Ciência de Dados Aplicada à COVID-19: Os Dados Implícitos em Meio à Pandemia

Gabriel Di iorio Silva, Victor Ströele, Mário Dantas, Fabrício Mendonça

[1]Departamento de Ciência da Computação - Universidade Federal de Juiz de Fora (UFJF)

{iorio, victor.stroele, mario.dantas, fabricio.mendonca}@ice.ufjf.br

Abstract. *Sars-Cov-2 drastically altered the standard of living of the global population and, with great prominence, the Brazilian. In a country of enormous dimensions like this, its socioeconomic inequalities are also notorious. In this context, each Federative Unit fights the impacts of the disease and reacts to it uniquely. To understand the effects and spread of the COVID-19, the analysis of statistical data based on data science is of great value in the current and future scenario. Initial experimental results indicate that the pandemic and its effects are closely related to the Brazilian states' discrepant realities. Social, economic, and educational indices can help to clarify points related to this issue.*

Resumo. *A Sars-CoV-2 alterou drasticamente o padrão de vida da população global e, com grande destaque, a brasileira. Em um país de grandes dimensões como este, também é notória suas desigualdades socioeconômicas. Nesse contexto, cada Unidade Federativa combate os impactos da doença e reage a ela de maneira singular. Assim, para compreender os impactos e disseminação da COVID-19, a análise de dados estatísticos baseada em Ciência de Dados é de grande valia no cenário atual e futuro. Resultados experimentais iniciais apontam que a pandemia e seus efeitos possuem relação próxima com as realidades discrepantes presentes nos Estados brasileiros. Índices sociais, econômicos e educacionais podem auxiliar na elucidação de pontos relativos a essa questão.*

1. Introdução

No final do ano de 2019 e início de 2020 uma nova doença começou sua disseminação na região de Wuhan, na China. A *Sars-CoV-2* teve seu primeiro artigo científico publicado ainda no mês de dezembro com os pesquisadores chineses descrevendo o caso de um paciente de 41 anos admitido no Hospital Central de Wuhan em 26 de dezembro [Wu et al. 2020]. Apesar da grande quantidade de artigos, pesquisas e estudos acerca da doença, não é possível, ainda, afirmar sua origem. Observa-se a relação de grande compatibilidade com vírus encontrados em morcegos e pangolins [Lam et al. 2020] [Zhou et al. 2020], que são animais comercializados em mercados na cidade de Wuhan. Entretanto, segundo Alexandre Hassanin, pesquisador da Universidade de Sorbonne, é possível que o vírus tenha sua origem de um produto entre um vírus próximo ao do morcego e ao do pangolim [Hassanin 2020]. O elo que explicaria a origem e ligação entre os vírus desses dois animais permanece desconhecido.

Ainda que haja uma grande quantidade de trabalhos que abordam a *Sars-CoV-2*, suas áreas de estudo são amplas e fazem com que seja possível visualizar a problemática dos mais variados ângulos. Dentre essas áreas, destacam-se: estudos biológicos, para

maior entendimento da doença em sua essência [Koyama et al. 2020]; e modelos matemáticos, para predição e estudos das curvas [Tuite et al. 2020] além de alternativas históricas a fim de realizar comparações com pandemias que acometeram a humanidade no passado [San Lau et al. 2020]. Entretanto, formas híbridas que combinem dados estatísticos com tópicos sociais ainda foram pouco exploradas, sendo o diferencial da pesquisa realizada neste artigo.

Uma das principais motivações para abordagem dessa temática é proveniente de estudos de atenção primária básica de saúde [do Nascimento et al. 2020]. No trabalho, desenvolvido pelo mesmo grupo de pesquisadores desse artigo, foi identificada a relevância das técnicas de Ciência de Dados no tratamento do procedimento proposto. Por conseguinte, uma ótica focada nesse aspecto é de suma importância na evolução e compreensão dessa área da pesquisa.

Neste trabalho, apresentamos uma pesquisa baseada em Ciência de Dados com relação a questões centrais na dissipação da doença diante das diversas Unidades Federativas do Brasil. São abordadas as relações de dados com a quantidade de casos, mortalidades e testes por região, a fim de se entender quais pontos são fatores determinantes para a disseminação da doença em um país de dimensões continentais e com enormes diferenças culturais, sociais e econômicas. A pesquisa objetiva esclarecer e contribuir para uma das questões centrais da disseminação da doença no Brasil: *"Por que os Estados brasileiros sofrem das consequências da Sars-Cov-2 de maneira tão singular?*".

O trabalho se encontra dividido da seguinte forma: Na Seção 2 são apresentados os trabalhos relacionados, na Seção 3 são indicado os materiais e métodos de pesquisa, na Seção 4 são retratados os resultados obtidos, na Seção 5 são apresentadas as considerações finais e os trabalhos futuros.

2. Trabalhos Relacionados

O trabalho investigatório promovido pela Ciência de Dados a respeito das informações da *Sars-CoV-2* almeja assimilar os elementos implícitos que os dados estatísticos oriundos da pandemia, como números de casos e mortes, carregam. Logo, trabalhos que permeiam campos da Ciência de Dados operadas para entendimento da nova pandemia e pesquisas que apontam fatores socioeconômicos agravantes para disseminação da doença serviram como material de embasamento nesse artigo.

Conforme constatado em [Patel et al. 2020], no Reino Unido as chances de pessoas com status socioeconômico mais baixo serem acometidas pela doença é maior. Posto que, dentre diversas características citadas, tais pessoas são geralmente empregadas em cargos que não possuem a oportunidade de *home office*. Além disso, suas condições de moradia facilitam seus acometimentos pela doença.

Outrossim, ainda é possível perceber essa ligação direta dentro da população brasileira. No trabalho [do Amaral Schenkel 2020] é evidenciado a medida diretamente proporcional entre renda per capita e gastos com saúde. Além disso, a concentração de regiões de saúde com leitos de Unidade de Tratamento Intensivo (UTI) dentro das conformidades recomendadas pela Organização Mundial da Saúde (OMS) na região sudeste e sul, principalmente, é outro fator averiguado e alarmante.

Além disso, por meio da Ciência de Dados, modelos de predição de dados acerca

da *Sars-CoV-2* já são desenvolvidos para mostrar o impacto da quarentena na redução do acometimento da população pela doença. No estudo feito em [Ray et al. 2020] é discutido o modelo produzido e seus resultados com base na proposta citada.

Outras perspectivas de Ciência de Dados apresentam a criação de um repositório contendo pesquisas e *datasets* que são atualizados frequentemente para criação, desenvolvimento e implantação de estratégias no combate ao novo coronavírus. Repositório, este, envolvendo Ciência de Dados como espécie de plataforma para disseminação de conhecimento sobre o novo coronavírus [Latif et al. 2020].

Portanto, considerando os trabalhos supracitados, o cruzamento de elementos da enfermidade de *Sars-CoV-2* com noções estatísticas disponibilizadas pelo IBGE, nos permite ratificar a conexão entre ambos e iniciar o processo de desmistificação da afirmativa: "a doença não discrimina". Esse trabalho carrega como diferencial a análise de correlação entre indicadores populacionais dos estados brasileiros com os casos confirmados, número de mortes e testes feitos do novo coronavírus.

Em contrapartida com os trabalhos mencionados, este artigo se posiciona a fim de apresentar uma perspectiva estatística e científica para ratificar as correlações do novo coronavírus com os dados socioeconômicos dos Estados brasileiros. Para tal, além de agregar as informações em um único *dataset* é interessante ressaltar as contribuições por meio do embasamento para outras pesquisas a partir dos resultados aqui obtidos.

3. Materiais e Métodos

Nesta seção a proposta deste artigo é apresentada com o objetivo de esclarecer as etapas e atividades relevantes para o estudo da correlação dos indicadores sociais com os impactos do novo coronavírus. O processo de descoberta do conhecimento KDD (*Knowledge Discovery in Databases*) [Han et al. 2011] foi adotado para a extração e consolidação de dados sobre indicadores populacionais e casos da pandemia da *Sars-CoV-2* nos Estados brasileiros. O *workflow* desta pesquisa é mostrado na Figura 1.

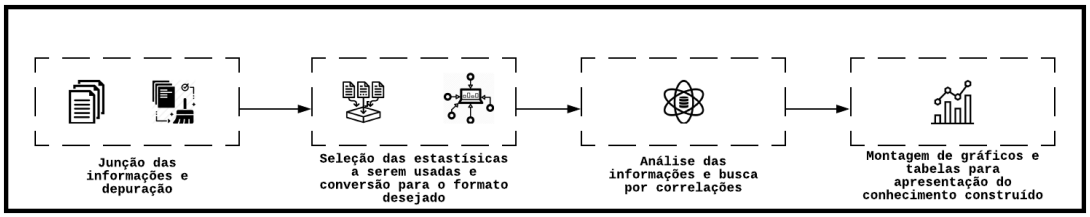

Figura 1. Fluxograma da metodologia KDD aplicada no estudo

Os dados utilizados são provenientes da *Our world in data*[1], publicação digital que expõe pesquisas e dados analíticos sobre mudanças ao redor do mundo. O *dataset* usado inicialmente consistia dos dados atualizados diariamente com averiguações a respeito dos impactos da *Sars-CoV-2* em cada país. Além disso, dados sociais de cada país também eram fornecidos para análise. Alguns dados não relacionados à doença são: população, densidade populacional, média de idade, população acima dos 65 anos, população acima dos 70 anos, renda per capita, população na faixa de extrema pobreza, além de outros.

[1]https://raw.githubusercontent.com/owid/covid-19-data/master/public/data/owid-covid-data.csv

Variáveis	Descrição
UF/Unidade da Federação	Sigla correspondente ao Estado
Total de mortes, testes e casos por 100.000 habitantes por Unidade Federativa	Total de mortes, testes e casos ocasionados pela doença por estado brasileiro a cada grupo de 100.000 habitantes
Índice de Gini	Medida de desigualdade que, neste caso, representa a desigualdade de renda de pessoas de 10 anos de idade ou mais e ocupadas no momento da pesquisa
Renda per capita	Renda média de um individuo daquela população
Índice de Desenvolvimento Humano (IDH)	Medida utilizada para aferir o grau de desenvolvimento de uma determinada sociedade nos quesitos de educação, saúde e renda
Taxa de desocupação	Número de pessoas desocupadas dividido pela população economicamente ativa (PEA)
Acesso à saúde privada	Porcentagem da população que tem acesso a planos de saúde
Trabalho informal	Taxa de informalidade da população ocupada
Pessoas com 25 anos ou mais e com 12 anos ou mais de estudo	Taxa de pessoas de 25 anos ou mais de idade que possuem 12 anos ou mais de escolaridade
Pessoas com ensino fundamental, médio ou superior completo por grupo	Taxa de pessoas de 25 anos ou mais de idade que possuem o ensino fundamental, médio, superior ou equivalentes completo
Domicílios em aglomerados subnormais	Taxa de domicílios que estão presentes em regiões de aglomerados subnormais
Acesso à água canalizada	Porcentagem de domicílios com acesso à água canalizada por estado brasileiro

Tabela 1. Tabela das variáveis empregues e suas descrições

No processo de estudo desse *dataset* foi possível observar que as variáveis associadas à doença, como número de casos, testes e mortes, não tinham correlação significativa com outras informações da mesma base de dados, inviabilizando a equiparação entre elas. Assim, um novo rumo de estudos direcionados a entender a doença e suas ligações em território brasileiro foi abordado. Um país com dimensões territoriais enormes e que experimenta, em suas diversas Unidades Federativas, realidades totalmente distintas também é impactado pela doença de forma bastante singular e, portanto, merece uma diligência própria.

Inicialmente, seguindo os passos do processo KDD, no passo de *Data Cleaning*, foi feito o pré-processamento dos dados coletados das diversas fontes utilizadas neste trabalho para tratar os valores em branco ou faltantes. Depois, foram reunidas informações que poderiam constituir correlações relevantes e interessantes quando comparadas com os dados estatísticos da doença conforme o procedimento de *Data Integration*. A partir disso foram selecionados, dentre os dados obtidos, aqueles mais relevantes para o trabalho além de se descartar as informações agrupadas por países, segundo o método de *Data Selection*, citado previamente. Os dados adotados na construção desse estudo assim como a descrição de cada e as referências de fontes são apresentados na Tabela 1. O *dataset* completo assim como as fontes dos dados usadas nesse trabalho estão disponíveis online [2]. É importante frisar que, como os dados da doença são atualizados diariamente, as correlações podem sofrer pequenas alterações.

Vale salientar que as informações referentes ao percentual da população com ensino fundamental, médio e superior completo acima de 25 anos de idade não é cumulativo. Dessa forma, uma pessoa acima de 25 anos que possui sua escolaridade com ensino superior completo não é contada como pessoa com ensino médio completo, por exemplo.

Com os dados tratados é necessária sua transformação para que assim possam ser estudados de maneira mais eficiente. Seguindo o passo de *Data Transformation* foi realizada a conversão dos mesmos para formato *CSV (Comma-separated values)* visando

[2]https://github.com/diiorio7/DadosCovid

a análise na plataforma *Jupyter*, empregada neste trabalho. Diversas bibliotecas para mineração e manipulação de informações a partir do *dataset* montado foram adotadas. Dentre elas podemos citar: *Pandas, Seaborn, Numpy e Matplotlib*, além de outras. Assim, fomos capazes de efetuar os processos de *Data Mining* e *Pattern Evaluation* com a identificação de correlações significativas entre as variáveis.

Vale frisar que todas as etapas presentes no processo KDD foram desenvolvidas sequencialmente formando uma pirâmide. Fundamentados em dados brutos coletados das mais variadas fontes, com técnicas de Ciência de Dados 'lapidamos' as informações para apresentar outras averiguações. Consequentemente, nos aproximamos do cume da pirâmide que nos mostra conhecimentos que, antes, eram de difícil percepção consolidando a construção da investigação.

4. Resultados

O produto das informações relacionadas à enfermidade e dos dados estatísticos disponibilizados pelo IBGE pode ser descrito como uma correlação entre essas variáveis. A correlação foi feita, nesse estudo, usando métodos da biblioteca *Pandas*, em especial, a função *corr()*.

No cálculo das correlações para variáveis quantitativas dois métodos são amplamente utilizados: *Pearson* e *Spearman*. O primeiro método mede o grau de correlação linear entre duas variáveis, enquanto o segundo não requer a suposição que a relação entre as variáveis é linear. Para melhor percepção das informações implícitas no *dataset* foi calculada a correlação usando os dois métodos para todas as variáveis. Dessa forma, a noção de como a variável se relaciona com os dados do novo coronavírus fica evidente, seja essa relação linear (*Pearson*) ou não (*Spearman*).

Deve-se frisar que a remoção de *outliers* não foi feita, pois, por se tratar de um conjunto de dados de 27 Unidades Federativas, a remoção de informações pode alterar radicalmente as correlações obtidas, além de não espelhar com veracidade os fatos aqui presentes. Para tratar esse problema, optou-se por usar dados relativos, ou seja, no caso dos dados do novo coronavírus, utilizou-se informações relativas à grupos de 100.000 habitantes. Nas demais, dados percentuais à população foram adotados.

O coeficiente de correlação varia de -1 a 1. Um coeficiente igual a 1 indica uma correlação perfeita e positiva, ou seja, ambas as variáveis caminham num caminho decrescente ou crescente de forma linear ou não segundo o método. Por outro lado, um coeficiente de valor igual a -1 nos mostra uma correlação perfeita e negativa, ou seja, ao passo que uma variável cresce ou decresce a outra caminha em um sentido contrário, crescendo ou decrescendo de forma linear ou não, segundo o método adotado.

Logo, com os métodos e atributos para interpretação definidos, os resultados das correlações encontradas são listados na tabela da Figura 2. Vale ressaltar que a técnica de *Spearman* também é capaz de identificar correlações lineares. Entretanto, por se tratar de um método mais abrangente, percebe-se que a correlação de *Spearman* possui um coeficiente menor do que a de *Pearson*.

Com base nos resultados, podemos observar o contraste que existe entre os dois métodos. Em alguns casos as técnicas se complementam, corroborando com a ideia de que existe ou não uma correlação entre as duas variáveis. Para compreender melhor as

Correlações baseadas no método de Pearson | Correlações baseadas no método de Spearman

Dados estatísticos IBGE	Total de casos/100.000 hab	Total de mortes/100.000 hab	Total de testes/100.000 hab	Total de casos/100.000 habitantes	Total de mortes/100.000 habitantes	Total de testes/100.000 habitantes
Índice de Gini	0.424922	0.428078	0.502872	0.390229	0.375267	0.320611
Renda per capita	-0.126290	-0.248962	0.312014	-0.320513	-0.162393	0.017094
Índice de Desenvolvimento Humano	-0.046202	-0.224247	0.259977	-0.224393	-0.096779	0.079377
Taxa de desocupação	0.433225	0.469997	0.258164	0.295403	0.433787	0.179319
Ensino superior completo	0.283868	0.035781	0.589772	-0.033288	0.077569	0.271797
Ensino médio completo	0.411807	0.403667	0.303539	0.073305	0.361332	0.104154
Ensino fundamental completo	-0.453855	-0.298687	-0.394238	-0.523453	-0.274103	-0.323300
Acesso a planos de saúde privados	-0.383763	-0.188234	-0.010531	-0.523199	-0.131868	-0.173993
Taxa de trabalho informal	0.243074	0.413348	-0.064242	0.341880	0.288156	0.040904
Acesso à água canalizada	-0.143204	-0.423083	0.038143	-0.261363	-0.339404	0.034277
Pessoas com 12 anos ou mais de escolaridade	0.400025	0.237054	0.516717	0.019844	0.215845	0.199969
Domicílios em aglomerados subnormais	0.217614	0.535869	0.077961	0.219780	0.569597	0.133700

Figura 2. Análise de correlações

informações dispostas vale frisar que: Correlações fortes entre as variáveis em ambas estratégias apontam uma relação linear uma vez que o método de *Spearman* compreende também o espectro da relações lineares. Correlações muito discrepantes entre as técnicas com uma sendo muito alta e a outra baixa nos mostram uma relação forte linear ou não linear (*Pearson* ou *Spearman*, respectivamente) dependendo de qual técnica a força da correlação foi evidenciada.

Para melhor ilustração dos resultados obtidos com o método de *Pearson*, além da tabela já apresentada e munido das informações supracitadas, é possível a exibição dos mesmos resultados mediante a gráficos na Figura 3. Nessa forma de visualização os gráficos descrevem uma reta crescente ou decrescente segundo sua correlação positiva ou negativa. Os pontos nele disponíveis remetem ao quão forte é a correlação entre as variáveis. Gráficos com sua maioria de pontos distribuídos próximos à reta apresentam uma correlação forte, seja positiva ou negativa.

No gráfico da esquerda é retratada a correlação entre as mortes por 100.000 habitantes e a porcentagem de domicílios presentes em aglomerados subnormais por Unidade Federativa. Nele é possível perceber que, possivelmente, devido à precariedade de muitos desses lares, seus moradores possuem uma maior vulnerabilidade à casos fatais da doença. Simultaneamente, a provável dificuldade de acesso à redes de saúde públicas ocasiona o diagrama exibido.

No gráfico central percebe-se a correlação, também forte, entre a concentração de renda, sintetizada pelo índice de gini, e o total de casos por 100.000 habitantes por Estado brasileiro. Tal índice pode significar, por uma visão macroscópica, um maior percentual da população em situações de pobreza e sem acesso à recursos. Logo, mais suscetíveis aos impactos da *Sars-Cov-2*.

Diferentemente do esperado, o gráfico da direta apresenta uma correlação fraca entre o percentual da população com acesso a planos de saúde privados e o total de testes feitos. Conforme verificado por matéria de órgão oficial, a Agência Nacional de Saúde (ANS) apenas incluiu o teste sorológico para o novo coronavírus no rol de coberturas obrigatórias recentemente [ans]. Diante disso, esse pode ser um dos motivos da baixa

correlação uma vez que muitos testes que poderiam ser feitos com cobertura num passado próximo não foram feitos.

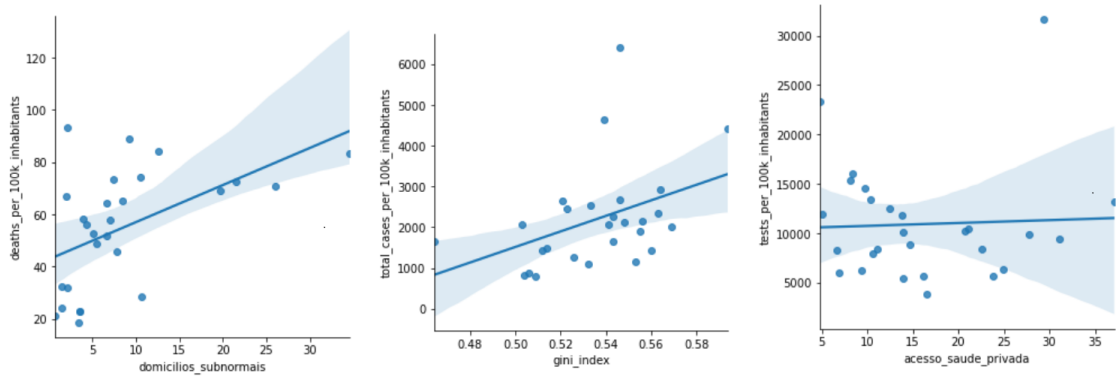

Figura 3. Gráfico da correlação do Índice de Gini

5. Considerações Finais e Trabalhos Futuros

Neste trabalho de pesquisa, baseado em Ciências de Dados, foi possível verificar a correlação dos impactos da pandemia com as desigualdades socioeconômicas presentes no Brasil assim como conhecimentos implícitos presentes. O trabalho auxilia em uma visualização diferenciada, além das estatísticas, posto à Ciência de Dados como forma de prover uma base para melhor entendimento dos dados. Um exemplo indicativo é o nível de acesso à água canalizada que, além de representar uma medida de saneamento, também aponta população em condições precárias e, possivelmente, de extrema pobreza.

Além disso, devemos sempre considerar a afirmativa: Correlação não necessariamente significa causalidade. Entretanto, com os diversos indicadores socioeconômicos abordados nesse trabalho, pesquisas relacionadas e a comparação dos dois métodos nesse artigo nos aponta indícios fortes de que, nesse caso, a correlação indica certa causalidade.

Como trabalhos futuros, é necessário entender com maior detalhe cada um dos dados correlacionados e quais outras estatísticas seriam necessárias para construção de cenários em cada uma das informações disponibilizadas pelo IBGE. Em outras palavras, imagina-se uma análise mais detalhada e com maior precisão executando-se essa proposta de Ciência de Dados em ambientes como [SIN] ou [Gri]. Novos experimentos do nosso grupo, que já executou em outros esforços de pesquisa, poderão nos auxiliar de maneira diferencial em casos acerca de um dos principais pilares para o acometimento tão grave do novo coronavírus perante a população brasileira.

Referências

[ans] Ans inclui teste sorológico para covid-19 no rol de coberturas obrigatórias. http://www.ans.gov.br/aans/noticias-ans/coronavirus-covid-19/coronavirus-todas-as-noticias/5648-ans-inclui-teste-sorologico-para-covid-19-no-rol-de-coberturas-c Accessed: 2020-08-10.

[Gri] Grid5000, 2020. https://www.grid5000.fr/w/Grid5000:Home. Accessed: 2020-08-10.

[SIN] Sistema de computação petaflópica do sinapad, 2020. https://sdumont.lncc.br/. Accessed: 2020-08-10.

[do Amaral Schenkel 2020] do Amaral Schenkel, M. (2020). Por que as desigualdades no brasil são ainda mais visíveis no cenário de enfrentamento ao covid-19? *Instituto de Filosofia e Ciências Humanas (IFCH) — UFRGS*.

[do Nascimento et al. 2020] do Nascimento, M. G., Iorio, G., Thomé, T. G., Medeiros, A. A., Mendonça, F. M., Campos, F. A., David, J. M., Ströele, V., and Dantas, M. A. (2020). Covid-19: A digital transformation approach to a public primary healthcare environment. Technical report, EasyChair.

[Han et al. 2011] Han, J., Kamber, M., and Pei, J. (2011). Data mining concepts and techniques third edition. *The Morgan Kaufmann Series in Data Management Systems*, pages 83–124.

[Hassanin 2020] Hassanin, A. (2020). Coronavirus origins: genome analysis suggests two viruses may have combined. In *World Economic Forum*.

[Koyama et al. 2020] Koyama, T., Platt, D., and Parida, L. (2020). Variant analysis of sars-cov-2 genomes. *Bulletin of the World Health Organization*, 98(7):495.

[Lam et al. 2020] Lam, T. T.-Y., Jia, N., Zhang, Y.-W., Shum, M. H.-H., Jiang, J.-F., Zhu, H.-C., Tong, Y.-G., Shi, Y.-X., Ni, X.-B., Liao, Y.-S., et al. (2020). Identifying sars-cov-2-related coronaviruses in malayan pangolins. *Nature*, pages 1–4.

[Latif et al. 2020] Latif, S., Usman, M., Manzoor, S., Iqbal, W., Qadir, J., Tyson, G., Castro, I., Razi, A., Boulos, M. N. K., Weller, A., et al. (2020). Leveraging data science to combat covid-19: A comprehensive review.

[Patel et al. 2020] Patel, J., Nielsen, F., Badiani, A., Assi, S., Unadkat, V., Patel, B., Ravindrane, R., and Wardle, H. (2020). Poverty, inequality and covid-19: the forgotten vulnerable. *Public health*, 183:110.

[Ray et al. 2020] Ray, D., Salvatore, M., Bhattacharyya, R., Wang, L., Du, J., Mohammed, S., Purkayastha, S., Halder, A., Rix, A., Barker, D., et al. (2020). Predictions, role of interventions and effects of a historic national lockdown in india's response to the covid-19 pandemic: data science call to arms. *Harvard data science review*, 2020(Suppl 1).

[San Lau et al. 2020] San Lau, L., Samari, G., Moresky, R. T., Casey, S. E., Kachur, S. P., Roberts, L. F., and Zard, M. (2020). Covid-19 in humanitarian settings and lessons learned from past epidemics. *Nature Medicine*, 26(5):647–648.

[Tuite et al. 2020] Tuite, A. R., Fisman, D. N., and Greer, A. L. (2020). Mathematical modelling of covid-19 transmission and mitigation strategies in the population of ontario, canada. *CMAJ*, 192(19):E497–E505.

[Wu et al. 2020] Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G., Hu, Y., Tao, Z.-W., Tian, J.-H., Pei, Y.-Y., et al. (2020). A new coronavirus associated with human respiratory disease in china. *Nature*, 579(7798):265–269.

[Zhou et al. 2020] Zhou, P., Yang, X.-L., Wang, X.-G., Hu, B., Zhang, L., Zhang, W., Si, H.-R., Zhu, Y., Li, B., Huang, C.-L., et al. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *nature*, 579(7798):270–273.