Enjeux sociaux & Big Data

Dr Michelle Kelly-Irving

Dir Equipe EQUITY, UMR1295 Inserm/ Université Toulouse III Paul Sabatier

Dir IFERISS, Université de Toulouse, France

ATELIER 2020-2021 de la PLATEFORME ETHIQUE ET BIOSCIENCES

Volet 3 Jeudi 20 mai 2021

Déclaration d'intérêts

Chercheuse en épidémiologie sociale sur les inégalités sociales de santé

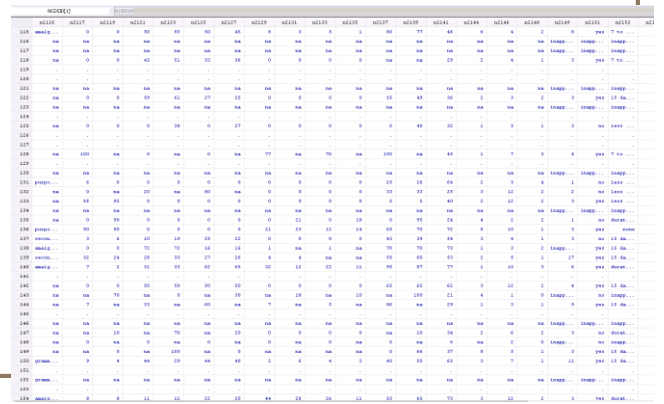
Habituée des données & statistiques mais pas une spécialiste de Big Data, l'intelligence artificielle ou machine learning etc

Motivée par l'équité et la justice sociale

Les grandes bases de données

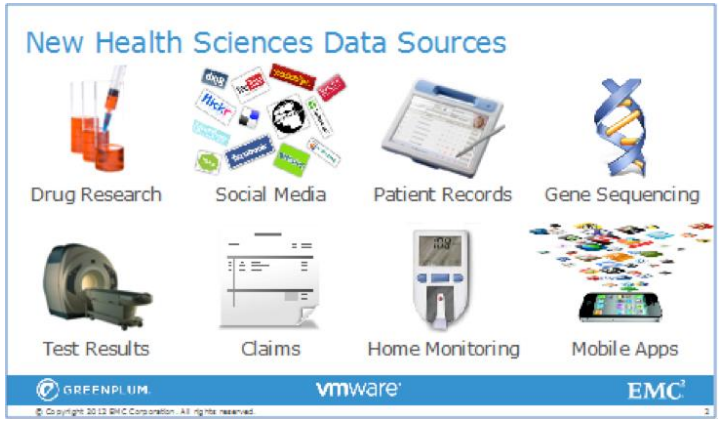
En épidémiologie populationnelle, en démographie, en sociologie quantitative etc les très grosses bases de données existent depuis longtemps…

-Sources d'information diverses & croisements de bases
-Gestion de beaucoup de variables compte tenu des nombres de cas: des milliers de variables par ligne dans des bases avec des dizaines / centaines de milliers de lignes
-Gestion de structures hiérarchiques et causales entre variables



Inserm

Une abondance de données observationnelles



Big Data +
Intelligence Artificielle
=
?

Deux problématiques

1. D'où proviennent ces données qu'on appelle 'Big Data', qui représentent-elles & quelles sont les enjeux sociaux & socio-économiques derrière l'utilisation de ces données via des outils d'intelligence artificielle?

2. Les données financées par les fonds publics en France sont souvent dans des silos, restant non interopérables. Leurs potentiels énormes restants limités par leur séparation. Quel est le devoir de rendre utile ces données?

1. Données: biais implicites & explicites

D'où proviennent ces données qu'on appelle 'Big Data', qui représentent-elles & quelles sont les enjeux socio-économiques derrière l'utilisation de ces données

a) Biais dans les données
b) Biais dans les algorithmes

Le risque de reproduire les inégalités sociales via l'utilisation de certaines données

Les biais dans les données

Les épidémiologistes sont formés sur une utilisation critique des données on apprend à questionner leur fiabilité & validité:

Peuvent elles nous donner une réponse à la question posée, ou nous induire en erreur?

Les biais

Les épidémiologistes sont formés sur une utilisation critique des données on apprend à questionner leur fiabilité & validité:

Peuvent elles nous donner une réponse à la question posé, ou nous induire en erreur?

https://catalogofbias.org/

« Les biais entrent dans les études de santé à tous les stades et influencent souvent l'ampleur et la direction des résultats»

Les biais

Ex: Le biais de sélection

se produit lorsque des individus ou des groupes dans une étude diffèrent systématiquement de la population d'intérêt, ce qui entraîne une erreur systématique dans une association ou un résultat

Most viewed biases

- Selection bias
- Collider bias
- Information bias
- Attrition bias
- Hawthorne effect
- Ascertainment bias
- Recall bias
- Detection bias
- Observer bias

Biais dans la recherche

Malgré les connaissances scientifiques sur les biais, ils se glissent dans les études

Ex: Etudes DOHAD
Biais dans le montage des études et renforcement de ces biais dans les résultats & interprétations

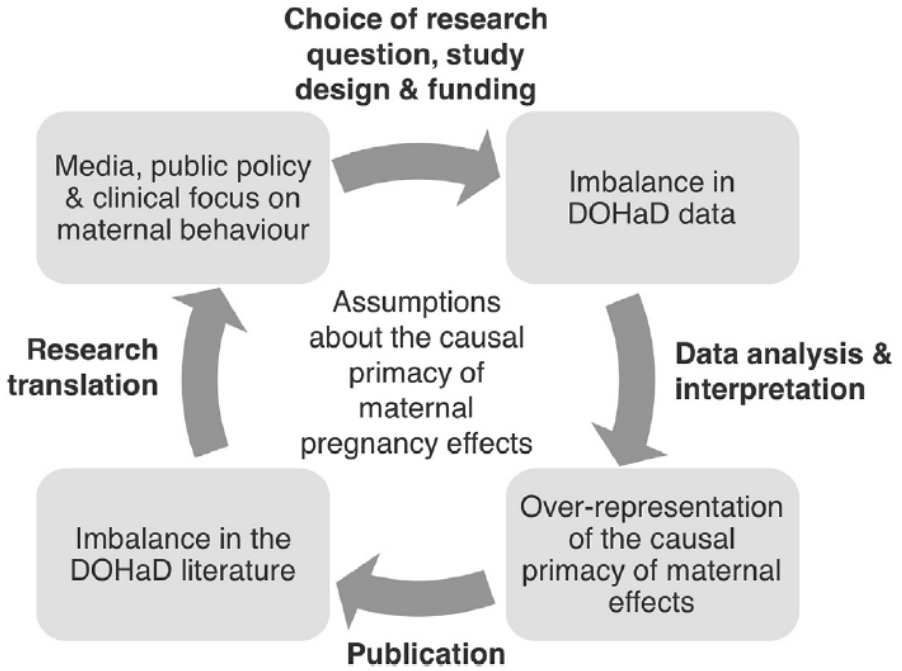


Fig. 2. Assumptions that the health, lifestyle and behaviours of mothers around the time of pregnancy have the largest causal influence on their children's health and risk of disease drives DOHaD research at all stages, from study design to research translation, and is also reinforced by DOHaD research itself.

Sharp, Lawlor & Richardson, SSM 2018

UK A-levels 2020

Biais dans les algorithmes, un exemple

UK A-levels 2020

En Aout 2020 les élèves en Angleterre & aux Pays de Galles ont reçu leurs résultats des A-levels (Baccalauréat)

Mais les examens avaient été annulés durant la pandémie

Au lieu de noter des examens, les notes étaient déterminées par un algorithme:

1. La répartition historique des notes des écoles au cours des trois années précédentes

2. le rang de chaque élève au sein de sa propre école pour une matière particulière, sur la base de l'évaluation par un enseignant de sa note probable si les A-levels avaient eu lieu comme prévu

3. les résultats des examens précédents pour un élève par matière

UK A-levels 2020

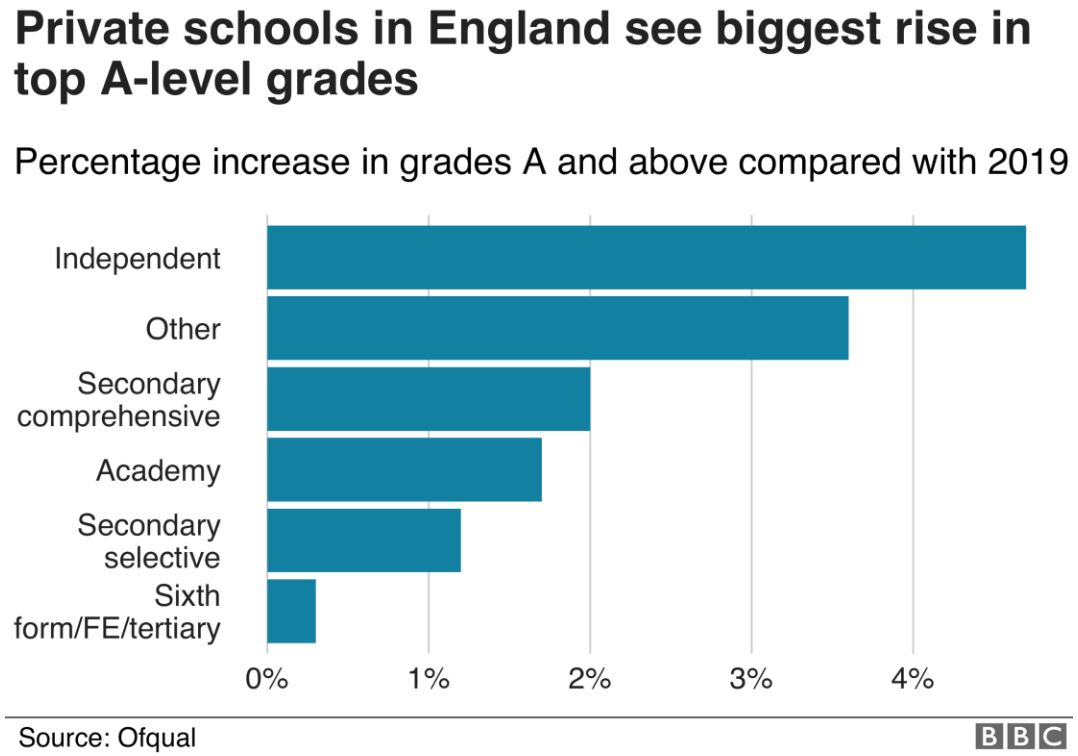
- En Aout 2020 les élèves en Angleterre & aux Pays de Galles ont reçu leurs résultats des A-levels (Baccalauréat)
- Mais les examens avaient été annulés durant la pandémie
- Au lieu de noter les examens réels, les notes étaient déterminées par un algorithme:

1. La répartition historique des notes des écoles au cours des trois années précédentes → Inégalités sociales systémiques
2. le rang de chaque élève au sein de sa propre école pour une matière particulière, sur la base de l'évaluation par un enseignant de sa note probable si les A-levels avaient eu lieu comme prévu → Biais explicites et/ou implicites
3. les résultats des examens précédents pour un élève par matière

UK A-levels 2020

- Près de 40 % des étudiants ont reçu des notes inférieures à celles qui avaient été prévues
- les élèves qui ont eu le privilège de fréquenter une école dont les performances globales étaient élevées depuis longtemps et de manière constante étaient encore plus privilégiés

→Annulation des notes via l'algorithme



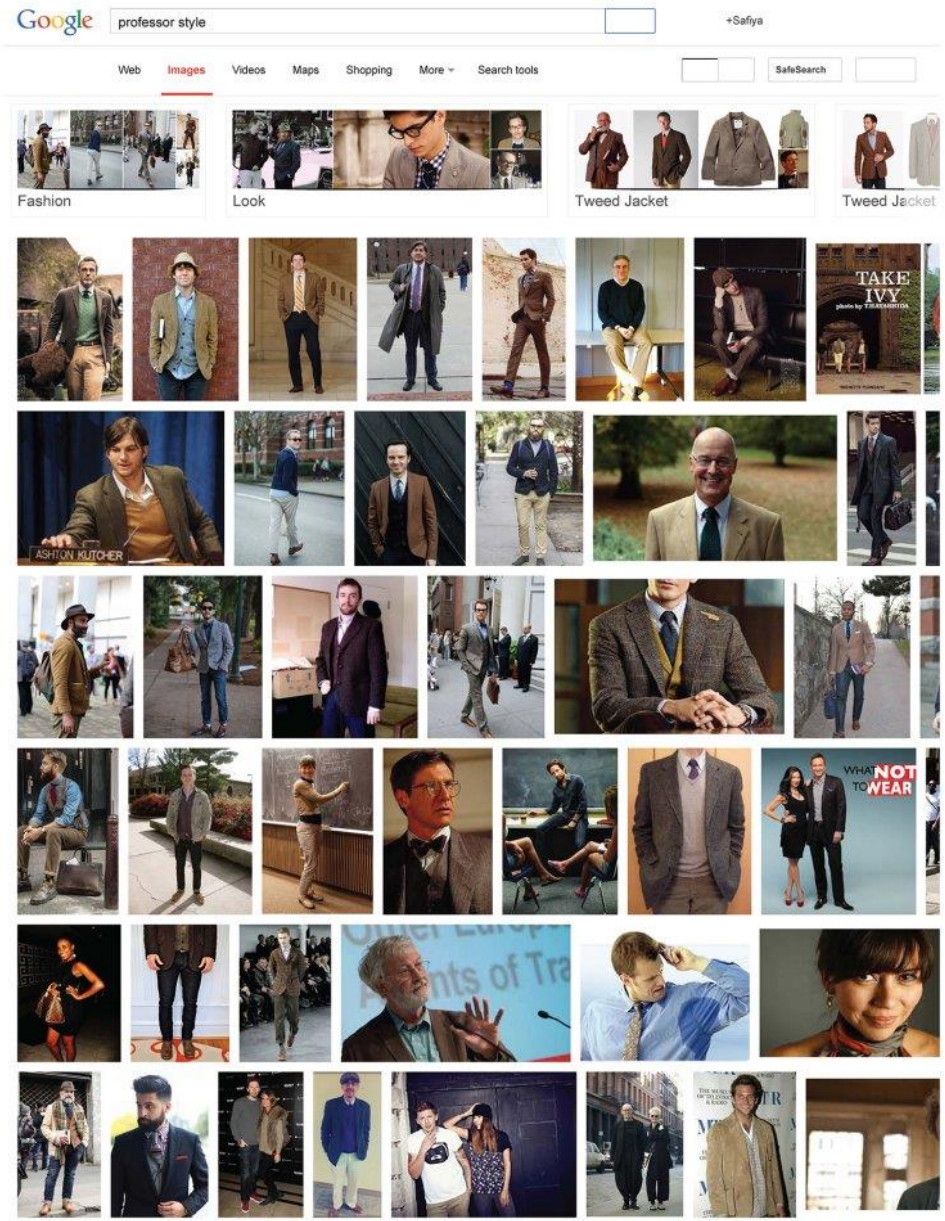
Algorithmes: biais

Biais dans les algorithmes, un exemple :les algorithmes de détection et de classification des visages

Algorithmes: biais

Exemple: requête Google d'images

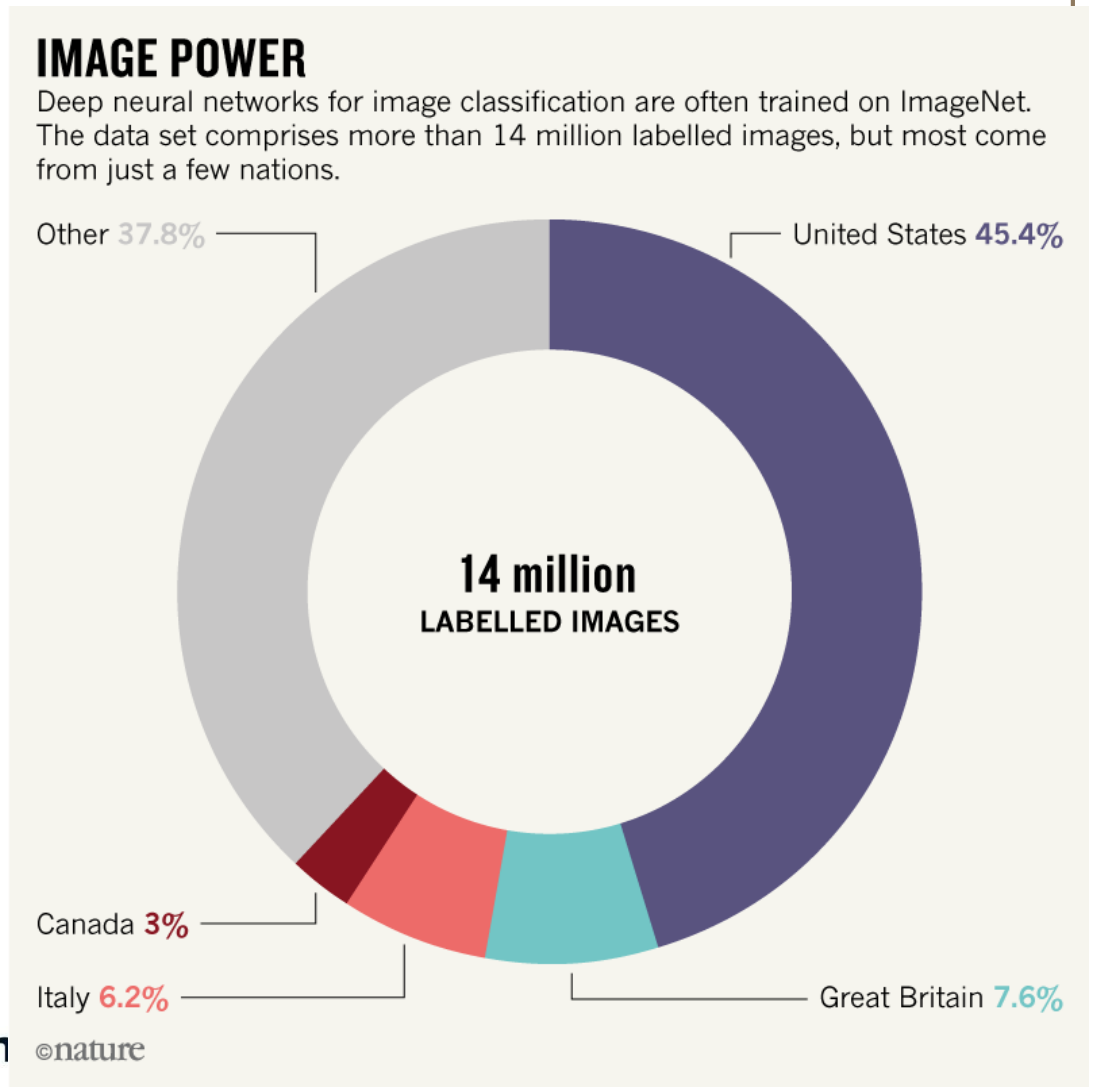
« Professor style »



Cf. Algorithms of opression, par Safiya Umoja Noble

Algorithmes: biais

Les algorithmes de détection et de classification des visages sont utilisés par les forces de l'ordre à des fins de surveillance et de prévention de la criminalité, par des sociétés commerciales et par les gouvernements



Une analyse intersectionnelle des classificateurs

Tous les classificateurs d'images (Microsoft, IMB etc) testés:

- sont plus performants sur les visages masculins que sur les visages féminins
- sont plus performants sur les visages à peaux claires que sur les visages à peaux foncées
- obtiennent les pires résultats sur les visages féminins à peaux foncées

Les classificateurs de Microsoft et d'IBM sont les plus performants sur les visages masculins à peaux claires

Les classificateurs Face++ obtiennent les meilleurs résultats sur les visages masculins à peaux plus foncées

Une analyse intersectionnelle révèle que tous les classificateurs sont moins performants sur les sujets féminins à peaux plus foncées

Buolamwini & Gebru, Proceedings of Machine Learning Research 81:1(15) 2018

Algorithmes: racisme & sexisme

Les algorithmes sont construits par des être humains qui peuvent être sujet aux valeurs et comportements discriminatoires de façon implicite et explicite

Des algorithmes existent pour beaucoup de processus:

-Sélectionner des candidats pour un travail

-Déterminer le salaire

-etc

Google staffer's hostility to affirmative action sparks furious backlash

'Manifesto' arguing against promotion of race and gender diversity attributes lack of women in tech to 'biological causes'

Bonnes pratiques pour limiter les inéquités?

1. S'investir sur la collecte et design de bases de données inclusives & réflexions sur les biais à ce niveau

2. Reconnaitre que les humains sont à la source des biais, pas les algorithmes eux-mêmes, donc soutenir une politique de diversité

3. Promouvoir une responsabilité algorithmique & une transparence dans leur développement pour activement chercher les possibles biais

Information Ethics & Equity Institute
https://ethicsequity.org/

2. Un potentiel sous-exploité

Les données financées & collectées par le secteur public en France sont souvent dans des silos, leur potentiel énorme restant limité par leur cloisonnement

a) Avons nous une obligation de rendre compte des inégalités de façon éthique?
b) Y a-t-il un système règlementaire qui empêche un devoir d'action?

Pourquoi les données Françaises posent problème?

- Multitude de systèmes non interopérables
- Volonté de partage des données parfois limitée : open data
- Absence de coordination globale, volontariste, de long terme, structurante, dépassant les clivages

Oxfam France, 10 avril 2020

SANTÉ Tous égaux face à la pandémie ? La France du Covid-19 en 10 cartes

VINCENT GRIMAULT | 06/04/2020 |

Foyers d'infections, capacités hospitalières, accès à un médecin, résidences secondaires... L'épidémie de coronavirus n'a pas touché tous les Français de la même manière selon le lieu où ils résident.

Alternatives économiques, 6 avril 2020

Coronavirus : « Les inégalités tuent aujourd'hui en Seine-Saint-Denis »

Le Monde, 11 avril 2020

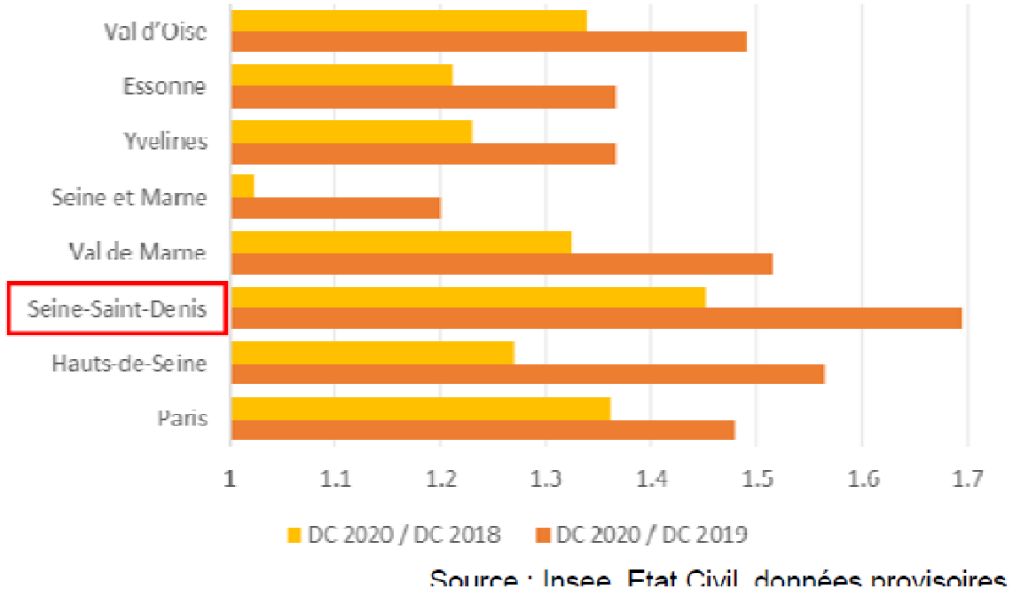
TRIBUNE

Covid-19, miroir des inégalités territoriales et sociales dans le 93

Libération, France, 5 avril 2020

Covid19 et les inégalités territoriales

Figure 3 – Rapport entre le nombre de décès domiciliés entre mars 2020 et mars 2019 et 2018



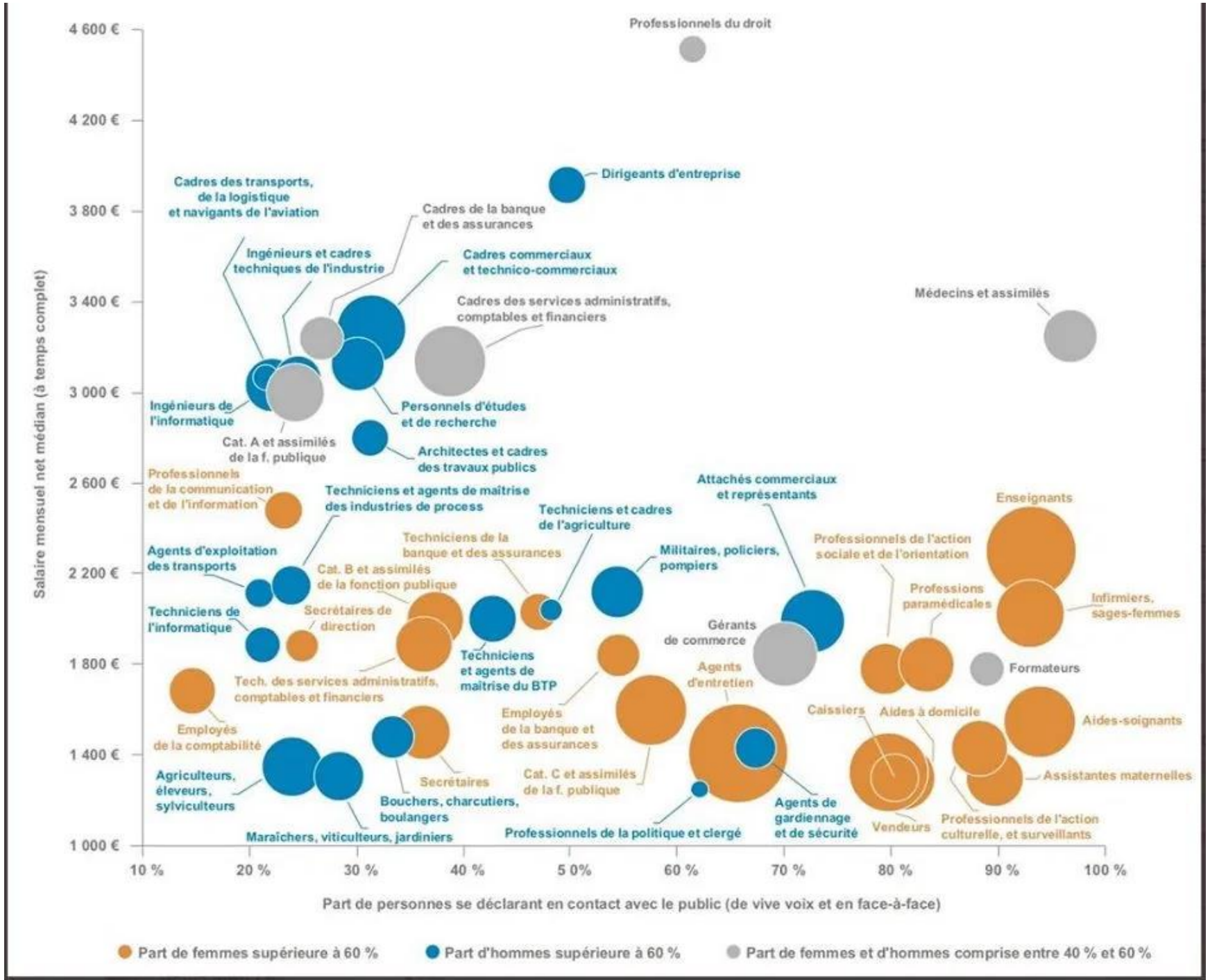
Source : Insee, Etat Civil, données provisoires

En France:

La surmortalité dans le département de Seine-St-Denis a été soulignée en mars-avril 2020

ORS-IDF: FOCUS SANTÉ EN ÎLE-DE-FRANCE | AVRIL 2020

Inserm



Expositions professionnelles & risque de Covid

Sources : France Stratégie, à partir des enquêtes Emploi 2016-2018 (Insee) et de l'enquête Conditions de travail 2013 (Dares)

Le puzzle : une vision incomplète

Les données collectées systématiquement sur
-l'incidence Covid19
-la vaccination
-la morbidité
-la mortalité

Ne sont pas chaînables avec les données sur la profession, les revenus etc

Essay

Importance of collecting data on socioeconomic determinants from the early stage of the COVID-19 outbreak onwards

Saman Khalatbari-Soltani[1, 2], Robert G Cumming[1, 2], Cyrille Delpierre[3, 4], Michelle Kelly-Irving[3, 4, 5]

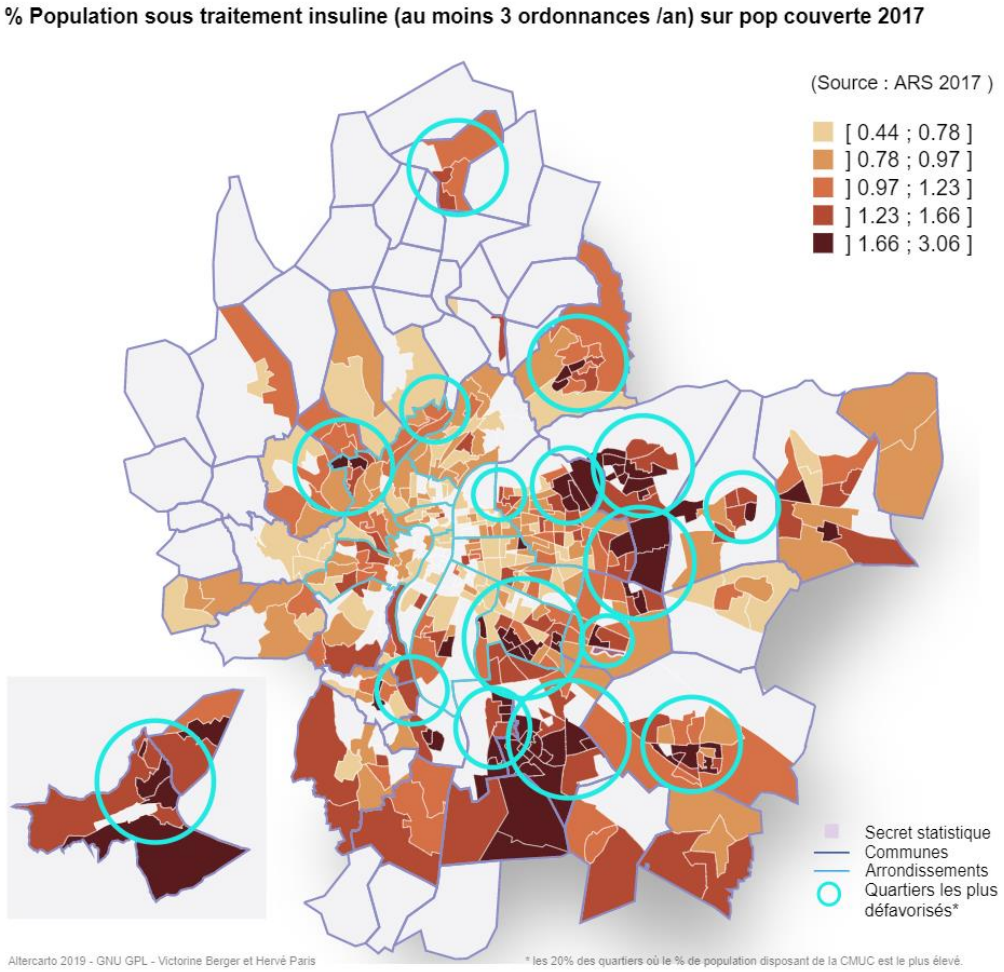
Journal of epi & community health, April 2020

France: éventuellement des données territoriales

Actuellement en France nous ne sommes pas en mesure d'utiliser l'infrastructure des données pour faire un bilan d'épidémiologie sociale rapide et de qualité

On reste à une échelle géographique très large et utilisant des variables incomplètes

On a du monter des études ad hoc
-EpiCov & Sapris



Le potentiel existant dans les données…

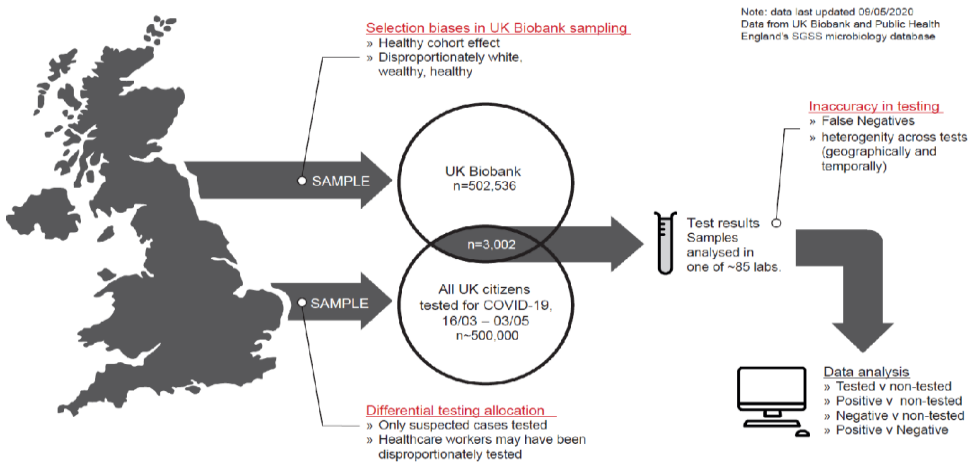
- Le recensement, les données socioéconomiques et administratives
- Les données du système de santé SNDS, données hospitalières, en médecine générale etc

Ensemble, ces sources de données nous permettraient une analyse fine des inégalités sociales en termes d'incidence, prévalence, suivi et morbidité, pouvant avoir un impact sur les politiques et les pratiques

La protection des données est essentielle, et possible…

Gouvernance des données: un potentiel

Les données sur le contexte social, socioéconomique au niveau individuel sont reliées aux données cliniques/de santé de qualité collectées en routine via le NHS



Risk factors for positive and negative COVID-19 tests: a cautious and in-depth analysis of UK biobank data

Marc Chadeau-Hyam ✉, Barbara Bodinier, Joshua Elliott, Matthew D Whitaker, Ioanna Tzoulaki, Roel Vermeulen, Michelle Kelly-Irving, Cyrille Delpierre, Paul Elliott Author Notes

International Journal of Epidemiology, Volume 49, Issue 5, October 2020, Pages

Ethnic differences in SARS-CoV-2 infection and COVID-19-related hospitalisation, intensive care unit admission, and death in 17 million adults in England: an observational cohort study using the OpenSAFELY platform

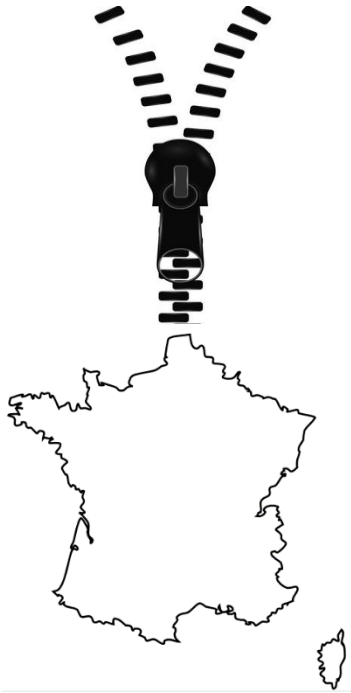
Rohini Mathur, Christopher T Rentsch*, Caroline E Morton*, William J Hulme, Anna Schultze, Brian MacKenna, Rosalind M Eggo, Krishnan Bhaskaran, Angel Y S Wong, Elizabeth J Williamson, Harriet Forbes, Kevin Wing, Helen I McDonald, Chris Bates, Seb Bacon, Alex J Walker, David Evans, Peter Inglesby, Amir Mehrkar, Helen J Curtis, Nicholas J DeVito, Richard Croker, Henry Drysdale, Jonathan Cockburn, John Parry, Frank Hester, Sam Harper, Ian J Douglas, Laurie Tomlinson, Stephen J W Evans, Richard Grieve, David Harrison, Kathy Rowan, Kamlesh Khunti, Nishi Chaturvedi, Liam Smeeth†, Ben Goldacre†, for the OpenSAFELY Collaborative*

Gouvernance des données et inégalités en matière de santé : débloquer les obstacles

Données sociales

- Recensement
- Travail & retraites
- Allocations familiales
- Impots etc…

Merge/croisement
-identifiant unique (NIR)
-Méthodes d'IA



Données de santé

- Collecte en routine
- Hôpital
- Médecine de ville

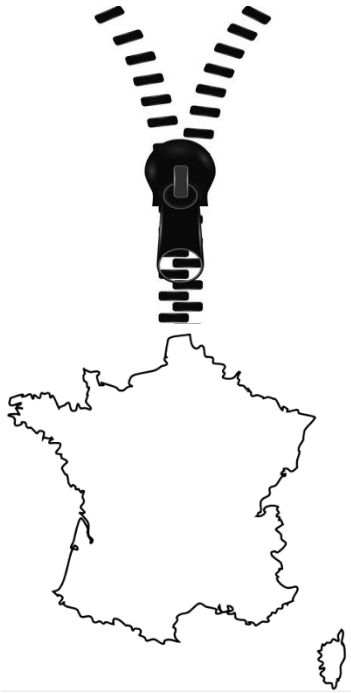
Inserm

Gouvernance des données et inégalités en matière de santé : Covid19

Données sociales

- La CSP
- Le niveau d'études
- Le logement
- Impots etc…

Merge/croisement
-identifiant unique (NIR)
-Méthodes d'IA



Données de santé

- SI-DEP; SI-VAC
- SI-VIC
- Médecine de ville

Inserm

Gouvernance des données éthique: le rêve

Chainage
-Données Sociales
-Données de Santé
-Données sociales collectées dans les dossiers cliniques

Données collectées systématiquement permettant :

-une surveillance
-des réponses aux questions ponctuelles
-informer les politiques publiques
-analyser en urgence

Données collectées de novo permettant :

-des questions approfondies
-tester des hypothèses causales
-informer la recherche et les politiques publiques

Merci

michelle.kelly@inserm.fr

@shell_ki

Ressources & references

https://catalogofbias.org/biases/

http://proceedings.mlr.press/v81/buolamwini18a.html?mod=article_inline

https://www.aclunc.org/blog/californias-equity-algorithm-could-leave-2-million-struggling-californians-without-additional

https://www.bbc.com/news/explainers-53807730

https://www.instituteforgovernment.org.uk/blog/a-level-algorithm-fiasco

https://blogs.lse.ac.uk/impactofsocialsciences/2020/08/26/fk-the-algorithm-what-the-world-can-learn-from-the-uks-a-level-grading-fiasco/

https://time.com/5209144/google-search-engine-algorithm-bias-racism/

https://www.nature.com/articles/d41586-018-05707-8

https://www.opensafely.org/research/2021/ethnic-differences-in-covid19-infection-icu-death/

Archives de données de sciences sociales en Europe

cessda

YHTEISKUNTATIETEELLINEN TIETOARKISTO
FINLANDS SAMHÄLLSVETENSKAPLIGA DATAARKIV
FINNISH SOCIAL SCIENCE DATA ARCHIVE

UK Data Service

Données :

- Quantitatives – source majeure de données individuelles
- Qualitatives
- Productions de projets majeurs
- Gouvernementales & politiques

Data Archiving and Networked Services
DANS

SND Swedish National Data Service

NSD NORWEGIAN CENTRE FOR RESEARCH DATA

QUETELET PROGEDO DIFFUSION

ČSDA Czech Social Science Data Archive
Institute of Sociology