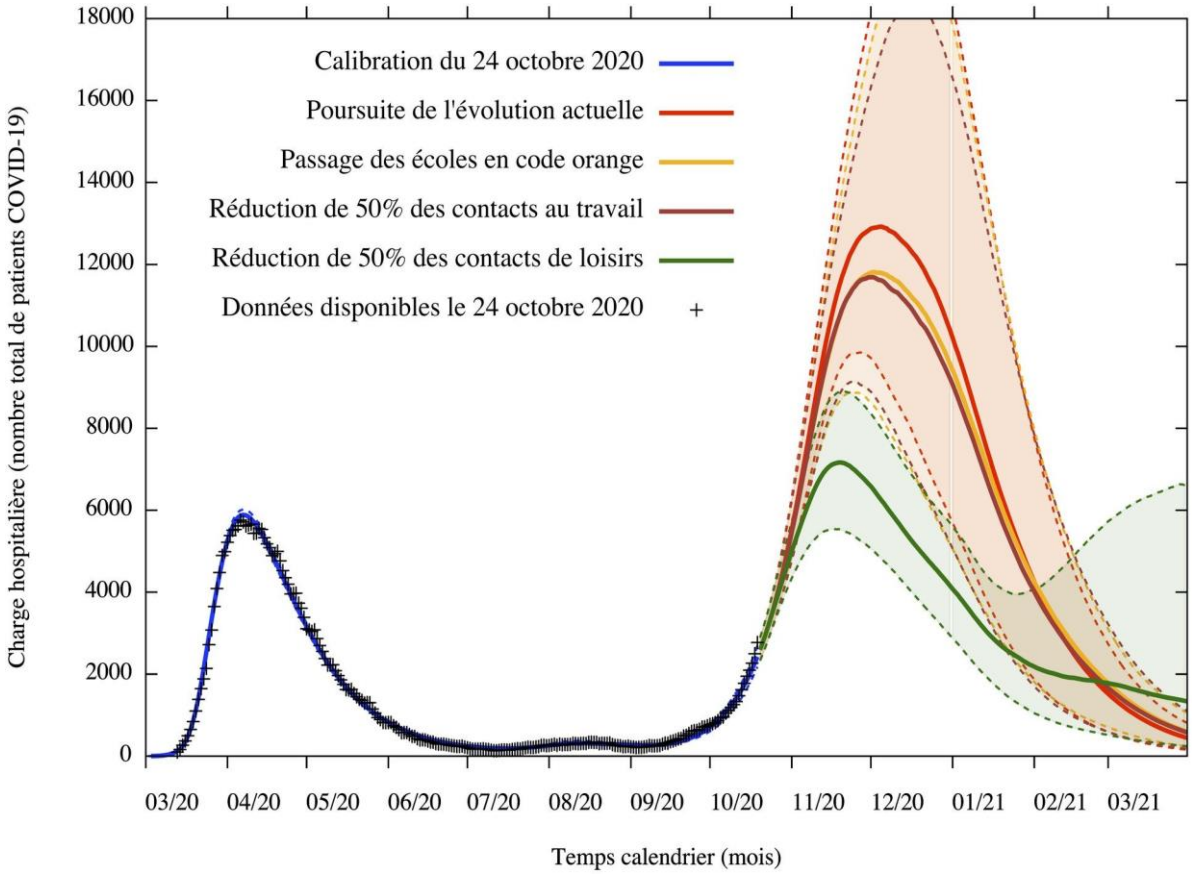


Source : Modèle UNamur (Franco, 2021)

Scénarios et non prévisions

Une idée fausse assez répandue est que ces modèles fournissent des prévisions. Or il n'en est rien. Ces modèles ne peuvent produire que deux sortes de résultats. Les premiers sont des projections théoriques de l'évolution de l'épidémie sur base des estimations actuelles mais ces projections ne sont valables que sous l'hypothèse d'une situation et d'un comportement constant. Ainsi, un changement de mesures politiques, une évolution du comportement de la population ou un effet non prévisible tel que l'apparition d'un variant peuvent modifier ces projections. Celles-ci sont cependant utiles pour avoir une idée de l'évolution théorique potentielle en cas d'absence d'intervention. Les seconds résultats possibles sont des scénarios, où des éléments inconnus sont supposés comme hypothèse. Dans ce cas, l'utilité est de pouvoir évaluer l'impact théorique potentiel de certaines mesures ou changements. Une illustration est présentée à la Figure 4 avec l'impact potentiel de différentes mesures envisagées lors de l'apparition de la 2e vague en Belgique enoctobre 2020, montrant une grande efficacité de la réduction des contacts lors des activités de loisirs.

Figure 4 : Projections du 24 octobre 2020 détaillant les effets séparés potentiels de différentes mesures. Ces mesures sont appliquées jusqu'au 3 janvier 2021



Source : Modèle UNamur (Franco, 2021)

Si ces modèles peuvent avoir un intérêt d'information au niveau des décideurs et sont régulièrement communiqués aux instances politiques, leur usage n'est pas restreint à cette utilisation. Durant les phases cruciales de l'épidémie de COVID-19 en Belgique, les résultats de ces modèles ont notamment été communiqués à différentes institutions telle que l'Agence fédérale des médicaments et des produits de santé afin de déterminer les stocks nécessaires, à l'Agence pour une Vie de Qualité pour estimer l'évolution du *testing* et du *tracing*, à différents réseaux hospitaliers pour estimer l'évolution des besoins en personnel ainsi qu'à divers centres de crises régionaux ou locaux. Les différents modèles étant basés sur des techniques et des hypothèses parfois différentes, ils sont régulièrement comparés à des fins de vérification (un processus quiest habituellement utilisé pour les modèles climatiques) et les projections combinées sont diffusées de façon publique (RESTORE, 2021).

Pour conclure, la modélisation mathématique est un outil particulièrement complexe qui constitue un apport important dans la gestion de la crise de la COVID-19, mais qu'il convient de relativiser : tous les modèles et les projections peuvent sans cesse être remis en cause par l'utilisation d'hypothèses différentes, par des changements de politique, par l'évolution du comportement de la population, par l'évolution de la médecine ou l'évolution naturelle du virus. L'incertitude scientifique doit également être intégralement prise en compte et non les valeurs moyennes ou médianes généralement communiquées dans les médias qui ne représentent aucune évolution réelle. Néanmoins, en l'absence de connaissances absolues, la modélisation mathématique constitue une réelle aide dans la compréhension de l'épidémie et dans la prise de décisions, à condition d'avoir conscience de ses

limitations.

Bibliographie

Abrams, S, J Wambua, E Santermans et al. (2021), "Modelling the early phase of the belgian COVID-19 epidemic using a stochastic compartmental model and studying its implied future trajectories", *Epidemics* 35,100449. http://doi.org/10.1016/j.epidem.2021.100449.

Alleman, TW, J Vergeynst, E Torfs et al. (2020), "A deterministic, age-stratified, extended SEIRD model for assessing the effect of non-pharmaceutical interventions on SARS-CoV-2 spread in Belgium", *MedRxiv*, http://doi.org/10.1101/2020.07.17.20156034.

Coletti, P, J Wambua, A Gimma et al. (2020), "CoMix: comparing mixing patterns in the Belgian populationduring and after lockdown", Scientic Reports 10, 21885. http://doi.org/10.1038/s41598-020-78540-7.

Coletti, P, P Libin, O Petrof et al. (2021), "A data-driven metapopulation model for the belgian COVID-19epidemic: assessing the impact of lockdown and exit strategies", *BMC Infect.* Dis. 21 (1), 503, https://doi.org/10.1186/s12879-021-06092-w

Franco N (2021), "COVID-19 Belgium: Extended SEIR-QD model with nursing homes and long-term scenarios-based forecast", *Epidemics* 37, 100490, http://doi.org/10.1016/j.epidem.2021.100490.

Mossong, J, N Hens, M Jit et al. (2008), "Social Contacts and Mixing Patterns Relevant to the Spread of Infectious Diseases ", *PLoS Med* 5(3): e74. http://doi.org/10.1371/journal.pmed.0050074.

RESTORE consortium (2021), Long-term scenarios for the number of new hospitalizations during the Belgian COVID-19 epidemic, rapports publics disponibles à l'adresse https://covid-en-wetenschap.github.io/restore.html.

Willem, L, S Abrams, PJK Libin et al. (2021), "The impact of contact tracing and household bubbles on deconfinement strategies for COVID-19", *Nature Commun*, 12 (1), 1524, http://doi.org/10.1038/s41467-021-21747-7.