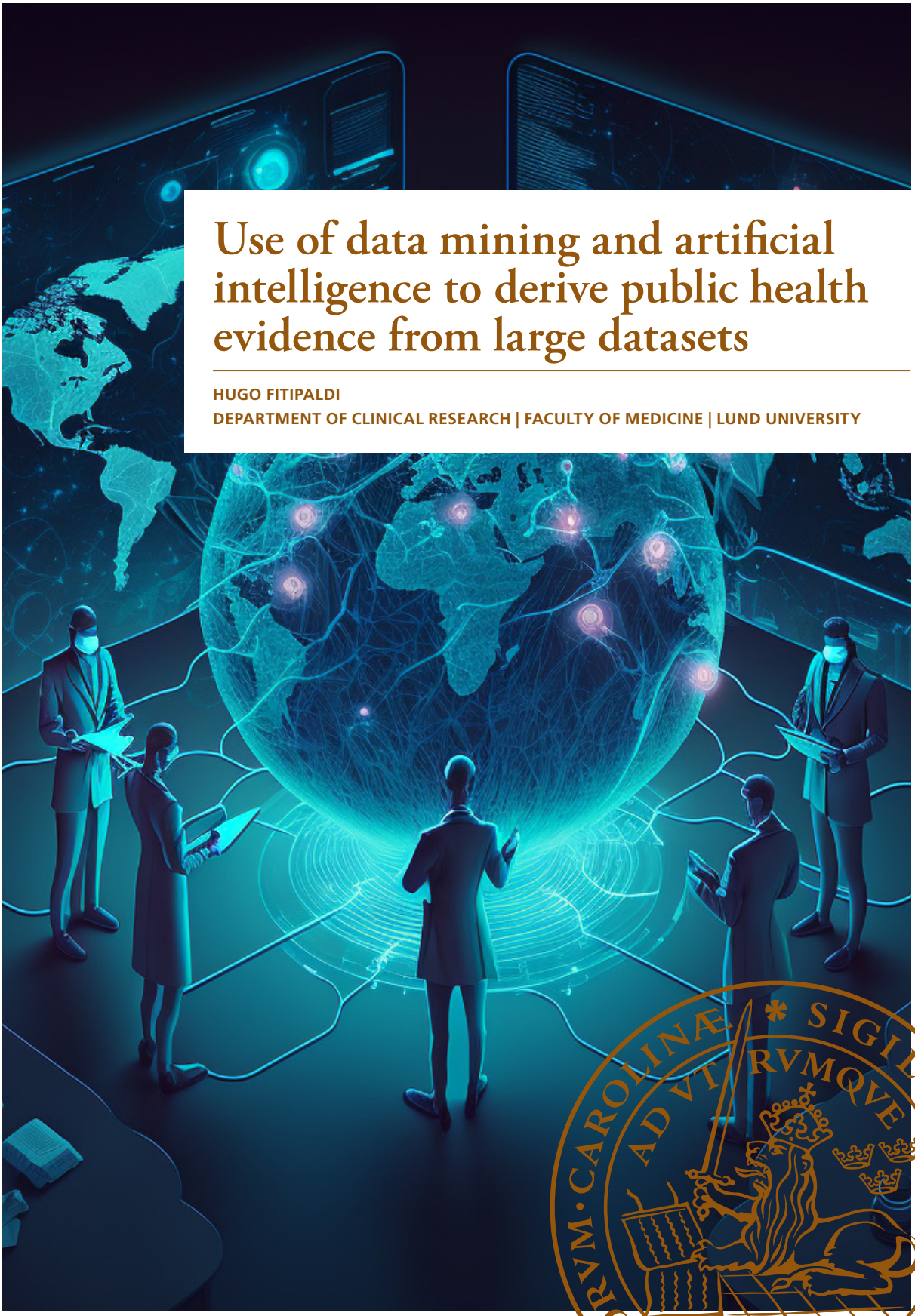
Use of data mining and artificial intelligence to derive public health evidence from large datasets

Fitipaldi, Hugo

2023

Document Version:
Publisher's PDF, also known as Version of record



Use of data mining and artificial intelligence to derive public health evidence from large datasets

HUGO FITIPALDI
DEPARTMENT OF CLINICAL RESEARCH | FACULTY OF MEDICINE | LUND UNIVERSITY

Use of data mining and artificial intelligence to derive public health evidence from large datasets

Use of data mining and artificial intelligence to derive public health evidence from large datasets

Hugo Fitipaldi

LUND
UNIVERSITY

DOCTORAL DISSERTATION

Doctoral dissertation for the degree of Doctor of Philosophy (PhD) at the Faculty of Medicine at Lund University to be publicly defended on the 2nd of March, 2023 at 13.00 in Agardhsalen, Clinical Research Centre, Jan Waldenströms gata 35, Malmö.

Faculty opponent
Prof. Claudia Langenberg
Thesis advisors
Maria F. Gomez, Paul W. Franks

Organization: LUND UNIVERSITY

Document name: DOCTORAL DISSERTATION **Date of disputation:** 02nd March 2023

Author(s): Hugo Fitipaldi **Sponsoring organization:**

Title and subtitle: Use of data mining and artificial intelligence to derive public health evidence from large datasets

Abstract:

This thesis explores the use of data mining and AI-tailored frameworks for extracting public health evidence from large health datasets. The research presented in this thesis demonstrates the potential of these tools for automating and simplifying the data mining process, and for providing valuable insights into various public health issues.

In Paper I, we used data mining and natural language processing to analyze the characteristics of genomic research on non-communicable diseases (NCDs) from the GWAS Catalog (2005 to 2022). We found that the majority of research institutions leading the work are often US-based and the majority of first, senior and all authors were male. The vast majority of complex trait GWAS has been performed in European ancestry populations, with cohorts and scientists predominantly located in medium-to-high socioeconomically ranked countries. This lack of diversity in both the data and the authorship of GWAS research has potential implications for the generalizability of genetic discoveries and the development of future interventions.

In Paper II, we analyzed data collected through the app-based COVID Symptom Study in Sweden. We then created a symptom-based model to estimate the individual probability of symptomatic COVID-19 and employed this to estimate daily regional COVID-19 prevalence. We also used this data to predict next week COVID-19 hospital admissions and compared it to a model based on case notifications. We found that the symptom-based model had a lower median absolute percentage error during the first wave of the pandemic and that the model was transferable to an English dataset. The findings of this study demonstrate the feasibility of large-scale syndromic surveillance and the potential for population-based participatory surveillance initiatives in future pandemics and epidemics.

In Paper III, we used data from over 500,000 participants in the COVID Symptom Study to investigate the impact of obesity and diabetes on the symptoms and duration of long-COVID. Using advanced data mining techniques, we found that individuals with higher BMI and diabetes had a higher burden of symptoms during the initial COVID-19 infection and a prolonged duration of long-COVID symptoms. We also found that vaccination had a protective effect against both COVID-19 symptoms and long-COVID symptoms in these at-risk groups. Our results demonstrate the disproportionate impact of COVID-19 on certain populations and the utility of app-based syndromic surveillance in providing timely and accurate information on the spread and impact of the virus.

Key words: artificial Intelligence, data mining, genome-wide association studies, covid-19

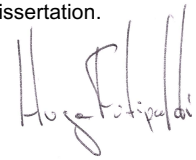
Classification system and/or index terms (if any) Supplementary bibliographical information

Language: English **ISSN and key title**: 1652-8220 Lund University, Faculty of Medicine Doctoral Dissertation Series 2023:24

ISBN: 978-91-8021-363-9

Recipient's notes **Number of pages**: 105

Price Security classification

Signature  Date 2023-01-18

Use of data mining and artificial intelligence to derive public health evidence from large datasets

Hugo Fitipaldi

LUND
UNIVERSITY

Faculty of Medicine
Department of Clinical Research

ISBN 978-91-8021-363-9
ISSN 1652-8220

Printed in Sweden by Media-Tryck, Lund University
Lund 2022

To Francisco, Ilze, Daniel, and Camila

Table of Contents

List of publications

Fitipaldi H, Franks PW. *Ethnic, gender and other sociodemographic biases in genome-wide association studies for the most burdensome non-communicable diseases: 2005-2022*. Hum Mol Genet. 2022 Oct 3:ddac245. doi: 10.1093/hmg/ddac245. Epub ahead of print. PMID: 36190496.

Kennedy B, **Fitipaldi H**, Hammar U, Maziarz M, Tsereteli N, Oskolkov N, Varotsis G, Franks CA, Nguyen D, Spiliopoulos L, Adami HO, Björk J, Engblom S, Fall K, Grimby-Ekman A, Litton JE, Martinell M, Oudin A, Sjöström T, Timpka T, Sudre CH, Graham MS, du Cadet JL, Chan AT, Davies R, Ganesh S, May A, Ourselin S, Pujol JC, Selvachandran S, Wolf J, Spector TD, Steves CJ, Gomez MF, Franks PW, Fall T. *App-based COVID-19 syndromic surveillance and prediction of hospital admissions in COVID Symptom Study Sweden*. Nat Commun. 2022 Apr 21;13(1):2110. doi: 10.1038/s41467-022-29608-7. PMID: 35449172; PMCID: PMC9023535.

Fitipaldi H, Oskolkov N, Kennedy B, Hammar U, Sudre CH, Canas LDS, Pujol JC, Maziarz M, Selberg CA, Tsereteli N, Varotsis G., Chan AT, Ourselin S, Wolf J, Spector TD, Steves CJ, Merino J, Fall T, Franks PW, Gomez MF. *Differential impact of COVID-19 vaccination on symptom presentation during SARS-CoV-2 infection and long-COVID depending on BMI and diabetes.* Manuscript in preparation, 2023.

Publications not included in this thesis

Allesøe RL, Lundgaard AT, Hernández Medina R, Aguayo-Orozco A, Johansen J, Nissen JN, Brorsson C, Mazzoni G, Niu L, Biel JH, Brasas V, Webel H, Benros ME, Pedersen AG, Chmura PJ, Jacobsen UP, Mari A, Koivula R, Mahajan A, Vinuela A, Tajes JF, Sharma S, Haid M, Hong MG, Musholt PB, De Masi F, Vogt J, Pedersen HK, Gudmundsdottir V, Jones A, Kennedy G, Bell J, Thomas EL, Frost G, Thomsen H, Hansen E, Hansen TH, Vestergaard H, Muilwijk M, Blom MT, 't Hart LM, Pattou F, Raverdy V, Brage S, Kokkola T, Heggie A, McEvoy D, Mourby M, Kaye J, Hattersley A, McDonald T, Ridderstråle M, Walker M, Forgie I, Giordano GN, Pavo I, Ruetten H, Pedersen O, Hansen T, Dermitzakis E, Franks PW, Schwenk JM, Adamski J, McCarthy MI, Pearson E, …**Fitipaldi H**.…, Banasik K, Rasmussen S, Brunak S;. *Discovery of drug-omics associations in type 2 diabetes with generative deep-learning models*. Nat Biotechnol. 2023 Jan 2. doi: 10.1038/s41587-022-01520-x. Epub ahead of print. PMID: 36593394.

Mutie PM, Pomares-Milan H, Atabaki-Pasdar N, Coral D, **Fitipaldi H**, Tsereteli N, Tajes JF, Franks PW, Giordano GN. *Investigating the causal relationships between excess adiposity and cardiometabolic health in men and women*. Diabetologia. 2023 Feb;66(2):321-335. doi: 10.1007/s00125-022-05811-5. Epub 2022 Oct 12. PMID: 36221008; PMCID: PMC9807546.

Slieker RC, Donnelly LA, **Fitipaldi H**, Bouland GA, Giordano GN, Åkerlund M, Gerl MJ, Ahlqvist E, Ali A, Dragan I, Elders P, Festa A, Hansen MK, van der Heijden AA, Mansour Aly D, Kim M, Kuznetsov D, Mehl F, Klose C, Simons K, Pavo I, Pullen TJ, Suvitaival T, Wretlind A, Rossing P, Lyssenko V, Legido Quigley C, Groop L, Thorens B, Franks PW, Ibberson M, Rutter GA, Beulens JWJ, 't Hart LM, Pearson ER. *Distinct Molecular Signatures of Clinical Clusters in People With Type 2 Diabetes: An IMI-RHAPSODY Study*. Diabetes. 2021 Nov;70(11):2683-2693. doi: 10.2337/db20-1281. Epub 2021 Aug 10. PMID: 34376475; PMCID: PMC8564413.

Slieker RC, Donnelly LA, **Fitipaldi H**, Bouland GA, Giordano GN, Åkerlund M, Gerl MJ, Ahlqvist E, Ali A, Dragan I, Festa A, Hansen MK, Mansour Aly D, Kim

M, Kuznetsov D, Mehl F, Klose C, Simons K, Pavo I, Pullen TJ, Suvitaival T, Wretlind A, Rossing P, Lyssenko V, Legido-Quigley C, Groop L, Thorens B, Franks PW, Ibberson M, Rutter GA, Beulens JWJ, 't Hart LM, Pearson ER. *Replication and cross-validation of type 2 diabetes subtypes based on clinical variables: an IMI-RHAPSODY study*. Diabetologia. 2021 Sep;64(9):1982-1989. doi: 10.1007/s00125-021-05490-8. Epub 2021 Jun 10. PMID: 34110439; PMCID: PMC8382625.

David A. Drew, Chuan-Guo Guo, Karla A. Lee, Long H. Nguyen, Amit D. Joshi, Chun-Han Lo, Wenjie Ma, Raaj S. Mehta, Sohee Kwon, Christina M. Astley, Mingyang Song, Richard Davies, Joan Capdevila, Mary Ni Lochlainn, Carole H. Sudre, Mark S. Graham, Thomas Varsavsky, Maria F. Gomez, Beatrice Kennedy, **Hugo Fitipaldi**, Jonathan Wolf, Tim D. Spector, Sebastien Ourselin, Claire J. Steves, Andrew T. Chan. *Aspirin and NSAID use and the risk of COVID-19*. medRxiv 2021.04.28.21256261; doi: https://doi.org/10.1101/2021.04.28.21256261

Atabaki-Pasdar N, Ohlsson M, Viñuela A, Frau F, Pomares-Millan H, Haid M, Jones AG, Thomas EL, Koivula RW, Kurbasic A, Mutie PM, **Fitipaldi H**, Fernandez J, Dawed AY, Giordano GN, Forgie IM, McDonald TJ, Rutters F, Cederberg H, Chabanova E, Dale M, Masi F, Thomas CE, Allin KH, Hansen TH, Heggie A, Hong MG, Elders PJM, Kennedy G, Kokkola T, Pedersen HK, Mahajan A, McEvoy D, Pattou F, Raverdy V, Häussler RS, Sharma S, Thomsen HS, Vangipurapu J, Vestergaard H, 't Hart LM, Adamski J, Musholt PB, Brage S, Brunak S, Dermitzakis E, Frost G, Hansen T, Laakso M, Pedersen O, Ridderstråle M, Ruetten H, Hattersley AT, Walker M, Beulens JWJ, Mari A, Schwenk JM, Gupta R, McCarthy MI, Pearson ER, Bell JD, Pavo I, Franks PW. *Predicting and elucidating the etiology of fatty liver disease: A machine learning modeling and validation study in the IMI DIRECT cohorts*. PLoS Med. 2020 Jun 19;17(6):e1003149. doi: 10.1371/journal.pmed.1003149. PMID: 32559194; PMCID: PMC7304567.

Wilman HR, Parisinos CA, Atabaki-Pasdar N, Kelly M, Thomas EL, Neubauer S… **Fitipaldi H**, … Mahajan A, Hingorani AD, Patel RS, Hemingway H, Franks PW, Bell JD, Banerjee R, Yaghootkar H. *Genetic studies of abdominal MRI data identify genes regulating hepcidin as major determinants of liver iron concentration*. J Hepatol. 2019 Sep;71(3):594-602. doi: 10.1016/j.jhep.2019.05.032. Epub 2019 Jun 19. PMID: 31226389; PMCID: PMC6694204.

Fitipaldi H, McCarthy MI, Florez JC, Franks PW. *A Global Overview of Precision Medicine in Type 2 Diabetes*. Diabetes. 2018 Oct;67(10):1911-1922. doi: 10.2337/dbi17-0045. PMID: 30237159; PMCID: PMC6152339.

List of abbreviations

AI	Artificial intelligence
ANOVA	Analysis of variance
API	Application programming interface
AUC	Area under the curve
BMI	Body Mass Index
CKD	Chronic kidney disease
CNN	Convolutional neural network
COVID-19	Coronavirus disease 2019
CRAN	Comprehensive R Archive Network
CRUSH Covid	Collaborative Research initiative in Uppsala on real-time Interventions of Suspected Hotspots of COVID-19
CSS	COVID Symptom Study
CSSS	COVID Symptom Study Sweden
CVD	Cardiovascular disease
DL	Deep Learning
EBI	European Bioinformatics Institute
EFO	Experimental Factor Ontology
eQTL	Expression quantitative trait locus
FTP	File transfer protocol
FOHM	*Folkhälsomyndigheten* (Public Health Agency of Sweden)
GDPR	General Data Protection Regulation
GWAS	Genome-wide association studies
HIC	High-income country
HIPAA	Health Insurance Portability and Accountability Act
HTML	Hypertext Markup Language
ICD	International Classification of Diseases
KDD	Knowledge discovery from data
LASSO	Least Absolute Shrinkage and Selection Operator
LIC	Low-income country
LMIC	Lower-middle income country
LUDC	Lund University Diabetes Center
MdAPE	Median absolute percentage error

MEDLINE	Medical Literature Analysis and Retrieval System Online
MERS	Middle east respiratory syndrome
ML	Machine Learning
NCBI	National Center for Biotechnology Information
NCD	Non-communicable disease
NER	Named entity recognition
NHGRI	National Human Genome Research Institute
NIH	National Institutes of Health
NLP	Natural language processing
NPR	National Patient Register
PCR	Polymerase chain reaction
PRS	Polygenic risk score
PHEIC	Public health emergency of international concern
PNG	Portable Network Graphics
ROC	Receiver operating characteristic
SARS	Severe acute respiratory syndrome
SARS-CoV-2	Severe acute respiratory syndrome coronavirus 2
SCB	*Statistiska centralbyrån* (Statistics Sweden)
SNP	Single-nucleotide polymorphisms
STEM	Science, technology, engineering, and mathematics
T1D	Type 1 diabetes
T2D	Type 2 diabetes
Tukey HSD	Tukey Honest Significant Difference
UK	United Kingdom
UMIC	Upper-middle income country
US	United States
WHO	World Health Organization

Chapter 1 - Introduction

The advancement of technology in society has greatly enhanced our ability to generate and collect data from a variety of sources. In healthcare, a vast amount of data is being generated by sources such as electronic medical records, wearable devices, and health-related apps (1). This "*big data*" is being collected and stored in large databases, which can then be used to support a variety of healthcare-related activities, such as research, population health management, and precision medicine. As the amount of healthcare and research data generated continues to escalate and become increasingly complex (2), it has become necessary to find new and innovative ways to manage, process, and analyze this data so it can be transformed into meaningful information and knowledge that can be used to transform the practice of medicine and the delivery of healthcare.

Artificial intelligence (AI) is a rapidly growing field that is transforming the way we handle *big data* in healthcare. The massive investment in developing fast parallel chips, specifically graphical processing units (GPUs), and the improvement of algorithms have together contributed to the growth and success of AI in the age of data. Common techniques in AI, such as *deep learning* (DL) and *machine learning* (ML), are being used to analyze complex healthcare datasets and uncover hidden patterns and insights (3-5). This has the potential to revolutionize the way we approach healthcare, enabling more accurate diagnosis and prediction of diseases, leading to more effective treatment and prevention strategies.

Another important tool in the *data science* toolkit for healthcare is *data mining*, which involves using algorithms and statistical models to discover trends, associations, patterns, anomalies, and feature of interest in large datasets (1, 6, 7). By applying *data mining* techniques to health data, researchers, data scientists and public health analysts can gain valuable insights into the factors that affect the health of individuals and communities, and use this information to develop strategies for preventing and controlling public health issues. ML, on the other hand, is a subset of AI that involves training algorithms on large datasets to make predictions or take actions based on those patterns (4, 8). In healthcare, ML can be used to identify patterns in patient data that can be used to make predictions about their health, such as the likelihood of developing a particular disease or the effectiveness of different treatments (3, 9-11). This allows healthcare providers to make more informed decisions about treatment and offer more personalized care. They can also help

researchers and public health officials identify trends and patterns in population health data, and use this information to develop strategies for improving the health of individuals and communities.

As technology continues to advance, there has been an increasing focus on collecting and analyzing genomic data. This has been made possible by the development of new technologies, such as *genome-wide association studies* (GWAS), which allow researchers to quickly and accurately collect and analyze large amounts of genomic data (12). During the past two decades, the development and implementation of technologies like GWAS have paved the way for the rise of *precision medicine*, an approach to healthcare that takes into account individual differences in people's genes, environment, and lifestyle to provide personalized and effective treatments (13). By leveraging ML and *data mining* techniques to facilitate the integration of these data, one can gain a better understanding of factors that might contribute to complex diseases which can be potentially used to develop targeted treatments and prevention strategies, leading to better health outcomes for individuals and communities and a more efficient healthcare system overall (6, 9, 14).

However, it is important to note that all these advances in technologies and tools are not happening without challenges. Bias in health datasets, including genomic data, can significantly diminish the accuracy and usefulness of the information they contain, potentially enhancing health inequalities. For instance, if the data used in GWAS studies is predominantly from a certain racial or ethnic group, the results of those studies may not be generalizable to individuals from other groups. This is because genetic variations can vary greatly among different populations, and therefore, the genetic variants identified in one population may not be the same as those identified in another population. As a result, as individuals from underrepresented groups may not have equal access to personalized and effective treatments, which can perpetuate existing health disparities (15, 16). Therefore, it is crucial that efforts are made to address ethnic and racial bias in health datasets and ensure that they are representative of the diverse populations they are intended to serve. This can involve a range of measures, from increasing the diversity of participants in studies to developing algorithms and techniques that can identify and mitigate bias in the data. Additionally, bias can be introduced through the data collection and analysis processes, leading to inaccurate results (9, 17). Ultimately, addressing biases is essential for ensuring that the benefits of precision medicine are available to all individuals, regardless of their race, ethnicity, or other factors.

The Coronavirus Disease 2019 (COVID-19) pandemic has highlighted the importance of *data science* tools in managing large amounts of data. The unprecedented nature of the situation, in terms of the rapid spread of the virus and the global impact, has resulted in an urgent need for information and knowledge. In

response, the research community, governments, and private sector have made a massive effort to collect and share data in almost real time. The World Health Organization (WHO) signaled that AI could be an important technology to manage the crisis caused by the virus (18). During the pandemic, *data mining* and ML algorithms have been used to analyze large datasets of health-related data, including data on the spread of the virus, the characteristics of infected individuals, and the impact of interventions and treatments (7, 19). App-based symptom trackers allowed individuals to self-report their symptoms, which helped public health officials to track the spread of the virus (20-22). This information was crucial for identifying hot spots and potential outbreaks, as well as for monitoring the effectiveness of various measures taken to control the outbreak. Data *dashboards* were important tools used by public and private institutions to rapidly disseminate knowledge acquired, allowing public health officials to make more informed decisions and to develop targeted interventions and strategies to control the spread of the virus and the public with access to up-to-date information on the pandemic, which was crucial for promoting awareness and understanding of the situation (23, 24).

AI and data mining

Artificial intelligence (AI) and *data mining* are both fields in computer science that are often used together to analyze large datasets. However, they are different in some key ways:

AI has its roots in the 1950s, when computer science pioneers began to ask whether computers could be made to "think" - a question that we are still exploring today. In short, AI is the effort to automate intellectual tasks normally performed by humans, such as problem-solving, learning, and decision-making (4, 25, 26). The broad field of AI includes many different sub-fields and approaches, including *machine learning* (ML), *deep learning* (DL) and *natural language processing* (NLP). These sub-fields focus on different aspects of AI and use different techniques to enable machines to learn from data and make decisions based on that learning.

ML is a sub-field of AI that focuses on the development of algorithms and statistical models that allow computers to improve their performance on a specific task over time (8, 19, 25). The goal of ML is to enable computers to learn from data without being explicitly programmed (27). There are two main categories of ML algorithms: *supervised learning* and *unsupervised learning*. *Supervised learning* algorithms use labeled data to learn a function that maps input data to output labels. For example, a supervised learning algorithm might be trained on a dataset of medical records and their corresponding labels (e.g., "diabetes" or "no diabetes") to learn to predict whether a new patient has diabetes (10). In contrast, *unsupervised learning*

algorithms use unlabeled data to learn patterns in the data. For example, an *unsupervised learning* algorithm might be trained on a dataset of medical records and learn to identify common patterns or risk factors for diseases without any prior labels (11).

DL and AI are terms often used interchangeably. However, DL is actually a sub-field of ML that focuses specifically on the use of *artificial neural networks* to learn from data (8, 25). Neural networks are computational models inspired by the structure and function of the human brain and are composed of many interconnected processing nodes or "neurons". DL algorithms use these neural networks to learn from data in a hierarchical manner, with each layer of the network learning to extract more abstract features of the data as the input passes through the network. This hierarchical structure allows DL algorithms to learn complex patterns in data that are difficult or impossible for other machine learning algorithms to detect (27). There are several different types of DL algorithms, including convolutional neural networks, recurrent neural networks, and generative adversarial networks. These different types of algorithms can be used to solve a wide range of problems in AI, from image recognition and NLP, to robotics and autonomous vehicles (8, 9, 25).

Data mining, also known as *knowledge discovery from data* (KDD), is the automated or convenient extraction of patterns representing knowledge that is implicitly stored or captured in large databases, data warehouses, the web, other large information repositories, or data streams (1). Data mining has been described as a combination of computer science, statistics, and database management with rapidly increasing utilization of AI and visualization (with advanced graphics), especially in the case of big data analytics. The knowledge extraction process usually involves the following steps: i) data cleaning - noise removal; ii) data integration – in which different source of data can be integrate; iii) data selection – filtering the relevant information; iv) data transformation – summary or aggregations; v) data mining – extraction of data patterns involving; vi) pattern evaluation – using proper metrics; and vi) knowledge presentation – data visualization.

By combining the power of data mining and AI, researchers can gain valuable insights from large amounts of data (19). One example is in the field of text analysis, where data mining techniques can be used to automatically extract key information from unstructured text (*text mining*) (1, 29). NLP - a subfield of AI, linguistics and computer science that focuses on making it possible for computers to read, understand, and generate human language in a way that is useful and effective – can then be used as a tool to analyze this information, identifying patterns and trends, and classifying the text into different categories (26)

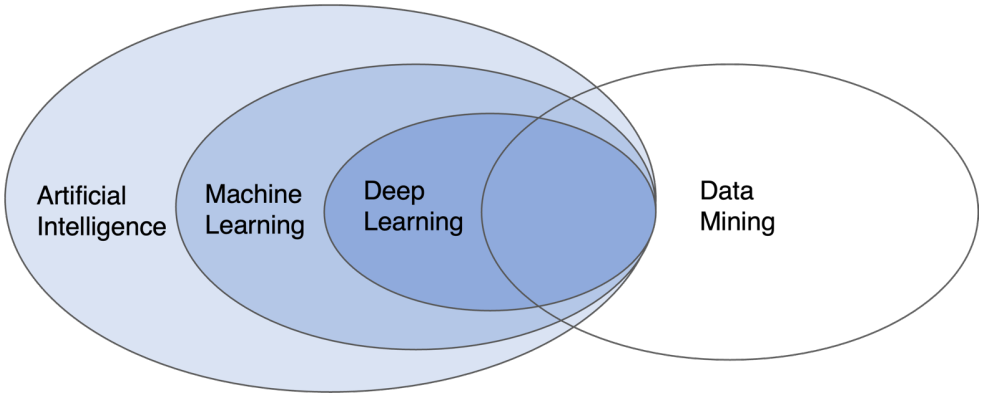


Figure 1 Venn diagram representing the relationships between AI, ML, DL and data mining. Adapted from (28)

As an example, we can consider a pipeline for analyzing health data from medical records. The first step would be to collect and preprocess the text data, which might involve cleaning and normalizing patient records or health reports to remove any unwanted characters or formatting. This process can be automated and scaled with text mining frameworks. Next, to extract useful information from the text data, such as key medical terms or phrases, grouping similar reports together based on the keywords and phrases that they contain, or summarizing the text data in a meaningful way, NLP algorithms can be applied. A next step could involve the use of a ML-based model and classify the text data into different categories, such as diagnosis, treatments, or symptoms. Similar pipelines to the one described are been widely used with COVID-19 data (30, 31).

By combining these fields, researchers can gain valuable insights from large datasets and make more accurate predictions and provide knowledge for data-driven decisions in public health. The evolution of these algorithms aligned with the fairly recent boom in genomic data, has driven efforts towards more personalized public health interventions (2, 5, 9). As these technologies continue to evolve and improve our understanding of complex data, efforts need to be made to ensure that the data used in AI and data mining algorithms is accurate and representative, and that the results of these analyses are used ethically and responsibly. It is also important to consider the potential drawbacks and limitations of these technologies, and to work towards minimizing any negative impacts they may have on individuals and society. Ultimately, the success of AI and data mining in public health will depend on a collaborative effort between researchers, policymakers, and other stakeholders to develop and implement these technologies in a way that is fair, effective, and sustainable.

Genome Wide-Association Studies

GWAS, or *genome-wide association studies*, is a type of observational study used in genetics research to identify genetic variations associated with specific traits or diseases. These studies analyze large amounts of genetic data, typically focusing on *single-nucleotide polymorphisms* (SNPs) and comparing the frequency of these markers in individuals with and without the trait or disease (12, 32-35). The results of GWAS identify specific genetic regions, genes or variations that are associated with the trait or disease, providing a starting point for further research to understand the underlying biology of the trait or disease, and potentially leading to the development of new treatments.

The history of GWAS can be traced back to 2002, when the first GWAS was published in with results that successfully identified a susceptibility gene for myocardial infarction (36, 37). This study, and subsequent ones (38-40), marked the beginning of large-scale genotyping capabilities, which eventually allowed for the quick and inexpensive genotyping of large numbers of genetic variations across the entire genome (41-43). Since then, GWAS has been used to identify genetic associations with a wide range of traits and diseases, including complex *non-communicable diseases* (NCDs) such as type 2 diabetes and cancer (12, 44, 45). The importance of these discoveries lies in its potential to improve our understanding of the genetic basis of disease and to develop new treatments based on this understanding. By identifying specific genetic variations that are associated with a particular disease, researchers can develop targeted therapies that are more effective and have fewer side effects than traditional treatments, also possibly changing public health strategies at the population level (32, 46).

Genetic studies of human disease, particularly GWAS, have been criticized for not adequately representing global diversity (47-54). This under-representation of ethnic diversity potentially hinds our ability to fully comprehend the genetic basis of human disease and can exacerbates health inequalities. This scarcity of ethnic diversity in human genomic studies may also limit our ability to apply findings to clinical practice and public health policy. For example, using estimates of genetic risk from European-based studies on non-Europeans may result in inaccurate risk assessment and inadequate interventions in under-studied populations.

To cite one example on how the focus on specific populations might create health disparities, we can turn to pharmacogenomics, a field of study that involves examining the genetic basis of an individual's response to drugs. For instance, warfarin, a commonly prescribed oral anticoagulant, has a narrow therapeutic range and considerable inter-individual variation in dosage effects. This dose is influenced by SNPs in *CYP2C9*, *VKORC1*, *CYP4F2*, and another variant near the *CYP2C* gene cluster (55). Studies have shown that in Europeans, these SNPs explain a significant

proportion of the variance in drug metabolism; however, in people of African descent, they explain much less (54, 56). Therefore, the algorithms derived from Europeans do not effectively translate to better and safer treatment across ethnic groups. Identifying genetic variants that influence drug metabolism across global populations is necessary to accurately predict drug response in individuals of diverse ethnicities

Overall, GWAS is an important tool in the field of precision medicine, helping to identify genetic variations that may be used to determine an individual's likelihood of developing a particular trait or disease, and to tailor treatment to the individual patient based on their specific genetic profile. However, it is important to note that, the majority of complex diseases are influenced by multiple genetic and environmental factors, so the results of GWAS alone are not sufficient to make accurate predictions or personalize treatment (57, 58). Additionally, potential bias in GWAS studies can harm public health by leading to health disparities and limiting the applicability of the findings to other populations (48, 50, 53). It is therefore essential to carefully consider these potential biases and to ensure that GWAS studies accurately reflect the diversity of the population to maximize the benefits of precision medicine and make it accessible to all individuals, regardless of their ethnicity or other factors.

The COVID-19 pandemic

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the causative agent of COVID-19, was first detected in December 2019 in Wuhan, China (59-61). In January 2020, the World Health Organization (WHO) declared the outbreak a Public Health Emergency of International Concern (PHEIC), and by March 11, 2020, the virus had spread to multiple countries around the world, leading the WHO to declare the outbreak a global pandemic (62). As of December 03, 2022, more than 648 million cases of COVID-19 and 6.65 million deaths have been reported globally (63).

SARS-CoV-2 is the third coronavirus to cause severe disease in humans and spread globally in the past two decades, after SARS and MERS (19, 59-61). The virus is primarily transmitted through person-to-person respiratory contact. This occurs when respiratory particles released by an infected person when they cough, sneeze, or talk come into contact with the mucous membranes of another person. Infection can also occur if a person touches contaminated surfaces and then touches their eyes, nose, or mouth. Prolonged exposure to an infected person is associated with a higher risk of transmission, while briefer exposures to individuals who are symptomatic are less likely to result in transmission (30, 64, 65).

Preventing the spread of SARS-CoV-2 requires measures to reduce person-to-person respiratory contact, such as physical distancing, wearing masks, and hand hygiene (64). WHO recommended that individuals maintained a distance of at least six feet (two meters) from each other, avoid large gatherings, and wear masks in public settings, especially when it was not possible to maintain physical distance. They also recommended frequent hand washing with soap and water or the use of alcohol-based hand sanitizers (66). In addition to these measures, many countries have implemented lockdowns or other restrictions on travel and public gatherings in order to control the spread of the virus (67). These measures can be effective in reducing transmission, but they can also have negative impacts on individuals and society, such as loss of income and social isolation (68). Therefore, it is important for public health officials to carefully consider the potential benefits and drawbacks of such measures when deciding on a course of action.

The incubation period for COVID-19 is between 5-6 days and symptoms may appear 2-14 days after exposure to the virus (69). The clinical presentation is heterogeneous, ranging from asymptomatic to severe or critical illness requiring hospitalization. Although respiratory problems are common in COVID-19, the disease affects multiple organ systems, resulting in a wide range of symptoms, including fever, cough, shortness of breath, fatigue, muscle pain and loss of taste and smell (19, 20, 30, 60, 64, 66).

Certain groups of people are at an increased risk of developing severe illness from COVID-19. These at-risk groups include older adults, people with underlying medical conditions, and people with weakened immune systems (70, 71). Older adults are particularly at risk, with over 81% of COVID-19 deaths occurring in people over age 65. People with underlying medical conditions, such as cancer, chronic kidney disease, chronic liver disease, chronic lung diseases, diabetes, heart conditions, and obesity, are also at an increased risk of severe illness from COVID-19. These conditions can weaken a person's immune system or affect their ability to fight off infections, making them more vulnerable to severe illness from the virus.

During the COVID-19 pandemic, an unprecedented effort was made to accelerate the pace of scientific progress and ensure rapid data sharing (24, 72-74). This allowed researchers around the world to collaborate and share information and insights in near real-time, which was particularly important in the fight against the virus. Thanks to advances in technology, the virus's entire genome was published online within days of its identification (75). Furthermore, the massive effort eventually culminated in the development of vaccines for COVID-19, which were ready to use and approved for emergency use in less than a year (73). This record time development was made possible through the collaboration of scientists, researchers, and pharmaceutical companies, as well as the support of governments and funding agencies.

The COVID-19 pandemic has been showing the potential of real-time and accurate surveillance data for adequate public health decision making and evaluation, as well as for healthcare system preparedness. In response to the pandemic, several app-based solutions have emerged that have facilitated the tracking and monitoring of the spread of the virus (20, 21, 23). In the early phase of the pandemics, when little was known about the symptomology of the COVID-19 and many countries lacked the power for massive testing for the disease, app-based symptom tracking allowed individuals to self-report valuable information such as symptoms and test results. By levering this data with the help of data mining and ML, researchers are able to identify trends and hotspots of infection and use this information to guide testing and other public health interventions (20, 76, 77).

As the world reaches the third year of the pandemic, new challenges have emerged, one of which is the phenomenon of "long-COVID," where individuals continue to experience symptoms for weeks or even months after their initial infection (78-83). Additionally, the emergence of different strains of the virus has raised concerns about the effectiveness of vaccines against the virus (84). Mining the data that was collected through app-based technology can help to better understand the long-term impact of the disease on at-risk populations, as well as the effects of immunization on the syndromic presentation of long-COVID for these populations. It is also crucial to continue collecting and analyzing data about the virus and its different strains to better understanding of its behavior and develop effective vaccines and treatments.

Effective pandemic preparedness requires robust surveillance systems that can quickly and accurately gather and analyze data about the spread of infectious diseases. App-based symptom tracking aligned with data mining and AI tools might help ensure healthcare systems are prepared to handle outbreaks and protect public health. Continued exploration of data related to the pandemic is essential to improving our understanding of COVID-19 and developing effective treatments and prevention strategies for future pandemics.

Objectives and aims

The rapid growth of available health-related data is a result of the computerization of our society and the rapid development of powerful data collection and storage tools. To uncover valuable insights from this "Big Data" and to transform it into organized knowledge, powerful and versatile tools are needed. Data mining tools and AI algorithms have emerged as key pillars in the age of data, with the potential to integrate health data from various sources and uncover hidden patterns in the data, ultimately helping to optimize healthcare. The COVID-19 crisis has brought these

new methods to the forefront, as researchers can use them to derive knowledge from the unprecedented amount of data have been produced.

The objective of this thesis is to explore the use of data mining and AI-based pipelines to mine large health datasets and derive public health evidence. The overarching aims of the papers included in this thesis are as follows:

- **Paper I** – in this paper, data mining and NLP methods are used to extract information from the 2300 GWAS papers published between 2005 and 2022 for the top 10 (of 11) NCDs causes of death. The geographic, ethnic, and socioeconomic characteristics of study populations and researchers was investigated to understand how past two decades of genomic discoveries in NCDs might be transferable across global populations and how high-impact genetics research can be equitably sustained in the future.
- **Paper II** – in this paper, using data from the app-based COVID Symptom Study, data mining pipelines and ML are applied to estimate the individual probability of symptomatic COVID-19, to map the spread of COVID-19 and to predict hospital admissions in Sweden.
- **Paper III** – this paper further explores the COVID Symptom Study data to investigate the difference symptomatology of COVID-19 in pre- and post-vaccination groups at risk (obese and diabetes) and its associations to long-COVID.

Chapter 2 - Methods

Data sources

NHGRI-EBI GWAS Catalog

The NHGRI-EBI GWAS Catalog (hereafter referred as 'GWAS Catalog') is a comprehensive, publicly available online database that compiles data from published GWAS analyses and provides an up-to-date, searchable, and visualizable resource for researchers (https://www.ebi.ac.uk/gwas/) (12). It was founded by the National Human Genome Research Institute (NHGRI) in collaboration with the European Bioinformatics Institute (EBI) in 2008 in response to the increasing number of published GWAS, which provide valuable insights into the genetic basis of complex diseases and traits. The GWAS Catalog is updated on a weekly cycle and serves as a central repository for this data, making it more accessible and easier for scientists, clinicians, and other users be able to integrate these data with other resources.

Initially, eligible published GWAS studies were identified through literature searches and assessed by expert scientists and trained curators, who extracted the reported trait, significant SNP-trait associations, and sample metadata. In 2013, an automated infrastructure was introduced to streamline the curation process and improve data extraction (12, 33). This process involves an automated PubMed search (using the terms "genome-wide" OR "genome AND identification" OR "genome AND association") and extraction of citation information, SNP RSIDs, traits, and P-value information from eligible papers.

Studies are eligible for inclusion in the GWAS Catalog if they include at least 100,000 SNPs in the initial stage before quality control filters are applied and have statistical significance (SNP-trait p-value $<1.0 \times 10^{-5}$) in the overall (initial GWAS + replication) population (12, 32). Studies are excluded if they are published in a language other than English, if the SNPs assayed are limited to those in candidate genes, if the samples are assayed to measure somatic variation, or if the study does not include any new GWAS data. Information on author, study date, PubMed URL, publication title, disease/trait information, sample sizes, platform, and number of SNPs passing quality control metrics is extracted for each eligible study. Curated

trait descriptions are mapped to Experimental Factor Ontology (EFO) terms, a highly adaptable and extensible ontology that enables richer querying than simple string searching and facilitates data integration across heterogeneous data sources such as data extracted from the scientific literature (12, 32, 33).

The GWAS Catalog website includes a search interface, API, and a GWAS diagram that maps SNP-trait associations onto the human genome by chromosomal location and displays them on the human karyotype (https://www.ebi.ac.uk/gwas/diagram). Summary statistics files are available for download from the Catalog's FTP site, and harmonized summary statistics are also available from the summary statistics database via API. In addition, the Catalog accepts submissions of unpublished GWAS data since 2020, which is available for download through a separate section of the website (34).

The GWAS Catalog has seen significant improvements in its data release frequency and functionality, making it a valuable resource for researchers studying the genetic basis of complex diseases and traits. The data from the Catalog has helped identify causal variants, the comprehension of disease mechanisms, and the analysis of expression quantitative trait loci (eQTL) (32-34, 85). As of December 2022, the Catalog contains information from 6130 unique publications and 447,939 variant-trait associations and continues to evolve and improve as a rich source of data for genetic research.

PubMed

PubMed, maintained by the National Institutes of Health (NIH), is a widely used database of biomedical literature and life sciences journals, including those indexed in MEDLINE (https://pubmed.ncbi.nlm.nih.gov/) (86). It holds millions of abstracts and full-text articles from scientific journals, making it a valuable resource for researchers and those interested in the latest advancements in the biomedical field. To extract metadata information from publications in PubMed, one can use the Entrez Programming Utilities (E-utilities), a public API maintained by the National Center for Biotechnology Information (NCBI).

The **first paper** of this thesis uses GWAS Catalog as the main source to identify target GWAS studies for its analysis, as well as a source of information about the participants (cohorts). Additional publication metadata for this paper was extracted from PubMed (figure 2.1).

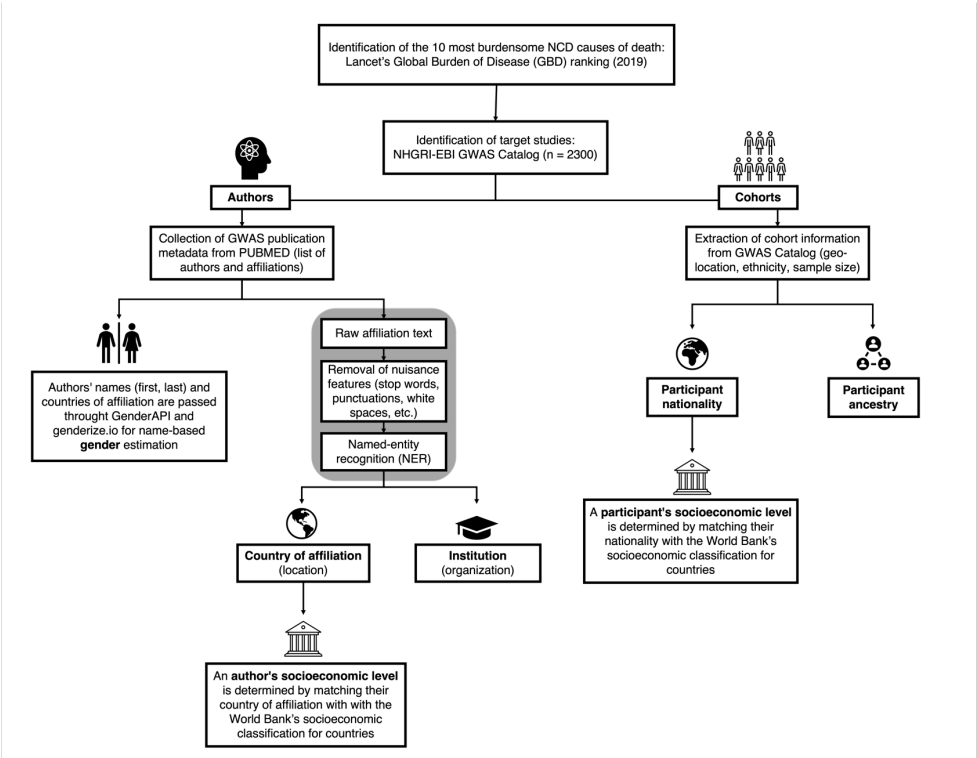


Figure 2.1 Study flowchart for paper I.

COVID Symptom Study

The COVID Symptom Study (CSS) is a large-scale, app-based research project that aims to map the spread of the of SARS-CoV-2 in the United Kingdom (UK), United States (US) and Sweden and to increase the knowledge about COVID-19, investigating symptomatology of the disease in different subgroups of participants and during different phases of the pandemic (20, 21, 87). The app was developed and is maintained by the health data science company ZOE Global Ltd, and the study is conducted in collaboration with King's College London (UK), Massachusetts General Hospital (US), Lund University (Sweden), and Uppsala University (Sweden). The app was launched in the UK on 24 March 2020, in the US on 29 March 2020, and in Sweden on 29 April 2020. By early 2022, the app had reached 4.7 million participants and over 500 million assessments across the three countries (Figure 2.2).

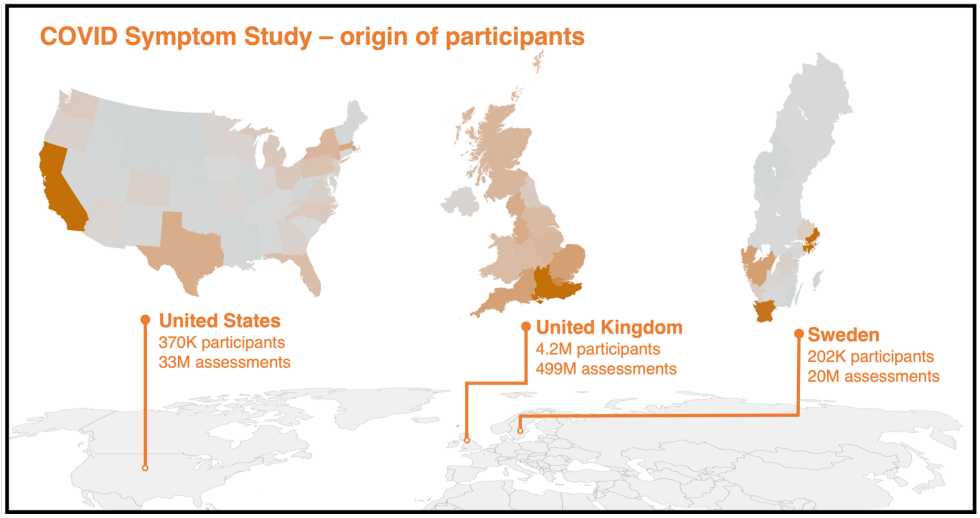


Figure 2.2 COVID Symptom Study population by April 2022. Country's choropleth are independents, with the color scheme representing participant density based on total number of participants per country. Darker colors indicate more participants in the region.

Upon registering for the study, participants were asked to provide baseline demographic information (e.g. age, sex, weight, and height), geographic location (postcode), and clinical history (comorbidities and lifestyle) (21, 88) . UK participants enrolled in ongoing epidemiological studies, clinical cohorts, or clinical trials had the possibility to provide informed consent to link survey data collected through the app to their existing study cohort data and any relevant biospecimens in a manner compliant with the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR). Daily prompts delivered through the app encouraged participants to provide updates on their health status, symptoms, health care visits, COVID-19 testing results, and vaccination doses (Figure 2.3). Throughout the study period, the app was adapted to address emerging research questions.

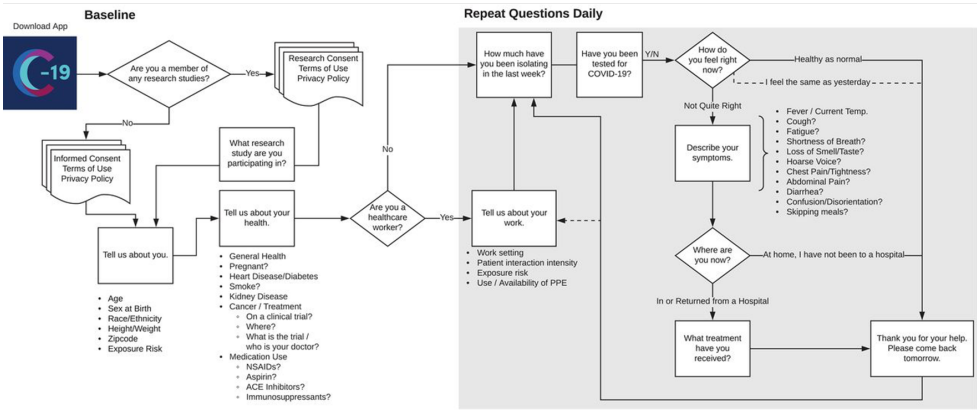


Figure 2.3 COVID Symptom Study app workflow. Source reference (21)

Paper II and III were conducted using data from the CSS. The second paper mainly used samples from the Swedish cohort (referred to as COVID Symptom Study Sweden or CSSS) while the third used data from participants from all three countries. In the next sections, I will shortly describe the databases used in the second paper for external comparison or validation of our results.

SMiNet

SmiNet is an electronic notification system of communicable diseases maintained by *Folkhälsomyndigheten* (FOHM), the Public Health Agency of Sweden (89). All of Sweden's infection control units are connected to the system, as are the microbiological laboratories and thousands of care units. The system is used to monitor diseases that are notifiable according to the Infection Control Act. When doctors and laboratories in Sweden report cases of infectious diseases to SmiNet, the regions' infection control units and FOHM can follow the epidemiological situation of various diseases both locally and nationally, enabling early detection of outbreaks and enabling measures to be taken. COVID-19 is a mandatory notifiable disease in Sweden, so all clinical laboratories are required by law to report positive PCR tests for SARS-CoV-2 to SmiNet.

NOVUS

NOVUS is a private company that conducts opinion polls and surveys using a panel of individuals recruited from the Swedish population using random sampling methods (90). The NOVUS Sweden Panel includes approximately 45,000 individuals who are representative of Sweden in terms of age, sex, and region of residence. Recruitment is mainly done through randomly determined phone

interviews, and personal invitations are sent to individuals who do not have registered phone numbers or do not answer their phones. Self-recruitment is not possible, and panel members who have not responded to at least one survey in the last three months are replaced. Since the early phase of the pandemic (March 2020), NOVUS has carried out repeated surveys on COVID-19-related symptoms.

CRUSH Covid

Collaborative Research initiative in Uppsala on real-time Interventions of Suspected Hotspots of COVID-19 (CRUSH Covid) is a research project in collaboration between Region Uppsala and researchers from five different departments at Uppsala University (91). The project mapped and sought to mitigate the spread of infection and local outbreaks of COVID-19 in Uppsala County, Sweden. The researchers have developed methods to combine information from several different data sources, such as surveys conducted after a PCR+ test, data from the Swedish CSS cohort, and measurements of the SARS-CoV-2 virus in sewage, to provide early signals of increased spread of infection in real time. The project also evaluates Region Uppsala's testing strategies and targeted measures and investigated which groups were most likely to experience surges in COVID-19 infection.

National Patient Register

The National Patient Register (NPR) is a database maintained by the National Board of Health and Welfare in Sweden (*Socialstyrelsen*) (92). It collects information on patient diagnoses from inpatient and outpatient visits to specialist care, and diagnoses are classified according to the International Statistical Classification of Diseases and Related Health Problems Tenth Revision (ICD-10). The register is used for statistics on diseases and treatments in Swedish specialist care, and it is also an important source of data for research and government evaluations. The register's data is used to improve the quality and safety of healthcare, and to ensure that healthcare resources are distributed fairly and on equal terms. NPR uses the ICD-10 codes U07.1 and U07.2 for COVID-19 cases.

Analytical methods

The analytical methods implemented in this thesis were performed using R software (versions 3.6.1 and 4.1.2) (93). Different R libraries and packages were used to build the analytical tools (see section below “Developing R packages” for more details),

which can be accessed in the open-source code available on the published papers' GitHub repositories.

Text mining and NLP: text pre-processing

Text pre-processing is a set of operations or techniques that are applied to raw text data to prepare it for further analysis. It is an important step in data mining and NLP because it helps improve the quality and consistency of the text data (1, 26). Some common steps in text pre-processing include:

1. Removing punctuation and special characters: This step helps remove text marks such as commas, periods, exclamation marks, and quotation marks from the text. This is important because punctuation marks can interfere with the analysis and processing of the text. For example, punctuation marks can create unnecessary breaks in the text and can cause words to be split up in ways that are not natural or meaningful. By removing punctuation marks, the text becomes more uniform and easier to process.
2. Lowercasing: This step involves converting all text to lowercase. This is important because many NLP algorithms and tools are case-sensitive, meaning that they treat words that are capitalized differently from words that are not. By converting the text to lowercase, the text becomes more uniform and easier to process.
3. Tokenization: This step involves breaking the text up into individual tokens (word or unit of meaning). This is important because many NLP algorithms and tools operate on individual words, rather than on the entire text. By breaking the text up into individual words, the text becomes more manageable and easier to analyze.
4. Removing stop words: This step involves removing common words that do not add meaning to the text. Stop words are words that are commonly used in everyday language, such as "a," "an," "the," and "of," but that do not convey any specific meaning. Because stop words do not add any meaning to the text, they can often be removed without affecting the overall meaning of the text. By removing stop words, the text becomes more concise and easier to analyze.
5. Stemming or lemmatization: This step involves reducing words to their base forms. This is important because many words in the English language have different forms, such as "run," "ran," and "running," which can cause confusion for NLP algorithms and tools. By reducing words to their base forms, related words can be grouped together, and the accuracy of NLP algorithms can be improved.

Named entity recognition (NER)

Named entity recognition (NER) is a subfield of NLP that uses artificial intelligence to identify and classify key pieces of information, known as entities, in text (94). These entities can be categorized into pre-defined categories, such as "person," "city," and "organization," or they can be more specific to the given task. To train a model for NER, a labeled dataset must be used, with the entities and their corresponding categories clearly defined.

SpaCy is a popular open-source NLP library that offers various text-processing capabilities including NER (95, 96). It is implemented in multiple languages and uses convolutional neural networks (CNNs) as its underlying model. The package offers pre-trained models that can be used out-of-the-box, such as the *en_core_web_sm* model, which is trained on web data and provides basic NLP capabilities. SpaCy is widely used in the NLP community for its efficiency, accuracy, and ease of use.

In **paper I**, text pre-processing was first used to clean and tidy PubMed metadata (authorship lists) and NER was implemented to identify entities in the text, such as organizations (institutions) and countries of affiliation.

Name-to-gender inference

The process of predicting the gender of an individual based on their name is known as name-to-gender inference (97). This task can be accomplished using machine learning algorithms that are trained on datasets such as censuses, birth lists, and self-labeled data from social media. In NLP, this technique is frequently employed in tasks such as language translation, as knowledge of a speaker or writer's gender can provide valuable contextual information or allow for the tailored production of output for specific audiences. Name-to-gender inference is also a rapid and cost-effective means of examining gender disparities in text-based documents, such as in scientific publications, books, conference proceedings and grant allocations (98). Name-to-gender inference also has limitations, as it can potentially perpetuate gender stereotypes or perpetuate discrimination against individuals whose names do not conform to the predictions made by the model (99).

In **paper I**, two popular web-based name-to-gender inference services, Gender API (https://gender-api.com/) and genderize.io (https://genderize.io/), are used to estimate the gender of authors in GWAS publications.

Linear Regression

Linear regression is a statistical method used to model the linear relationship between a dependent variable and one (simple linear regression) or more (multiple

linear regression) independent variables (100). It is a widely used technique in statistics and ML, especially common in predictive analysis.

The basic idea behind linear regression is to fit a straight line (or hyperplane in the case of multiple independent variables) to the data in such a way that the distance between the data points and the fitted line is minimized. This distance is measured using a loss function, such as the mean squared error, which is the sum of the squares of the differences between the predicted values and the true values. The mathematical formula for simple linear regression is as follows:

$$y = \beta 0 + \beta 1 x$$

Where y is the dependent variable, x is the independent variable, and $\beta 0$ and $\beta 1$ are the coefficients (or parameters) of the model. $\beta 0$ is the intercept, which is the value of y when x is 0, and $\beta 1$ is the slope of the line, which determines the rate at which y changes as x increases.

Simple linear regression is a useful tool for understanding the relationship between two variables and is often used as a starting point for more complex regression models. The method has some key limitations such as the assumption that the relationship between the variables is linear, which may not be true.

In **paper I**, simple linear regression is used to evaluate temporal trends in percentages of authors' geographic location, gender, and socioeconomic levels, as well as origins and ethnicity of participants in GWAS publications in NCDs. In **paper II**, a weighted linear regression model was used to predict future hospitalizations due to COVID-19 in Swedish regions based on the CSSS prevalence estimates and current rate of hospitalizations.

The Cochran-Armitage test for trend

The Cochran-Armitage test for trend is a statistical test used to determine whether there is a statistically significant trend in a categorical data set (101). It is often used to test for a trend in proportions over time, or in response rates across different groups, as it allows for an assessment of whether the odds of a particular outcome increase or decrease as the independent variable changes.

The test is based on the Cochran-Armitage statistic, which is calculated as the sum of the products of the number of observations in each category and the rank of that category. The ranks are assigned based on the order of the categories, with the smallest category receiving a rank of 1 and the largest category receiving a rank equal to the number of categories. The test statistic is then compared to a critical value, which is determined based on the sample size and the number of categories.

If the test statistic is greater than the critical value, this indicates that there is a statistically significant trend in the data. If the test statistic is less than the critical value, there is not a statistically significant trend.

The Cochran-Armitage test for trend is a widely accepted and commonly used method for analyzing trends in categorical data. One of its key advantages is that it is relatively easy to compute and does not require any assumptions about the distribution of the data. However, it is important to note that the test can only detect linear trends and may not be appropriate for detecting more complex patterns in the data. Moreover, it is important to note that this test is one-tailed, thus it only tests for a linear trend in one direction (increasing or decreasing). If the aim is to test for a trend in both directions, it is necessary to perform two separate tests, one for each direction and then adjust the p-value accordingly.

There are a few R packages that can be used to perform the Cochran-Armitage test for trend. One popular package is *DescTools*, which provides the CochranArmitageTest() function for performing the test (102).

In **paper I**, Cochran-Armitage test for trend is used in parallel with the simple linear models for the evaluation of trends in proportions of authors' geographic location, gender, and socioeconomic levels, as well as origins and ethnicity of the data points in GWAS publications in the area of NCDs.

L1 Penalized Logistic Regression

Logistic regression is a statistical model used for binary classification tasks, where the goal is to predict the probability that an instance belongs to a certain class (100, 101). In logistic regression, the relationship between the dependent variable (the class label) and the independent variables (the features) is modeled using the logistic (sigmoid) function, which is defined as follows:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

where z is the linear combination of the independent variables and the model parameters, also known as the logit. The logistic function maps the input values z to the range between 0 and 1, which can be interpreted as the probability of the instance belonging to the positive class.

The logistic regression model can be represented mathematically as follows:

$$\hat{y} = \sigma(w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n)$$

where $\hat{y}$ is the predicted probability of the instance belonging to the positive class, and $w_0, w_1, \ldots, w_n$ are the model parameters (also known as weights or coefficients)

that are learned from the training data. The model parameters are usually estimated using maximum likelihood estimation or maximum *a posteriori* estimation.

LASSO, or Least Absolute Shrinkage and Selection Operator, is a regularization method for regression models that aims to reduce the complexity of the model and prevent overfitting (100). It does this by introducing a penalty term, which is defined as the sum of the absolute values of the model parameters. This encourages the model to set some of the parameters to zero, effectively selecting a subset of the most important features. The objective function of the LASSO model is defined as follows:

$$\min_{w} \frac{1}{2} \left|\left|Xw - y\right|\right|_2^2 + \alpha \left|\left|w\right|\right|_1$$

where X is the design matrix, y is the target vector, w is the model parameters vector, and α is a hyperparameter that controls the strength of the regularization. The LASSO objective function can be minimized using an optimization algorithm such as coordinate descent.

L1-penalized logistic regression is a variant of logistic regression that combines the logistic function with the LASSO regularization (103). It is commonly employed for binary classification tasks when the goal s not only to predict the probability of an instance belonging to a certain class, but also to select a subset of the most important features. The objective function of L1-penalized logistic regression can be formulated as follows:

$$\min_{w} \frac{1}{m} \sum_{i=1}^{m} [-y_i \log \hat{y}_i - (1 - y_i) \log(1 - \hat{y}_i)] + \alpha \left|\left|w\right|\right|_1$$

where m denotes the number of training examples, y_i is the target value of the i-th example, $\hat{y}_i$ is the predicted probability of the i-th example belonging to the positive class, w represents the model parameters vector, and α is a hyperparameter that controls the strength of the regularization.

In **paper II**, an L1-penalized logistic regression model was used to predict the probability of a CSSS participant having COVID-19, selecting variables which include a range of reported symptoms. In **paper III**, logistic regression is used to estimate the propensity scores in the propensity score weighting analysis of diabetes.

Multinomial regression

Multinomial regression, a generalization of logistic regression, is a type of classification model that is used when the response variable has more than two

categories (100, 101). In multinomial regression, the model estimates the probability of each category of the response variable based on a set of predictors. The model is trained using a dataset comprising observations and their corresponding values for the dependent and predictor variables. Subsequently, the model utilizes these estimates to predict the category of the dependent variable for new observations.

The general form of the probability function for multinomial regression can be written as:

$$P(Y = k \mid X) = \frac{\exp(\beta_k^T X)}{\sum_{j=1}^{K} \exp(\beta_j^T X)}$$

where:

Y is the dependent variable with K categories.

X is a vector of predictor variables.

β_k is a vector of coefficients for the kth category of the dependent variable.

K is the total number of categories in the dependent variable.

This general formula assumes that the predictor variables are independent of one another and that the model is linear in the parameters. Additionally, it uses the exponential function to estimate the probability of each category, but other forms of multinomial regression could use other functions to estimate the probability.

In **paper III**, multinomial regression is used to estimate the propensity scores in the propensity score weighting analysis of BMI.

Propensity score weighting

Propensity score weighting is a statistical method that can be used for addressing confounding in observational studies (104). It is based on the concept of propensity score analysis (105), which is the conditional probability of assignment to a treatment condition given a set of observed covariates: e = p(z=i|X). By utilizing propensity scores, the resulting groups will have similar characteristics to those created through random assignment.

Matching is one of the most common applications of propensity scores (Thoemmes & Kim, 2011). However, a potential drawback of using propensity scores for matching is that a large number of subjects may be needed, particularly in the control group (Guo & Fraser, 2015), and depending on the specific matching technique, caliper, and number of subjects matched to each subject in the control group, a significant number of subjects in the control group may not be used.

Alternatively, propensity scores can also be employed as weights (known as propensity score weighting).

In essence, propensity score weighting is a method for adjusting for confounding by estimating the probability (or propensity) that an individual will be exposed to a certain treatment or factor, based on baseline information. This probability is estimated using a statistical model (such as logistic regression or multinomial regression). Once the propensity scores have been calculated, they can be used to weight the observations in the study so that the exposed and unexposed groups are more similar on average. This helps to control for the influence of confounding factors and can provide a more accurate estimate of the relationship between the exposure and the outcome.

In **paper III**, propensity score weighting is used to control for confounders and examine the prevalence of COVID-19 symptoms among CSS participants with different body mass index (BMI) or diabetes status.

Developing R packages

In R, a package is a collection of functions, data, and documentation that are bundled together and made available for use (106). These packages are an essential component of the R ecosystem and are designed to facilitate the sharing and reuse of code among users. There are >18,000 packages available on the Comprehensive R Archive Network (CRAN), a public repository that serves as a clearinghouse for R packages. The large number of available packages is a major factor in the popularity of R, as it allows users to benefit from the work of others by easily downloading and using packages that solve specific problems.

The development of R packages is important for several reasons, particularly in the realm of open-source and reproducible science. First, it allows researchers to share their work with others, enabling other scientists to replicate and extend prior findings. This is critical for the advancement of knowledge and the validation of scientific results. Second, R packages can provide a standardized and structured way to distribute and document scientific code, making it easier for others to understand and use. This is especially important in the context of reproducibility, as it allows others to replicate the results of a study by following the same steps that were used to generate them.

Using the usual pipeline methodology to develop R packages (figure 2.4), two R packages were developed in conjunction with **paper I**, *affiliation* and *genderAPI* (107, 108); and one package was developed in conjunction with paper II, *covidsymptom* (109).

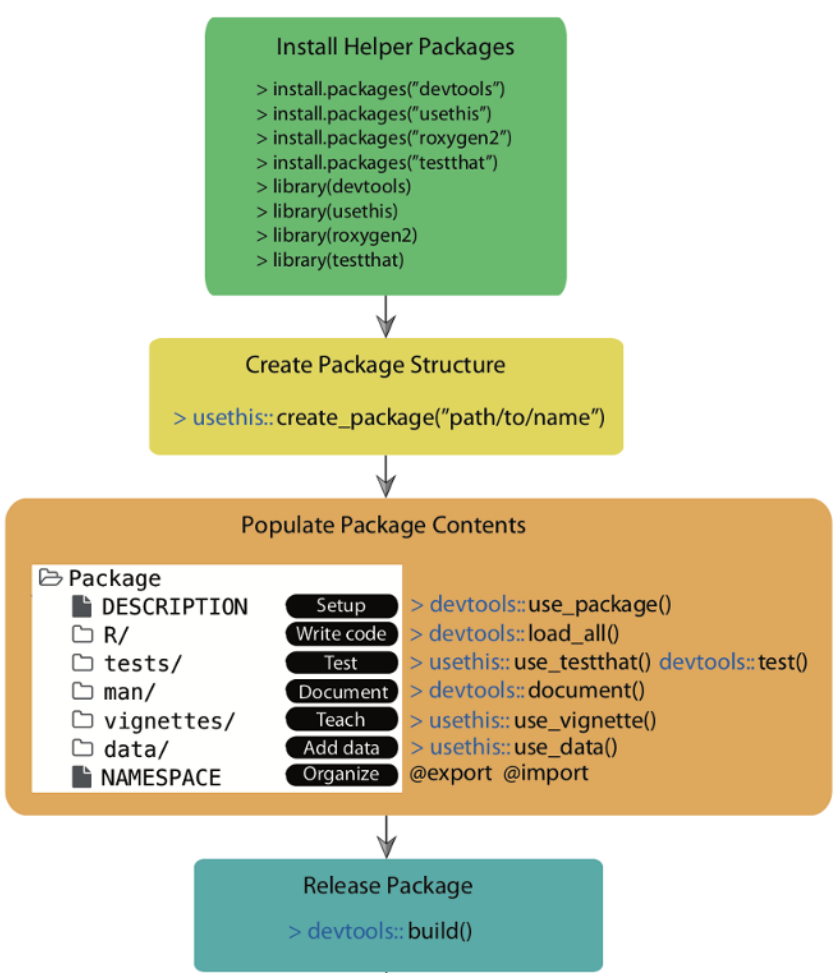


Figure 2.4 R package development workflow. Source reference (110)

Chapter 3 – Results

This following section summarizes the main results of the papers included in this thesis. Some of the methods are contextualized to each study and briefly overviewed. More detailed results are provided in the papers and manuscript at the end of this thesis. Additionally, at the end of this chapter, I will introduce the COVID Symptom Study Sweden (CSSS) Dashboard, an interactive web-tool which I developed to keep the general public and health authorities in Sweden informed about CSSS's latest results. The dashboard was updated daily, providing information that helped increase understanding about the spread of the virus in Sweden.

Paper I

In the first paper, we used data-mining and NLP techniques to examine data from the GWAS Catalog (2005-2022) concerning the top-10 (of 11) non-communicable causes of death identified in The Lancet's Global Burden of Disease ranking (2019) (111). We contrasted the geographic, ethnic, and socioeconomic characteristics of the study populations and the geographic, socioeconomic, and gender characteristics of the researchers. Utilizing EFO codes, we identified 2300 target studies in the GWAS catalog and obtained additional metadata on these studies through PubMed for the following disease categories: cardiovascular disease (CVD), cancer, chronic respiratory diseases, diabetes, and chronic kidney disease (diabetes and CKD), digestive diseases, mental disorders, musculoskeletal disorders, neurological disorders, skin diseases, and substance use. Additionally, we included a category called "*all traits*" that encompassed all identified studies on non-communicable causes of death, excluding duplicates.

Utilizing our R package, *affiliation* (107), I extracted and analyzed the authorship data from the target studies. Text mining and NER techniques were employed to determine the institutions and nations in which the authors were affiliated. The authors were then classified as *all authors*, *first authors*, or *senior authors*, according to their placement on the author list. Our findings demonstrated that, for *all traits*, the United States (US) had the most co-authors affiliated with its institutions, comprising 37% of *all authors*, 40% of *first authors*, and 41% of *senior authors* (Figure 3.1). The United Kingdom (UK), China, Japan, and Germany followed in corresponding order. Within specific disease areas, the US generally

had the most affiliated authors, with the exception of 'musculoskeletal disorders', where the UK had the highest representation. We also showed that, throughout the analysis period (2005-2022), for *all traits*, the US experienced a decline in authorship dominance, experiencing a 2.4% mean decline in authorship dominance each year (95% CI: −4.05, −0.84%; P < 0.01). Countries such as Australia, China, South Korea, and Spain have increased their dominance, with annual average increases of 0.14% (95% CI, 0.06, 0.22; P < 0.01), 0.74% (95% CI, 0.45, 1.04%; P < 0.01), 0.30% (95% CI, 0.13, 0.47%; P < 0.01) and 0.16% (95% CI, 0.05, 0.28%; P = 0.01), respectively.

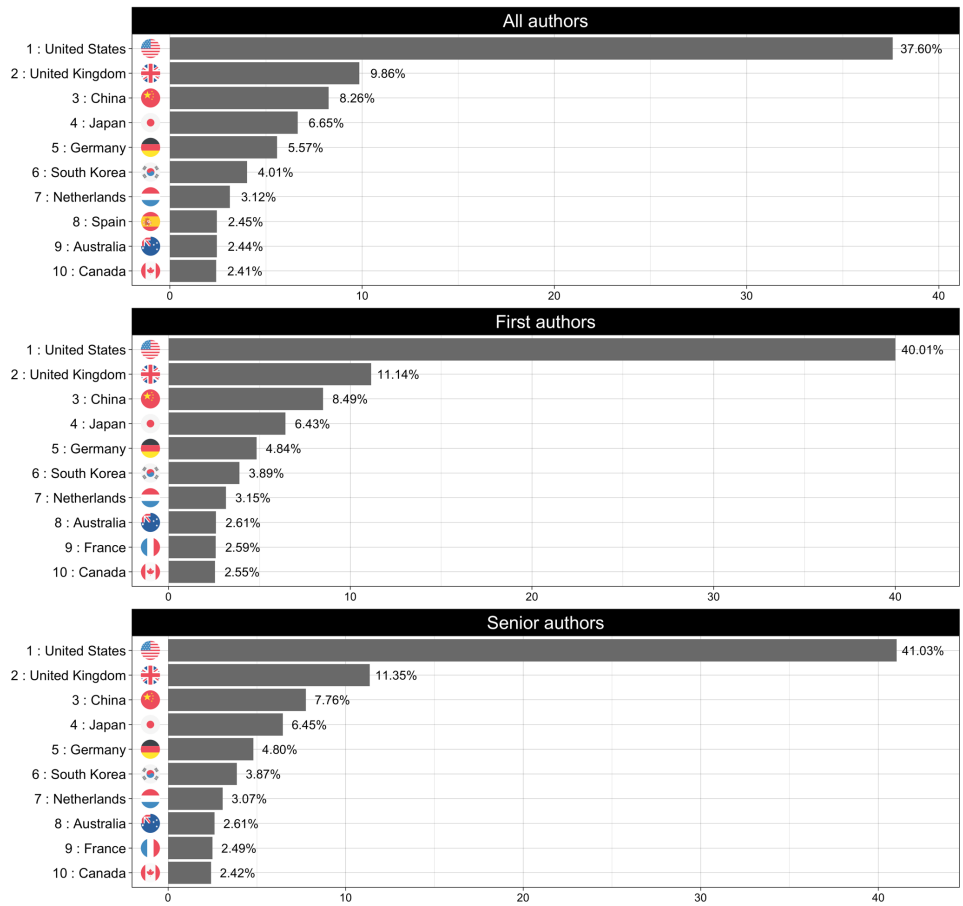


Figure 3.1 Top 10 countries of affiliation for the "all traits" disease category. The panels show the ranked list of countries of affiliation for each of the defined authors categories (all, first and senior).

The preeminence of US-based researchers is also evident in the ranking of institutions of affiliations. The *dominance score*, which measures the proportion of

co-authors from a given institution relative to all co-authors, showed that Harvard Medical School in the US ranked first for *all traits*, with 1.2% of *all authors* affiliated with this institution. deCODE genetics, in Iceland, ranked first for *first authors*, and Harvard Medical School ranked first for *senior authors*. The ubiquity score, which measures the frequency with which a given institution appears within co-author affiliations across GWAS publications, showed that seven out of the top ten institutions for *all traits* were located in the US, with Harvard Medical School ranking first and having co-authors on 15% of such publications. The second and third ranked institutions were Karolinska Institutet in Sweden and the Broad Institute in the US.

In this study, the gender of authors of scientific papers was estimated using two online platforms, genderize.io and Gender API, both accessed through R, and the last using an R package that I developed for this study, *genderAPI* (108). Our results showed that male authors were overrepresented in all disease categories. For *all traits*, the gender imbalance in *first* and *senior* authorships improved over time, with an increase in the proportion of female first and senior authors from 2005 to 2021. The proportion of female *first authors* increased from 33.3% in 2006 to 42.2% in 2021, with an average annual increase of 0.93% (95% CI, 0.52, 1.35%; $P < 0.01$) (Figure 3.2).

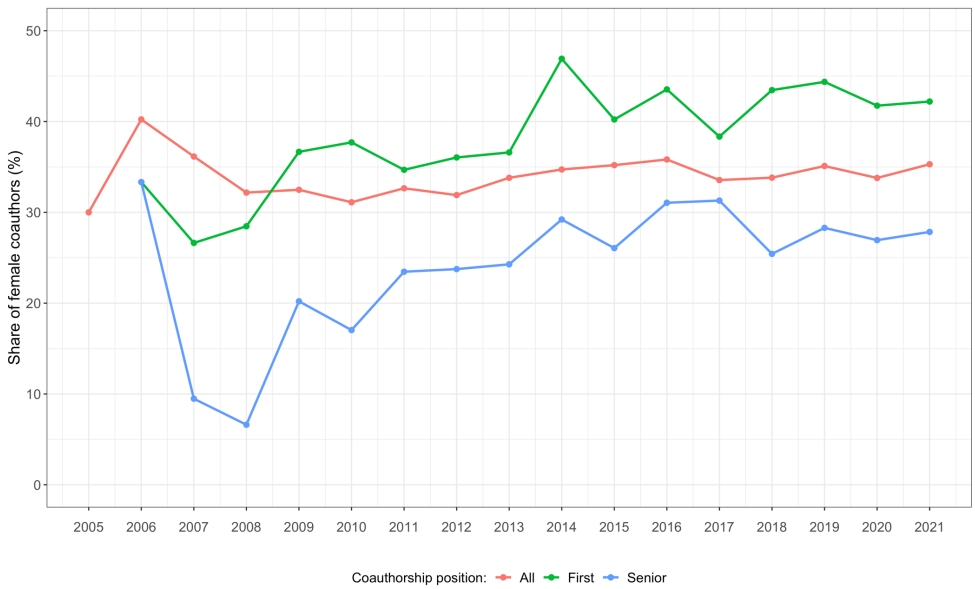


Figure 3.2 Trends of female authors in "all traits" category across the studied period (2005–2021). The figure shows changes in the representation of women in all (red), first (green) and senior (blue) authorship positions by year.

Data on study participants within each GWAS paper was mined from the resources offered by the GWAS Catalog. Our results show that the UK contributed the most data overall (*all traits* category), at 34%, followed by the US at 16%. However, in the specific areas of diabetes and CKD, the US contributed the most data at 35%, followed by the UK at 21%. During the analysis period, contributions from Germany, the Netherlands, and the US have decreased, while the contributions from Norway and the UK have increased. No other country's data contributions have changed significantly over time.

We have also analyzed the information about the ancestry of the study participants for the GWAS included in this study. Our findings show that, for *all traits*, the majority of participants were of European descent (91%), with East Asians being the second most well-represented at 4.9%. This pattern remained constant across all disease areas, and European ancestry accounted for over or close to 90% of the representation of ancestry (figure 3.3).

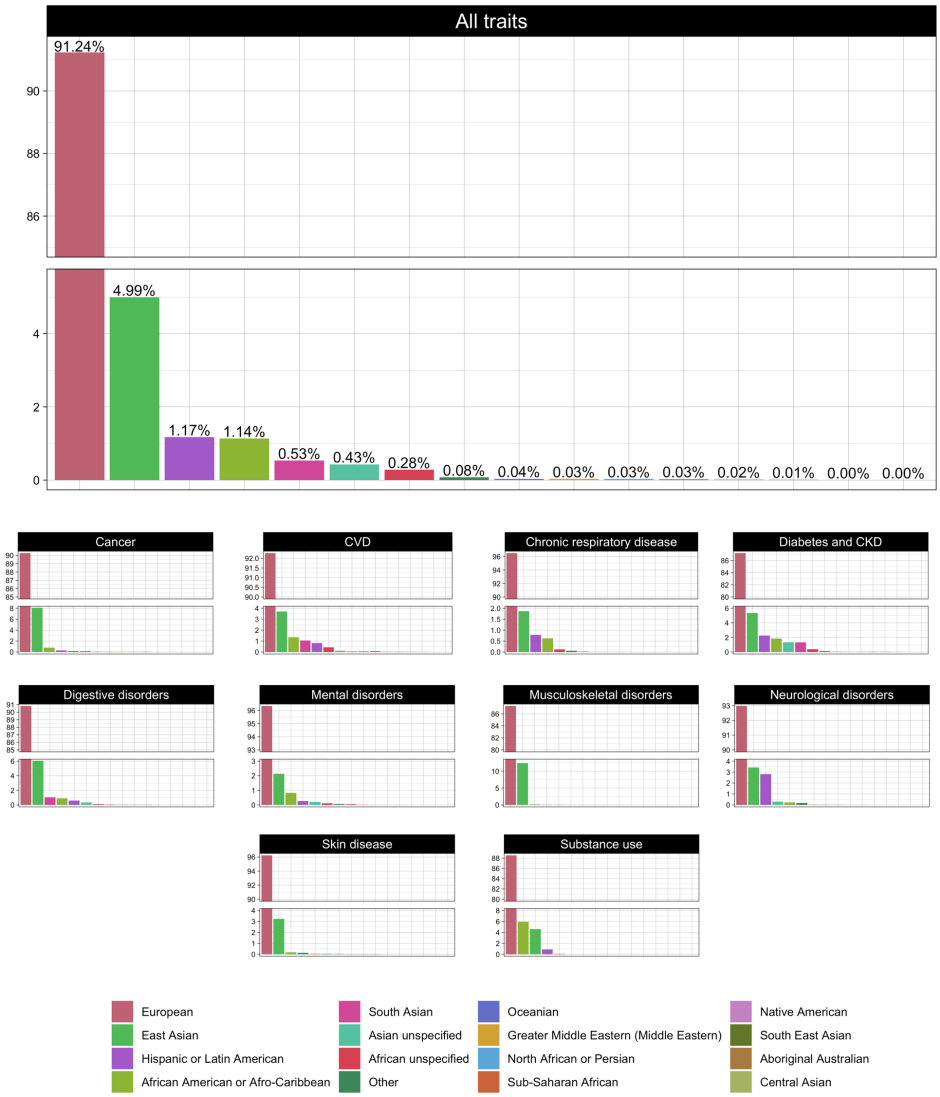


Figure 3.3 Ancestry of study participants: the proportion of ancestry groups in "all traits" (upper panel) and across the 10 disease areas (lower panel).

Finally, the results of economic background analysis showed that across all disease categories, co-authors were affiliated to high-income countries (HIC) and participants were also mostly originated from the same socioeconomic category. A temporal analysis across the study period for the *all traits* category showed a slight decrease in authors affiliated with HICs, with average annual decrease of −0.99%

per year (95% CI, −1.30, −0.68% per year; P < 0.01) for all authors. During this period, for *all authors*, upper-middle income countries (UMICs) and lower-middle income countries (LMICs) had average annual increases of 0.91% (95% CI, 0.61, 1.22% per year; P < 0.01) and 0.06% per year (95% CI, 0.02, 0.10% per year; P < 0.01), respectively.

To disseminate the full set of results from this project, I built an interactive R Shiny web application (Figure 3.4), available at:

https://hugofitipaldi.shinyapps.io/gwas_results/.

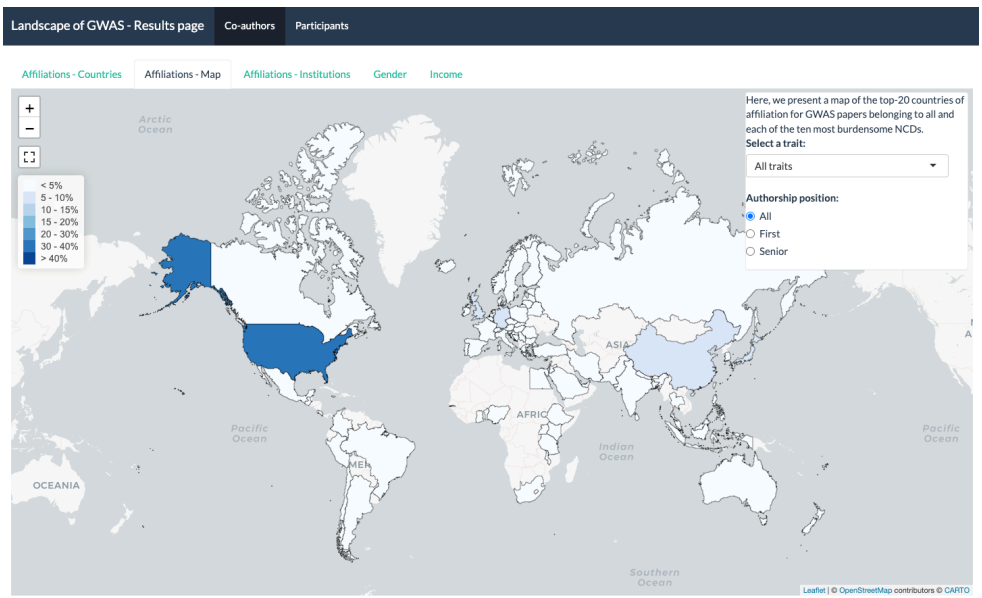


Figure 3.4 Paper I results' page. The page features two primary tabs: Co-Authors and Participants. The figure illustrates the Affiliations-Map sub-tab, which displays an interactive choropleth map of the countries of affiliation for all studied traits.

Paper II

The aim of **paper II** was to develop and evaluate a syndromic surveillance-based framework to estimate the regional prevalence of COVID-19, and to predict subsequent trends in COVID-19 hospital admissions.

This study utilized data from the CSSS and included 143,531 individuals aged 18-years or older who contributed at least one daily report between April 29, 2020, and February 10, 2021. The study cohort included more female participants and a lower proportion of individuals over 65 years of age, smokers, and residents of deprived postal-code areas compared to the general population. The highest CSSS participation rates were observed in the regions of Skåne, Uppsala, and Stockholm.

Our statistical analysis strategy had five steps (Figure 3.5). The first step involved training a L1-penalized logistic regression (LASSO) model to select variables predicting symptomatic COVID-19 using data from 19,161 participants who reported at least one PCR test result between April 29 and December 31, 2020, and who reported at least one symptom within 7 days before or on the test date. The final model included 17 symptoms and sex, as well as two-way interactions between loss of smell and/or taste and 14 other symptoms, and a two-way interaction between loss of smell and/or taste and sex. The model had a receiver operating characteristic (ROC) area under the curve (AUC) of 0.76 (95% CI 0.75-0.78) during the training period (April 29-December 31, 2020) and 0.72 (95% CI 0.69-0.75) during the evaluation period (January 1-February 10, 2021). The AUC in an external dataset of 943 symptomatic individuals from the CRUSH Covid survey was 0.78 (95% CI 0.74-0.83).

The second step involved estimating the prevalence of COVID-19 in the study population using the model trained in the first step and data from all CSSS participants who reported at least one symptom between April 29 and December 31, 2020.

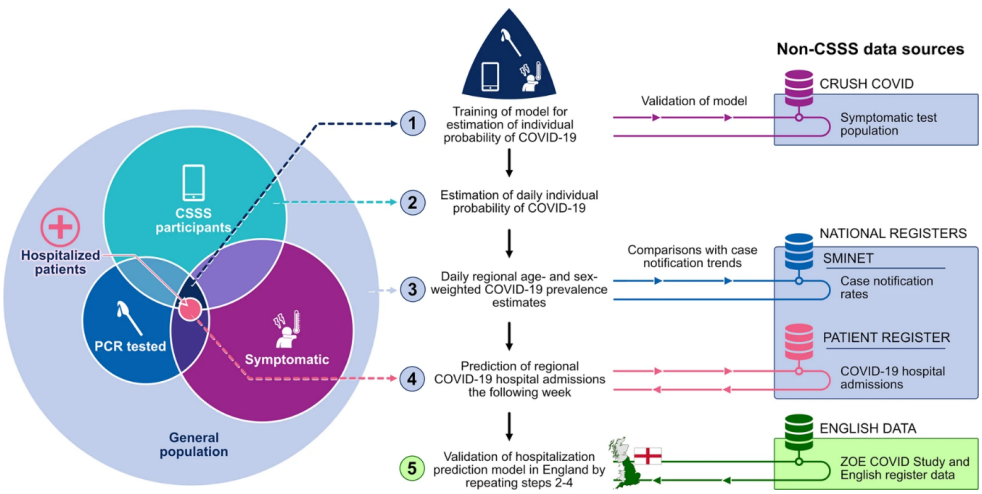


Figure 3.5 Analysis strategy and data sources (paper II).

In the third step, we used the individual probabilities derived in the second step to estimate the prevalence of COVID-19 for each of the 21 regions in Sweden. The model was adjusted to account for differences in the age and sex distribution of the CSSS participants compared to the general population in each region. The resulting estimates of COVID-19 prevalence showed similar waves as the first and second waves of COVID-19 hospitalization. We also observed a peak in the estimated prevalence of COVID-19 in September 2020 based on data from the CSSS app, but this peak was not reflected in other national data on COVID-19 case notification

rates or hospital admissions. To better understand this discrepancy, we developed a retrospective time-dependent model for the probability of having symptomatic COVID-19, which showed higher concordance with national COVID-19 case notification and hospital admission trends than the main model.

In the fourth step, we used the CSSS prevalence estimates to predict COVID-19 hospitalizations in Sweden seven days in advance. We used a weighted linear regression model using data on the CSSS prevalence estimates and current regional rate of COVID-19 hospitalizations. The model was trained and updated iteratively throughout the study period (May 11 to November 29, 2020), during which 16,752 individuals were hospitalized with a diagnosis of COVID-19. The model demonstrated a median absolute percentage error (MdAPE) of 25.9% for the first pandemic wave and 26.8% for the second wave when applied to the five most populated regions in Sweden. The accuracy of the model diminished as regional daily number of hospital admissions dropped. A similar prediction model using daily case notifications from the national Swedish register (SmiNet) had MdAPEs of 30.3% and 25.9% for the first and second waves, respectively. Overall, the CSSS hospital prediction model had higher accuracy than the SmiNet-based model, particularly in the most populous region, Stockholm, where the MdAPE was 12.2% for the first wave and 16.6% for the second wave. When applied to all 21 regions in Sweden, the MdAPEs for the CSSS hospital prediction model were 37.0% and 42.4% for the first and second waves, respectively (Figure 3.6).

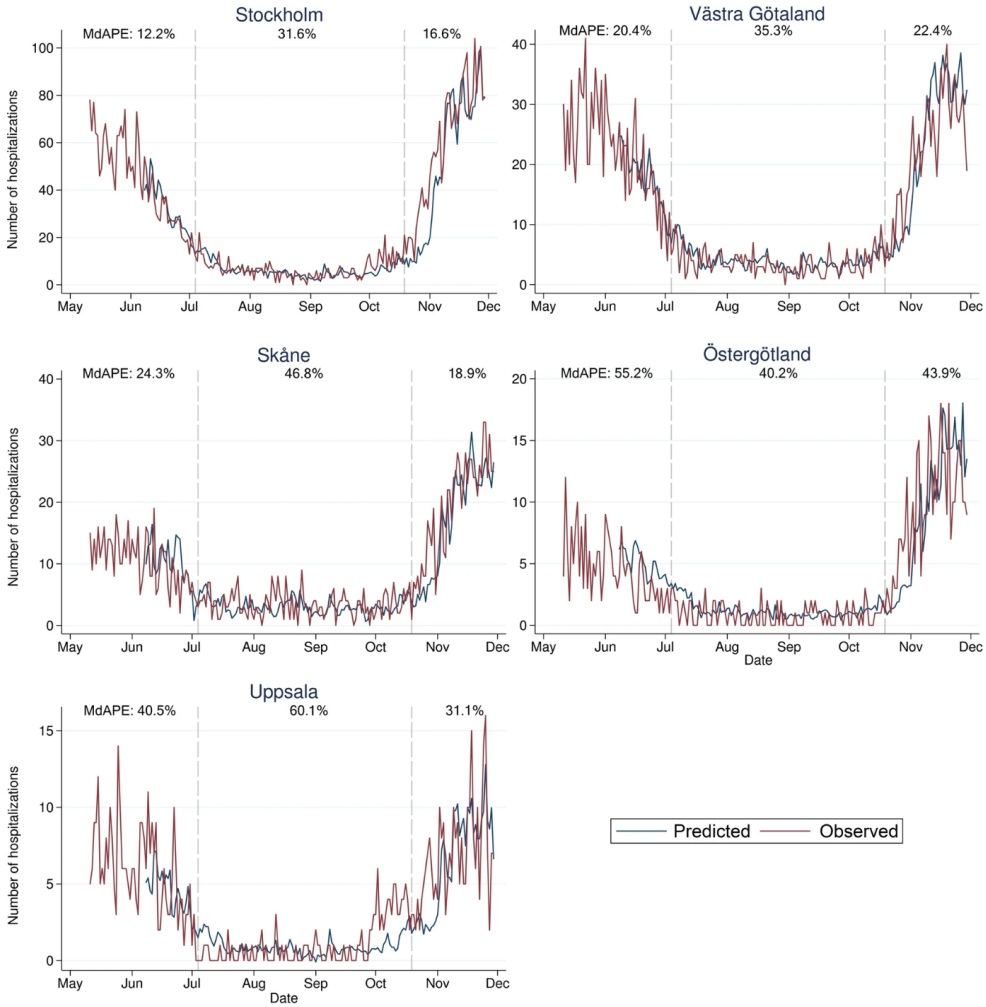


Figure 3.6 Predicted number of daily hospital admissions 7 days ahead across the five most populated regions in Sweden ordered by population size. The median absolute percentage errors (MdAPEs) of the predictions are denoted for the first pandemic wave (June 8–July 3, 2020), the summer period (July 4–October 18, 2020), and the second pandemic wave (October 19–November 29, 2020).

In Step 5, the hospitalization prediction model developed in the first step for Sweden was validated for the UK by repeating Steps 2 and 3 and parts of Step 4 on a dataset of UK participants from the CSS. The model was applied to daily reports from 2,638,536 English study participants from March 30, 2020, to January 31, 2021 and hospital admission data for individuals ≥18 years from April 6, 2020 to February 7, 2021. The model was used to estimate the daily age- and sex-weighted COVID-19 prevalence across the seven English healthcare regions and to predict hospital admissions the following week using an iterative time-

updated prediction model. The model showed an MdAPE of 22.3% for the part of the first English pandemic wave captured in the data (May 4–June 19, 2020) and an MdAPE of 19.0% for the second English wave (September 20, 2020–February 7, 2021). The predicted number of hospital admissions were overestimated when daily regional hospital admissions were low.

Paper III

The third study in this thesis analyzed data from >520,000 participants from the UK, US, and Sweden who reported their PCR test results, vaccination status, and symptoms in the CSS mobile app between May 2020 and November 2021. The aim of the study was to investigate whether the symptom profile of COVID-19 during the acute phase of the infection and the presentation and duration of the more persistent long-COVID symptoms, differed between different subgroups of the population (based on self-reported BMI and diabetes diagnosis) before and after vaccination.

We normalized the data by aligning the date of each participant's PCR test as day 0. This allowed us to compare the symptoms reported by different participants at the same relative time point, 30 days before and 30 days after their PCR test. By re-centering the data around the date of the PCR test, we were able to more accurately understand the progression of symptoms (prevalence) in relation to the onset of the infection in the compared groups (Figure 3.7). We then analyzed these symptom trajectories around PCR test dates (positive and negative) for each BMI category (underweight, normal, overweight, and obese) and diabetes status (non-diabetic, T1D and T2D), both before and after vaccination. Moreover, acute COVID-19 and long-COVID symptoms were organized into symptom domains.

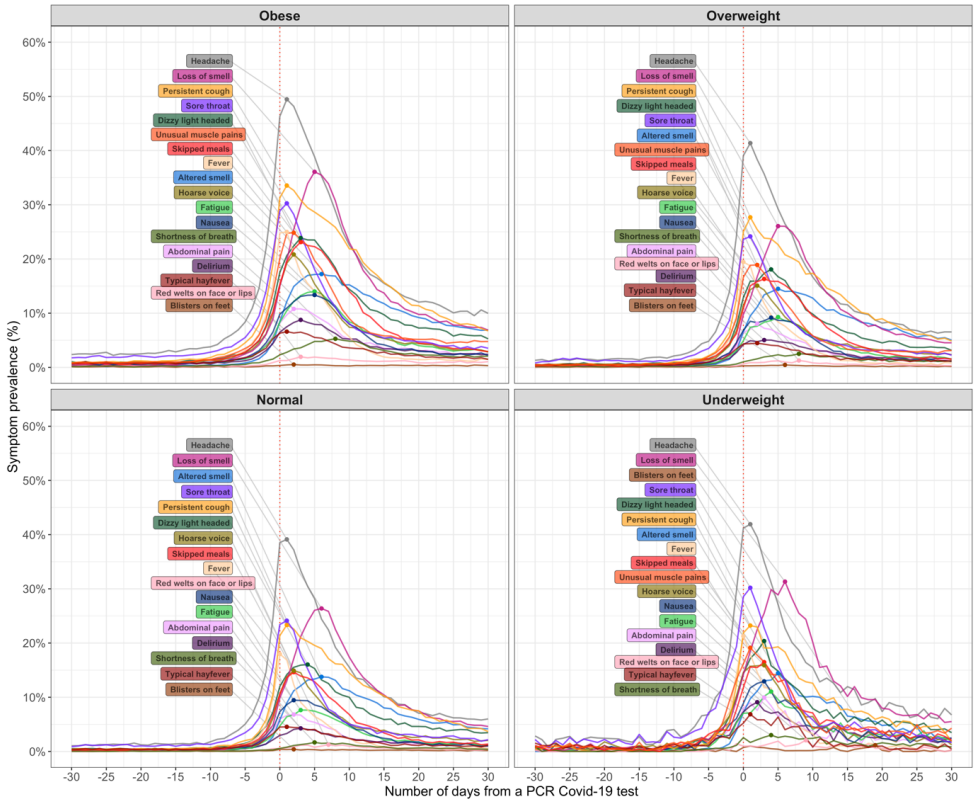


Figure 3.7 Symptoms prevalence around PCR positive test across BMI categories (pre-vaccination).

To minimize the impact of confounding and other types of bias when comparing symptom curves in different categories of BMI and diabetes status, we employed a series of statistical procedures. These included propensity score weighting, which adjusts the statistical weight of each individual based on their susceptibility to potential confounding factors (age and sex in this case). Additionally, we used bootstrapping method with 100 re-sampling iterations to understand the variability and uncertainty of symptom dynamics in the different BMI and diabetes categories, as well as for each vaccination status and PCR test result. We calculated AUC for each of the bootstrapped symptom curves for each category, to obtain a point estimate of the total symptom prevalence for that category with associated uncertainty. Finally, we performed Mann-Whitney U test to check if the difference between symptom curves were statistically significant.

Our results showed that, prior to vaccination, the group with the highest average symptom scores (as measured by AUC) was typically the obese group (BMI > 30 kg/m^2) except for *delirium*, in which the underweight group (BMI < 18.5 kg/m^2)

had the highest AUC of 4.0 (95% CI 3.9 – 4.1). Obese and underweight individuals did not significantly differ in terms of shortness of breath, with both groups showing higher AUCs compared to individuals who were classified as overweight (BMI between 25 kg/m2 and 29.9 kg/m2) or normal weight (BMI between 18.5 kg/m2 and 24.9 kg/m2).

After vaccination, the differences in symptom AUCs between the different BMI groups decreased significantly (Figure 3.8). This is mainly driven by a larger impact of vaccination in the obese group as visualized in the radial plots of Figure 3.8, where the obese group is clearly separated from the other groups in panel A prior to vaccination, and the same obese group in panel B shrinks towards the center of the plot and converges with the other groups after vaccination. For instance, before vaccination, the AUC for fatigue was 4.7 (95% CI 4.7 – 4.7) for the obese group, which was approximately 67% greater than the AUC of 2.9 (95% CI 2.9 – 2.9) for the underweight group. After vaccination, the proportional difference between the groups dropped to half, with the AUCs decreasing to 2.1 (95% CI 2.1 – 2.1) and 1.6 (95% CI 1.6 – 1.6) for the obese and underweight groups, respectively. For headache, symptom that presented the highest AUC across all groups, the AUCs dropped from 16.2 (95% CI 16.2 – 16.3), 11.6 (95% CI 11.6 – 11.7), 9.9 (95% CI 9.9 – 10.0), 9.6 (95% CI 9.5 – 9.7) to 11.1 (95% CI 11.1 – 11.1), 10.3 (95% CI 10.3 – 4), 9.4 (95% CI 9.4 – 9.4) and 6.7 (95% CI 6.6 – 6.7) for the overweight, normal weight, and underweight groups, respectively.

We have also compared the differences in the prevalence of symptoms before and after vaccination for individuals with type 1 diabetes (T1D), type 2 diabetes (T2D), and non-diabetic. Overall, the AUCs decreased after vaccination for all the groups. Shortness of breath was the symptom with the biggest proportional difference in AUC between T1D and non-diabetic groups, with 2.0 (95% CI 1.9 – 2.1) and 0.9 (95% CI 0.9 – 0.9) and also between T2D and non-diabetic, with 2.2 (95% CI 2.2 – 2.2) and 0.9 (95% CI 0.9 – 0.9) respectively. Post-vaccination, these AUCs dropped to 0.6 (95% CI 0.6 – 0.6) and 0.5 (95% CI 0.5 – 0.5) for the T1D and non-diabetic groups; and 1.0 (95% CI 1.0 – 1.0) and 0.5 (95% CI 0.5 – 0.5) for T2D and non-diabetic groups respectively.

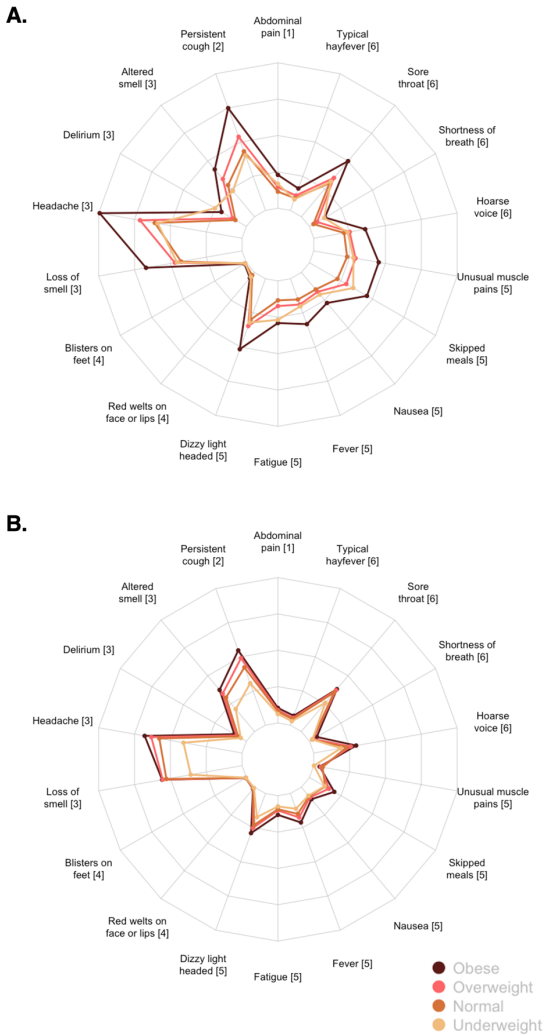


Figure 3.8 Mean area under the curve (AUC) of COVID-19 symptoms in different BMI groups before and after vaccination. Panel A displays the pre-vaccination mean AUCs for COVID-19 symptoms in four BMI groups, represented by different colors. Higher BMI groups generally displayed higher mean AUCs for most symptoms. Panel B displays the post-vaccination mean AUCs for COVID-19 symptoms in the same four BMI groups. In comparison to the pre-vaccination AUCs shown in Panel A, the post-vaccination AUCs for most symptoms are significantly lower. The numbers that follow the symptom labels represents the pre-defined symptom domains: [1] - Abdominal; [2] Cardiorespiratory; [3] Central neurological; [4] Immune related/cutaneous; [5] Systematic/inflammatory and [6] Upper respiratory.

Moreover, we sought to determine whether the prevalence and duration of long-COVID (defined as symptoms lasting >28 days) differs across different categories of BMI and diabetes status. The data was split into vaccinated and unvaccinated sets, with post-vaccination infection defined as a positive test at least two weeks after the participant had completed their vaccination regimen. We used the same propensity score weighting derived for the first part of our analysis to adjust the data on long-COVID and used analysis of variance (ANOVA) and Tukey Honest Significant Difference (HSD) tests to compare the mean duration of symptoms and symptom domains across the different categories of BMI and diabetes before and after vaccination.

Prior to vaccination, there were statistically significant differences in the mean duration of various long-COVID symptoms between at least two of the four BMI groups. Most of these differences were between the obese and normal weight groups. For *shortness of breath*, results showed statistically significant differences in duration (adjusted days) between the obese and normal groups (95% CI 3.86 – 14.61, $p < 0.01$) and between the overweight and normal weight groups (95% CI 0.60 – 10.60, $p < 0.05$). After vaccination, the mean duration of long-COVID symptoms decreased across all BMI groups, but there were still significant differences between the groups for most symptoms, with shortness of breath having significant differences between (95% CI 3.74 – 6.09, $p < 0.01$), obese and overweight (95% CI 0.91 – 3.01, $p < 0.01$) and overweight and normal (95% CI 1.73 – 4.18 , $p < 0.01$) groups (Figure 3.9). Our results also showed that, prior to vaccination, there were statistically significant differences in the mean duration of various long-COVID symptoms between at least two of the three diabetes status groups. The largest difference was for shortness of breath, which had a significantly longer duration in the T2D group compared to the non-diabetic group (95% CI 2.72 – 22.93, $p < 0.01$). After vaccination, the mean duration of long-COVID symptoms decreased for all diabetes status groups.

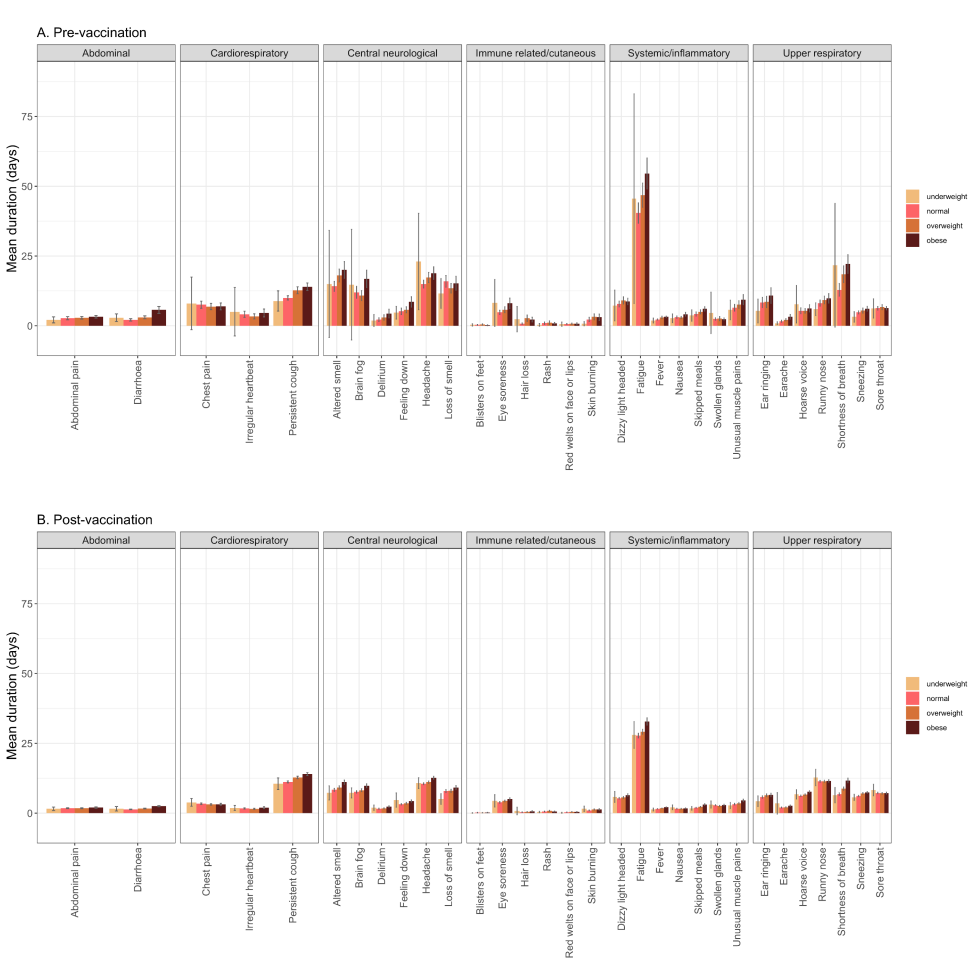


Figure 3.9 Comparison of long-COVID symptom duration between BMI groups before and after vaccination. Panel A shows the mean duration of long-COVID symptoms in four BMI groups (obese, overweight, normal weight, and underweight, represented by different colors) before vaccination. Panel B shows the mean duration of long-COVID symptoms in the same BMI groups after vaccination. Error bars represent the 95% confidence interval for each mean. In both panels, higher BMI groups generally have higher mean symptom duration, particularly the obese group. However, post-vaccination, the mean duration of all symptoms is reduced for all BMI groups. Symptoms are grouped in symptom domains.

COVID Symptom Study Sweden Dashboard

The COVID-19 pandemic has highlighted the importance of well-crafted data visualization in addressing global health challenges. Dashboards, which are

interactive visualizations of data, have played a crucial role in tracking the spread of the virus and informing public health decision-making. Examples of widely used dashboards include the John Hopkins University COVID-19 dashboard (63) and the WHO Coronavirus (COVID-19) dashboard (https://covid19.who.int/).

The development of dashboards has been facilitated by open-source tools such as *R Shiny* (112, 113), a web application framework for the R programming language. *R Shiny* enables users to create interactive, real-time dashboards from data and has proven invaluable in tracking the COVID-19 pandemic and providing transparent, up-to-date information to the public (114-117). *R Shiny* is an R package, making it extremely easy to build web applications for the large R community in bioinformatics and data science, as the usual knowledge required to build such tools (e.g. HTML, CSS - Cascading Style Sheets or JavaScript) is integrated in the R framework (112).

Here, I describe the COVID Symptom Study Sweden Dashboard (referred to as the CSSS Dashboard), a Shiny app that tracks the spread of COVID-19 in Sweden using data from the CSSS. The application, which is developed in R Shiny and hosted on shyniapps.io, is intended for use by health policy makers, the healthcare and scientific community, and the public. The app is designed to be easily accessible and user-friendly, making it a valuable resource for all users.

Overview

CSSS Dashboard was launched in July 2020 as an interactive online solution for presenting CSSS results. The transition from publishing fixed figures on Lund University's website (https://www.covid19app.lu.se/covid-symptom-study-sverige) to using the CSSS Dashboard expedited the dissemination of CSSS results to the public, delivering daily rather than weekly updates. This increased frequency was made possible through the automation of several steps in the data analysis process and the utilization of the R framework already in use by the data analysis team.

The dashboard is in Swedish and is available at: https://csss-resultat.shinyapps.io/csss_dashboard/. All codes and past versions of the dashboard are accessible at: https://github.com/hugofitipaldi/CSSS_dashboard.

Throughout the study period, the CSSS Dashboard was regularly updated to reflect the latest knowledge and insights about the pandemic. As new information became available, the dashboard was modified to include new findings and adjust figures accordingly. In addition, it was also modified based on feedback from users to improve its effectiveness and usability. These updates ensured that the CSSS Dashboard remained a reliable and useful resource for understanding and tracking the impact of the pandemic in Sweden.

In July 2022, the CSSS completed the data collection phase, and the CSSS Dashboard stopped its automatic updates. The dashboard now serves as a data archive, providing a historical record of CSSS predictions. Despite no longer being actively updated, the Dashboard remains a valuable resource for researchers, policymakers, and the public, as it provides a comprehensive and detailed overview of the pandemic and its effects in Sweden.

CSSS model overview

As described previously in this thesis, we used an L1-penalized logistic regression model (LASSO) to predict the individual probability of having a symptomatic COVID-19 infection. The final model chosen by LASSO included 17 symptoms, sex, and two-way interactions between loss of smell and/or taste and 14 symptoms, as well as a two-way interaction between loss of smell and/or taste and sex. The model was used to estimate the daily probability of having symptomatic COVID-19 for all participants in the CSSS study, including those who were not tested. We estimated the daily prevalence of COVID-19 in different Swedish regions (counties and 2-digit postal-code areas) by taking into account the individual probabilities of having the disease and adjusting for the differences in the age and gender compositions of the study participants compared to the general population in each region.

Dashboard layout

Using the third package *shinydashboard* (118), we created a classic dashboard layout, with a main body (central page) and a lateral sidebar navigation menu, in which subtabs can be accessed. In the following sections, the main CSSS Dashboard's subtabs are described.

1. "Om COVID Symptom Study" (About CSS)

The landing page of the CSSS Dashboard presents general information about the Dashboard, as well as any pertinent information about the data displayed on the page. Moreover, this page also includes information boxes with data on the total number of study participants in Sweden, the number of active participants in Sweden (those who utilized the app at least once in the past seven days), the latest prediction of the prevalence of COVID-19 in Sweden with confidence intervals, and the total number of app accesses by Swedish participants.

The Dashboard's most recent version includes a pie chart, created using the *Plotly* R package (113), a data visualization library that allows users to create interactive, publication-quality plots and charts. This pie chart shows the proportion of actively participating users who have received COVID-19 vaccination (Figure 3.10). Interactive actions for the pie chart include hovering over the segments to display

more information, toggling the display of different data categories using the legend and easy download of the chart as PNG.

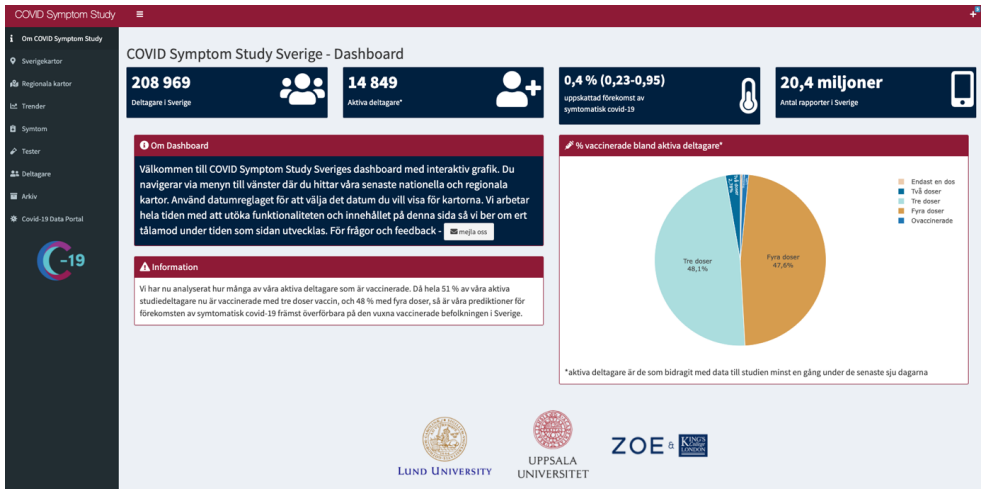


Figure 3.10 CSSS dashboard landing page.

2. “Sverigekartor & Regionala kartor” (National and Regional maps)

These pages showcase the COVID-19 prevalence predictions for Sweden’s 21 regions (counties) and 83 2-digit level postal-code regions, respectively, in interactive choropleth maps and interactive tables (Figure 3.11). Predictions are shown to regions that have at least 200 hundred active participants reporting.

The interactive maps were designed using *leaflet* library (119), a widely used open-source JavaScript library for creating interactive maps on the web. To plot the maps, we used shapefiles from Statistics Sweden (statistiska centralbyrån, SCB) (120) and *Postnummerservice* Norden AB (121), which contain vector-based geographic data. Interactive actions for the maps include zooming, panning, and hovering over regions to display more information.

Tables were created using the *DT* package in R (122), which provides functions for generating HTML tables with the jQuery *DataTables* library. *DT* is an R wrapper package that simplifies the use of jQuery DataTables from within R. The tables can be sorted, filtered, and searched, and users can select which columns to display.

Users can also access past maps and tables through an interactive date slider widget.

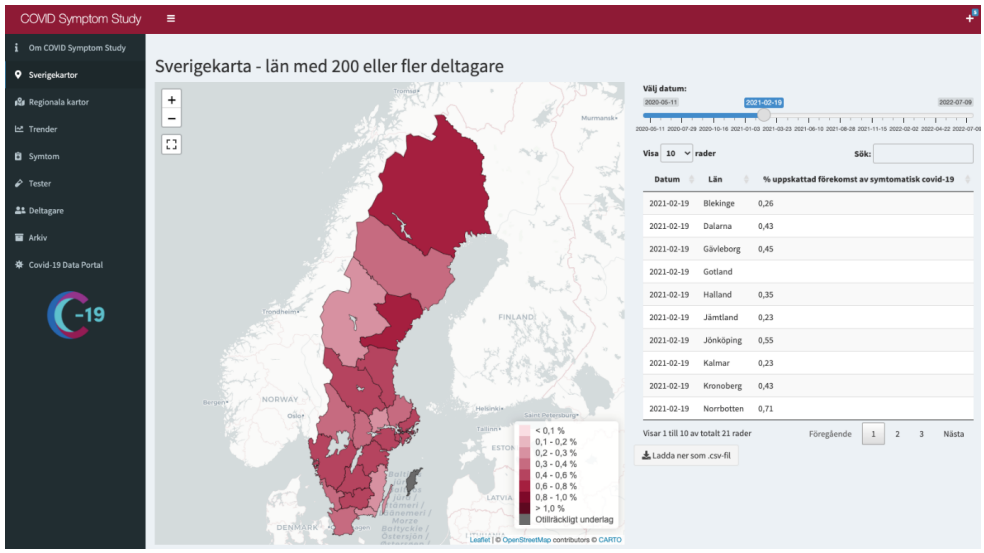


Figure 3.11 CSSS interactive map for 21 regions of Sweden.

3. “Trender” (Trends)

This page features a dual y-axis interactive scatterplot that displays the estimated prevalence of COVID-19 symptoms (with confidence intervals) for Sweden and the number of active participants over the course of the study, as well as an interactive web tool that enables users to compare and plot the estimated prevalence of COVID-19 symptoms (with confidence intervals) for any of Sweden's 21 counties (Figure 3.12).

The dual y-axis plot was created using the *Plotly* R package (113). Interactive features for the dual y-axis plot include hovering over the data points to display more information, zooming, and isolating trends for specific time periods. The legend can be used to toggle the display of different data categories. A function in the plot also allows users to easily download it as a PNG file.

To plot county trends, we utilize the JavaScript library *selectize.js* as interface, which allows the user to type and search in the pre-defined options (Swedish counties) as well as to control the number of options/items to show/select. The selected county is automatically plotted using the *ggplot2* library (123), a widely utilized data visualization tool in the R programming language known for its capacity to produce high-quality, visually appealing plots and charts.

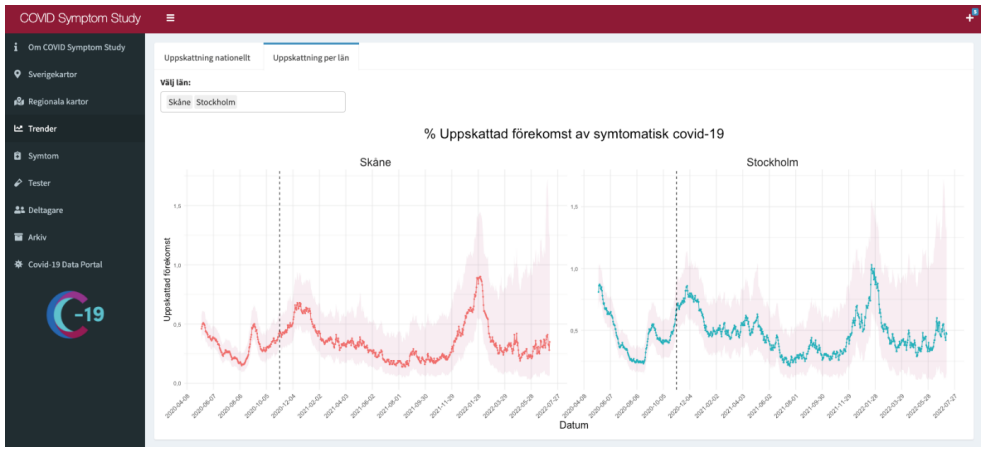


Figure 3.12 CSSS dashboard trends page.

4. “Symtom” (Symptoms)

This multi-tab page presents two scatterplots depicting the prevalence of symptoms that were identified as positively or negatively correlated with a COVID-19 infection by our model (Figure 3.13). The page also includes a bar plot displaying the weight of each symptom in the prediction model. All plots in this page were created using *Plotly* (113). Interactive features for the scatterplots and bar plot include hovering over the data points and bars to display more information, zooming, and the ability to isolate and examine specific data categories using the legend.

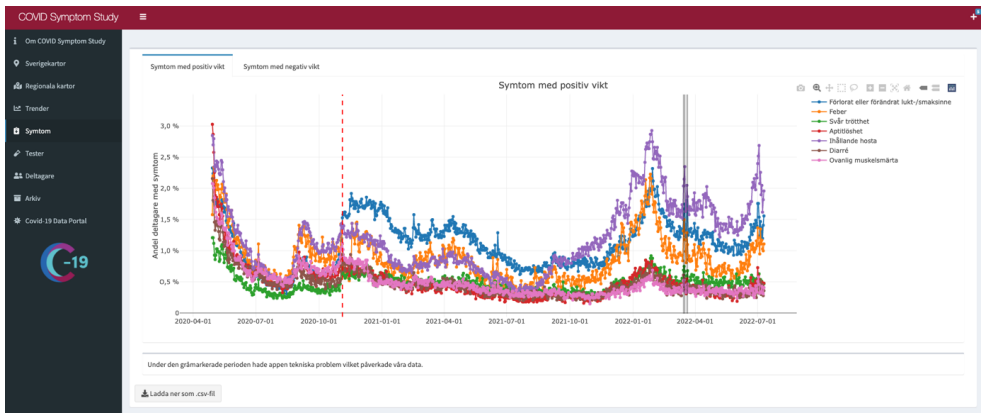


Figure 3.13 CSSS dashboard symptoms page.

5. “Tester” (Tests)

In January 2022, FOHM, the Public Health Agency of Sweden, announced changes to the recommendations for testing for suspected COVID-19 in Sweden. As there

were no national statistics available on antigen tests (rapid tests) taken at home or at private test centers, it would have been difficult to interpret trends in the infection situation based on test statistics from that period forward. To provide additional insights into COVID-19 test results, this multi-tab page was introduced to the dashboard (Figure 3.14).

This page includes two interactive scatterplots with dual y-axes, accessible through a subtab switch button on the top of the page. The scatterplots display the total number of tests and the percentage of positive results over the study period for both antigen and PCR tests. Both plots were built using the *Ploty* package (113). Interactive actions for the scatterplots include hovering over the data points to display more information, zooming, and the ability to isolate and examine specific data categories or group of points using the legend or the panel.

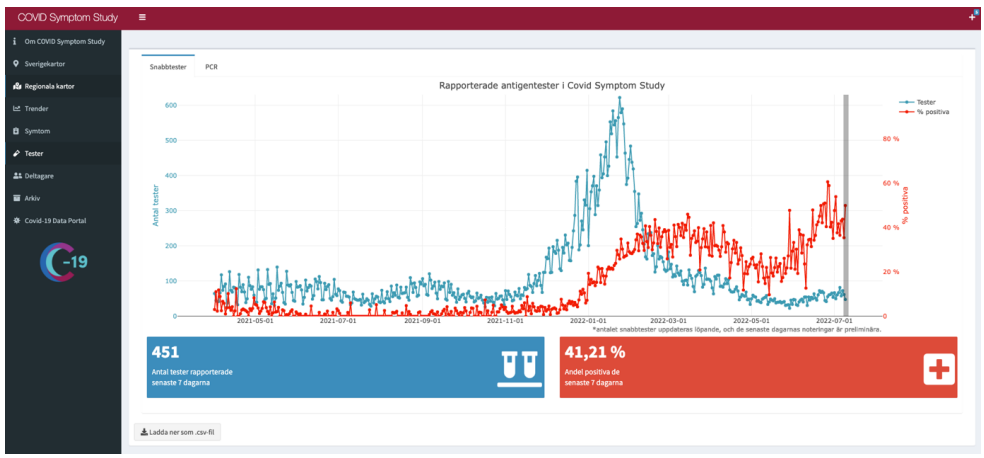


Figure 3.14 CSSS dashboard tests page.

6. “Deltagare” (Participants)

This page presents some demographic information on the CSSS population through two interactive plots: a bar chart and a pie chart (Figure 3.15). The bar chart displays the age and sex distribution of participants, while the pie chart illustrates the proportion of male and female participants. The plots were created using the *Plotly* R package (113), with interactive actions that include hovering over the data points and segments to display more information, zooming, and the ability to toggle the display of different data categories using the legend or the plot panel. Plots can also be downloaded as PNG.

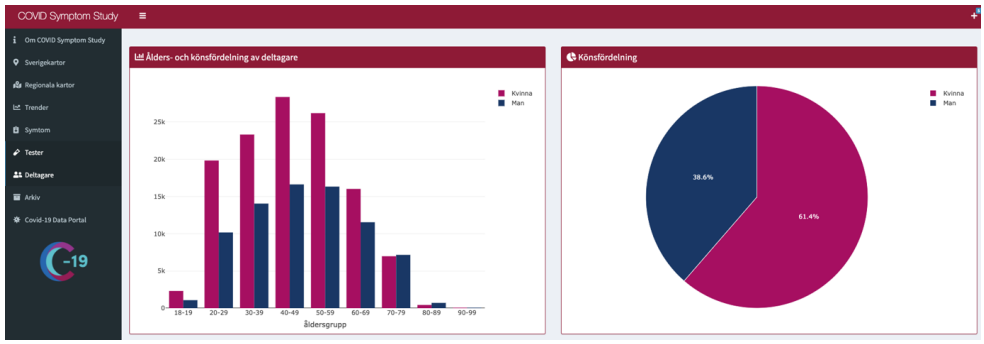


Figure 3.15 CSSS dashboard participants' demographics page.

7. "Arkiv" (Archive)

This page offers downloads of all the CSSS predictions (national, regional, and 2-digit) used in the CSSS dashboard. Additionally, we present the *covidsymptom* R package (109), which allows users to access CSSS' latest predictions directly within the R environment.

Deployment

The process of collecting data through the CSS mobile app and publishing these data through the Dashboard involved several intricate steps (Figure 3.16). The data collected by the app was stored and managed by the health data science company ZOE Ltd. Data related to Swedish participants was transferred daily to the servers of the Lund University Diabetes Center (LUDC). Automated R scripts were implemented to make daily predictions and aggregations with the new data, and the CSSS was then automatically updated and meticulously reviewed by an analyst before publication. This ensured that the dashboard was consistently updated with the most current and accurate data. The CSSS dashboard was hosted on shinyapps.io, a platform developed by the RStudio team specifically for hosting and deploying Shiny applications.

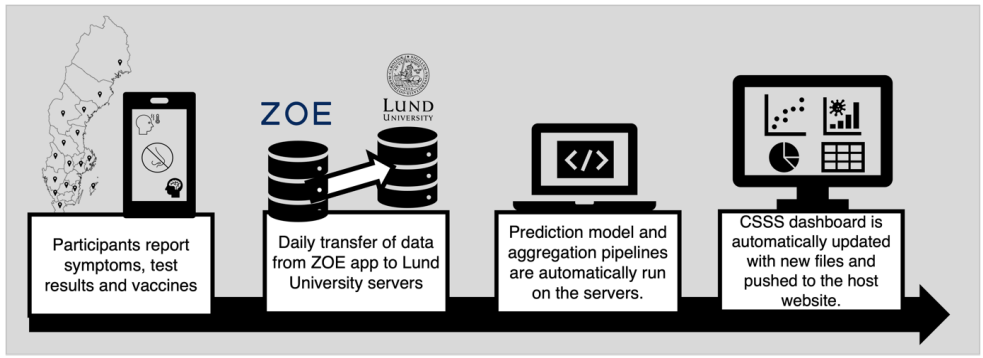


Figure 3.16 CSSS dashboard deployment pathway.

covidsymptom R package

In addition to the dashboard, we developed an R package called *covidsymptom* (109). The package was launched in early 2021 and designed to facilitate the import of open data from the CSSS. It was written entirely in R and published on the Comprehensive R Archive Network (CRAN), a repository for open-source R packages. The package contains the main aggregated datasets used in the CSSS dashboard, such as daily estimates of the incidence of symptomatic COVID-19 in Sweden and its regions, as well as daily prevalences of COVID-19 tests and symptoms reported by CSSS users. The package is a useful tool for researchers and analysts who want to analyze and visualize the data collected by the CSSS. As of December 2022, the package had around 8000 downloads on CRAN. The code for the package and examples of its use can be found at: https://github.com/csss-resultat/covidsymptom

Chapter IV – Discussion

The present thesis explored the utilization of data mining- and AI-based tools as means of managing extensive health datasets and generating evidence for public health interventions. In **paper I**, we developed a framework for text mining and NLP to process and analyze metadata from research publications, using it to effectively mine all the GWAS literature in NCDs from 2005-2022. The outcomes of this analysis unveiled gender, economic, and geographical biases among the authors of the research papers, as well as in the ethnicity of the study participants used. In **papers II** and **III**, we introduce the CSS, a large epidemiological initiative that utilized a mobile app to assist in tracking and understanding COVID-19 in the UK, US and Sweden. Using ML-based algorithms, we were able to predict symptomatic COVID-19 in Swedish participants based on self-reported symptoms and predict hospitalization waves seven days in advance in the major regions of Sweden. Furthermore, we mined and pre-processed CSS data obtained from the three countries, demonstrating that the symptom profile of COVID-19 and the duration of long COVID varied among different subgroups, revealing that participants with higher BMI and/or diabetes had a high prevalence of symptoms during a COVID-19 infection and a longer duration of long COVID symptoms. We also demonstrated that vaccination mitigated these differences between groups. Finally, I also introduced the **CSSS dashboard**, a web-based application designed to disseminate CSSS results, facilitating public health decision making and updating the general public in Sweden about the most recent CSSS findings.

Paper I

GWAS has played a significant role in advancing our understanding of the biological basis of NCDs over the past two decades. In **paper I**, we analyzed 2300 published GWAS papers on NCDs using text mining and NLP to describe the characteristics of the analyzed cohorts and the researchers who conducted the work.

The results of this analysis indicated that more than 90% of GWAS data originated from individuals of European ancestry, particularly from Nordic countries and the UK, while approximately 5% of GWAS analyses were conducted on individuals of East Asian ancestry. Other ethnicities, such as African and Afro-Caribbean populations, are infrequently included in GWAS of NCDs. Similarly, previous

studies (51-54) noted the lack of diversity in GWAS publications and have emphasized the importance of addressing this issue in order to achieve a more comprehensive understanding of the genetic basis of diseases and to promote equity in the field of genetics research.

The use of automated text mining analysis tools used in this study enabled us to quantify this issue in a large scale. The lack of diversity in GWAS data has implications for the generalizability of genetic discoveries to global populations and hinder the development of tools such as polygenic risk scores (PRS) that can be used to predict disease risk and inform precision healthcare approaches (35). For example, a study (124) that evaluated the performance of seven breast cancer PRSs in three distinct ancestral groups (European, African, and Latinx) found that the scores derived in women of European ancestry did not generalize well to African ancestry cohorts. Furthermore, other studies have reported that predictions of PRS in African ancestry populations are only ~20–40% as accurate as in European populations when using European-based GWAS (50, 125).

In terms of authorship, researchers based at institutions in HICs, particularly the US, contributed the largest proportion of co-authorships across GWAS papers for most disease areas, comprising an average of around 40% of authorships. Among the top-ranking institutions, Harvard Medical School was one of the most frequent contributors. However, there were some exceptions to the dominance of US institutions, such as the Wellcome Trust Centre for Human Genetics in Oxford (UK) and Karolinska Institutet (Sweden).

US institutions are known for their high levels of productivity in various fields of science (126, 127), but these analyses often only consider the affiliations of first or senior authors. By including all authors in the authorship lists, the text mining and NER-based framework presented in **paper I** of this thesis allows for a more comprehensive understanding of scientific networks and improves upon the traditional method of analysis. In GWAS research, the US prolificacy can be attributed to its extensive capacity to generate and analyze genetic data from within and outside its borders, driven by international collaborative networks and well-funded national programs (128). Our results also shown that there has been a slight decline in US dominance in recent years, while China, South Korea, and other countries are gradually increasing their leadership in GWAS research.

Moreover, the analysis in this study revealed that the origin of the data utilized in these GWAS did not consistently align with authorship of the publications. As one example, in the case of GWAS research on chronic respiratory disease, despite more than half of participants originating from the UK, UK-based scientists only occupied 7% of first and senior authorships, while US-based scientists disproportionately contributed nearly half of these co-author positions. This type of disparity has previously been documented (129), in which it was shown that a disproportionate

amount of genomic research conducted in Africa has been conducted and led by researchers from Europe and the US, rather than African investigators.

Undoubtedly, this issue needs further investigation, particularly when it involves non-HICs, as it raises concerns about the potential for exploitation of local populations and resources, as well as the unequal distribution of benefits from the research. It is important to examine the root causes of this discrepancy and to identify ways to address it, such as through increased collaboration and capacity-building efforts within these countries, so that the benefits of GWAS research are more evenly distributed and to address the most pressing health needs of the region. The framework implemented in this study, which includes a R package (107) for the analysis of authorship list of research papers published in PubMed, can be of use for unveiling more of the issues in different fields of health research.

Furthermore, gender disparities in co-authorship of GWAS papers were identified in this analysis, with males consistently comprising a disproportionate number of authorships, particularly in senior positions. Across all disease areas, male researchers accounted for an average of about 60% of first authorships and 75% of senior authorships. While the gender balance of first authorships varied by disease area, senior authorships consistently favored male researchers, with the most balanced being musculoskeletal disorders (65% male) and the least balanced being chronic respiratory diseases and digestive diseases (both with 79%). While the gender gap in co-authorship has narrowed over the 16-year study period, with females comprising an increasing proportion of first and senior authorships, it remains significant and reflective of the overall representation of women in research. These findings are consistent with broader trends in the scientific community, where male researchers continue to dominate leadership roles and female researchers remain underrepresented, particularly in fields such as STEM (130-132). This gender imbalance can have various negative consequences, including hindering the development of diverse perspectives and approaches to scientific inquiry and potentially limiting the impact and generalizability of research findings. Therefore, it is important for the scientific community to address these imbalances and promote gender equality in all aspects of research, including authorship of GWAS papers.

In summary, our analysis revealed that GWAS research in NCDs has been dominated by male researchers based at institutions in HICs, particularly the US, while the data used in these studies has been largely derived from individuals of European ancestry, with a significant proportion coming from Nordic countries. A limitation the analysis conducted in this paper was that it only contained publications written in English, as the primary source of data was the GWAS Catalog. Additionally, the method used to predict author gender, name-to-gender

inference, has limitations as it is not a perfect proxy for self-reported gender, although it has been validated before (98).

Our study also demonstrates the utility of data mining and AI-based techniques in conducting large-scale, comprehensive analyses of complex data. The use of text mining and NLP tools, including custom R packages that I developed (107, 108), were essential to extract relevant information from the GWAS Catalog and PubMed and to identify trends and patterns in the data. These computational methods also enabled us to perform analyses that would have been infeasible without their assistance, such as extracting and pre-processing co-authorship lists and linking authors' names to their affiliations.

Paper II

Effective regional COVID-19 surveillance has been a challenge since the onset of the pandemic, requiring multiple sources of data to address this challenge (74). The spread of the SARS-CoV-2 virus and the impact on the healthcare system can vary significantly within a country or region, and real-time, accurate data is essential for informed decision making and preparedness efforts. Conventional surveillance methods, such as laboratory-based testing, may not be sufficient to capture the complete extent of the outbreak, and alternative approaches, such as participatory syndromic surveillance, whereby individuals self-report their symptoms, may be useful in complementing these efforts (20-22). In **paper II**, we developed and evaluated an app-based framework for syndromic surveillance of COVID-19 at the national and regional level in Sweden during the first year of the pandemic.

The results of our analysis revealed that the CSSS prevalence estimates of symptomatic COVID-19 displayed wave patterns comparable to those observed during the first and second waves of COVID-19 hospitalization in Sweden. These findings are consistent with the results of a prior studies conducted in England (21, 133) and the US (134), which demonstrated that self-reported data from users could effectively identify emerging COVID-19 hotpots.

However, our results also highlight the importance of accounting for other respiratory infections in the interpretation of app-based surveillance data. The resulting prevalence of our main model showed a peak in CSSS-based COVID-19 prevalence estimates in Sweden for September 2020 that was not reflected in national case notification or hospital admission data, likely due to the concurrent peak in other respiratory infections. To address this issue, we developed a retrospective time-dependent model that showed better concordance with national COVID-19 case notification and hospital admission trends.

In addition to monitoring disease trends, our app-based framework for syndromic surveillance demonstrated the potential for predicting regional levels of COVID-19

hospital admissions with a moderate level of accuracy. This ability to forecast future demand for hospital resources is crucial in times of crisis, as it allows for the proactive allocation of healthcare resources. Importantly, our results show that the CSSS-based hospital prediction model is transferable to a different setting, as we were able to successfully validate the model using data from England. However, the accuracy of the model varied by region, with higher accuracy observed in more highly populated areas. This suggests that factors such as total population size and the number of study participants may affect the accuracy of the model.

In summary, the current study demonstrates the utility of app-based COVID-19 syndromic surveillance for monitoring disease trends and predicting hospital admissions at a regional level. The CSSS app was able to provide accurate and timely estimates of COVID-19 prevalence and hospital admissions in Sweden, and showed potential for transferability to other settings. This approach has several advantages, including the rapid collection and analysis of data, and the ability to continuously disseminate results to the public.

Limitations of this study include reliance on self-reported data, which may be subject to bias and may not accurately reflect the true prevalence of COVID-19. There is also the possibility that some individuals with COVID-19 may not have reported symptoms or may have been asymptomatic, leading to an underestimation of the prevalence of COVID-19. Additionally, the study was limited to data from app users, which may have resulted in selection bias and limited the generalizability of the findings to the broader population. Finally, unmeasured confounders may have influenced the results.

To analyze the data collected through the CSSS app we employed data mining and ML techniques, such as a LASSO model. This was used to identify key symptoms and interactions that were associated with an increased probability of having COVID-19 and later estimate daily regional prevalence of the disease in Sweden. We also developed a prediction model that utilized these regional prevalence estimates and hospital data to forecast COVID-19 hospital admissions in Sweden and England. These approaches demonstrate the usefulness of data mining and ML techniques in COVID-19 research and surveillance. Furthermore, they may be refined and developed further to improve prediction accuracy and adapt to changing circumstances.

Paper III

The COVID-19 pandemic has brought to light the significant impact that underlying health conditions can have on the severity of the disease. Prior to the pandemic, it was already well-established that individuals with certain underlying health conditions, such as obesity and diabetes, were at an increased risk of developing

severe illness or complications from infectious diseases (135). However, the magnitude and global impact of the COVID-19 pandemic has highlighted the critical need for further research in this area. Understanding how underlying health conditions may affect the symptomatology and outcomes of COVID-19 is crucial for developing targeted interventions and guidelines for individuals at higher risk. In the **third paper** of this thesis, using data from the CSS, we investigated the differences in the symptomatology of both acute COVID-19 and long-COVID across different categories of BMI and diabetes before and after vaccination.

Our results showed that individuals with higher BMIs had a higher prevalence of symptoms around the time of COVID-19 infection. These results align with the growing body of evidence indicating that obesity may play a role in exacerbating the severity and outcomes of COVID-19, including higher rates of hospitalization and mortality (135-137). We have also found that obese individuals experienced longer duration of long-COVID symptoms. Despite limited research on the link between obesity and long COVID, there is increase evidence suggesting that the severity of COVID-19 infection is a predictor of prolonged manifestations of the disease (79, 81). Notably, it is estimated that long COVID manifestations affect between 50-70% of hospitalized cases of COVID-19 (27).

Moreover, our findings suggest that individuals with diabetes, particularly those with T2D, may be at increased risk for severe acute symptoms and prolonged symptom duration in the long-COVID. These findings are consistent with previous research indicating that individuals with diabetes may be more susceptible to severe illness and complications from COVID-19, including higher rates of hospitalization and mortality (70, 138, 139). Furthermore, T2D have been also previously showed to be both a risk factor and new onset condition for long-COVID (140). The mechanisms underlying this increased susceptibility for severe manifestations of COVID-19 and long-COVID in this population are yet to be determined, but may be related to the underlying immune dysfunction and altered blood flow associated with diabetes (141).

By comparing the resulting symptoms' AUCs pre- and post-vaccination across the BMI and diabetes categories, our study provided evidence for the effectiveness of COVID-19 vaccines for these at-risk populations. Although the AUCs for the COVID-19 symptoms decreased across all BMI and diabetes categories, the obese and the T2D were the groups that benefitted the most from vaccination in each separate analysis. Vaccination impacted positively all groups, reducing the prevalence of symptoms of a COVID-19 infection, and reducing the duration of long-COVID symptoms. These findings are in line with previous research indicating that COVID-19 vaccines provide high levels of protection against severe illness and death in at-risk populations, including those with underlying medical conditions such as obesity and diabetes (137). Given the increased risk for severe illness and

complications in these groups, it is crucial that individuals with obesity and diabetes are prioritized for vaccination.

Overall, the results of this study highlight the importance of considering the influence of underlying medical conditions such as obesity and diabetes on the presentation and course and consequences of SARS-CoV-2 infection. These analyses also add to the growing body of evidence highlighting the effectiveness of vaccination in reducing the burden of disease in these at-risk populations.

The integration of data mining and statistical techniques in this study highlight the potential of these tools in understanding and addressing public health crises such as the COVID-19 pandemic. These techniques were vital in enabling us to uncover patterns and trends in symptom prevalence and duration across various population subgroups, as well as to quantitatively assess the impact of vaccination on these outcomes. The future application of these tools in the realm of public health has the potential to greatly enhance our ability to effectively respond to and mitigate the consequences of future infectious disease outbreaks.

CSSS dashboard

During the COVID-19 pandemic, access to transparent and easily understandable data has been crucial for informing both public health policy makers and the public (63, 73-75). Interactive dashboards that visualize data have played a vital role in this regard by providing real-time information on the spread of the virus. Elegantly crafted visualization tools such as those developed using open-source tools like *R Shiny* (112, 113) have empower end-users to comprehend and assimilate intricate information and unearth new correlations without necessitating programming expertise or exhaustive data-analytic capabilities. These tools have proven valuable for tracking the COVID-19 pandemic and informing effective responses (114, 115, 117).

The **CSSS dashboard** was developed with the aim to provide a centralized, user-friendly report and open data sharing of CSSS results. Since its creation, the dashboard has received thousands of page views from users in Sweden. Its rapid development is the results of the rapid evolution of programming languages and frameworks that have made it easier to create feature-rich applications.

The utilization of dashboards as a public health surveillance tool in times of crisis has been undoubtedly demonstrated to be efficacious, as illustrated by the proliferation of COVID-19 dashboards developed by a range of organizations globally, including governments, media outlets, private enterprises, and academic institutions. The John Hopkins University (JHU) COVID-19 dashboard (63) and the WHO Coronavirus (COVID-19) Dashboard (66) have been widely used by researchers, the public and policy makers as sources of real-time data on global

COVID-19 cases. In Sweden, FOHM, the public health agency, has also created its own data dashboard on COVID-19, which provides updates on the pandemic situation in the country (142). However, during the first years of the pandemic, FOHM only updated the dashboard a few times a week, with data lagging a few days behind. CSSS dashboard, on the other hand, provided daily updates, with data from participants that reported one day before, thus serving as a reliable source for COVID-19 estimations in Sweden when the official government data was not available.

Additionally, within the CSSS dashboard we have also created the *covidsymptom* R package (109), which allows analysts to access CSSS aggregated data and predictions directly within their programming environment. Available on CRAN, the main R package distributor in the world, *covidsymptom* was downloaded by approximately 8000 users worldwide.

The development of such tools for disseminating research results has been vital during the COVID-19 pandemic, as it facilitates the rapid sharing of data and analytical tools among researchers and analysts. This not only helps to expedite the research process, but it also ensures that the data and tools being used are open and transparent, which is essential for reproducible research. Furthermore, by making these tools accessible on platforms such as CRAN and shinyapps.io, it allows for widespread use by researchers and analysts globally. The creation of these types of tools not only serves the current pandemic, but also has long-term implications for pandemic preparedness. By creating open-source projects and repositories of data and analytical tools, it allows for easier and quicker access to these resources in the event of future pandemics, speeding up future research processes.

Overall summary and conclusions

The objective of this thesis was to investigate the application of data mining and AI-based pipelines to extract insights from large health datasets and generate public health evidence. By building frameworks based on these tools, I demonstrated how the analysis of large datasets can be optimized, and the potential for automating and simplifying the data mining process.

In **paper I**, by using advanced data mining methods and creating frameworks to extract and analyze information from 2300 GWAS papers, it was possible to characterize the last two decades of genomic research in NCDs and identify biases in terms of the populations studied and the institutions and researchers involved. Specifically, we found that these studies have been predominantly undertaken by male researchers affiliated to institutions in HIC, particularly the US, and have largely utilized data from individuals of European ancestry. The insufficient diversity in both the data and the authorship of GWAS research has potential

implications for the generalizability of genetic discoveries and the development of future interventions. For instance, the lack of representation of certain populations in these studies may lead to an incomplete understanding of the genetic basis of diseases and how they affect different populations, potentially limiting the effectiveness of interventions and policies aimed at addressing these diseases in diverse populations. It is important to address these biases to produce more robust and generalizable public health evidence that can be effectively applied to improve the health of all populations.

In **paper II**, an app-based framework for syndromic surveillance was developed to track the early spread of COVID-19 in Sweden. In this study, data mining and ML techniques were used to analyze a large amount of data collected through the COVID Symptom Study Sweden (CSSS) app and develop a model to estimate the probability of having symptomatic COVID-19. This model was then used to estimate daily regional prevalence and predict hospital admissions due to the disease. Real-time and accurate disease surveillance data is essential for effective public health decision-making and evaluation, as well as for preparing healthcare systems. In times of global health crisis, the integration of data from multiple sources, including participatory syndromic surveillance, is vital for early detection of disease outbreaks. The results of this study may enhance our understanding of the feasibility of large-scale syndromic surveillance and inform the creation of population-based participatory surveillance initiatives in future pandemics and epidemics.

In **paper III,** the impact of BMI and diabetes on COVID-19 symptom presentation and the development of long-COVID in this at-risk populations were investigated. The advanced data mining techniques implemented in this study were able to process and analyze the large amount of data collected by the COVID Symptom Study (CSS) and reveal that individuals with higher BMIs and diabetes had a higher burden of symptoms at the time of COVID-19 infection and prolonged symptom duration in long-COVID manifestations. Moreover, the results provided evidence for the effectiveness of COVID-19 vaccines for at-risk populations, with the obese and those with T2D experiencing the greatest benefits in terms of reduced symptom prevalence and long-COVID symptom duration. The findings of this study have significant public health implications as they demonstrate the impact of COVID-19 vaccines in at-risk populations, in terms of both COVID-19 and long-COVID presentations. This information can aid in the development of targeted vaccination strategies and prioritization, particularly in populations that are at elevated risk for severe illness and morbidity from COVID-19.

Future perspectives

As technology and research tools continue to advance, we can expect to see an increase in the volume and variety of healthcare data being collected. To effectively harness this data, it will be essential to utilize data mining and AI-based tools, as well as any new methods of automating the process of analyzing and extracting insights. By making these tools openly available to researchers and analysts, collaboration can be facilitated, potentially hastening the pace of innovation in the medical field. Open-source tools in the form of programming packages or dashboards have the potential to democratize the process of analyzing healthcare data, making it more inclusive and accessible to researchers from diverse backgrounds and institutions, thereby mitigating the potential for biases in the research-implementation chain.

The work initiated during my PhD that culminated in the papers and projects presented in this thesis has potential for further exploration and extension of these ideas. Using the developed framework, we can continue to track and address biases in GWAS research for NCDs, including biases related to the geographic and gender distribution of authors and the lack of ethnic diversity in research samples.

Additionally, the data generated by the framework implemented in **paper I** can be further explored with other advanced data analysis tools. Applying network analysis and other NLP techniques to the collected data or undertaking similar data mining analyses of published papers could potentially provide a more comprehensive understanding of patterns and trends within research communities. Network analysis, a technique that can be used to analyze the relationships between different entities in a network, could be applied to co-authorship patterns to identify clusters of collaboration or isolation within a field. By visualizing the network of relationships within this field, it would be possible to gain a more intuitive understanding of the structure and dynamics of the field and could identify key nodes (such as countries or institutions) that play a particularly influential role. Network analysis could be particularly useful for examining patterns of collaboration and the flow of knowledge and resources within the specific field of research. Additionally, other NLP techniques such as text classification algorithms, part-of-speech tagging, dependency parsing, and topic modeling could also be useful for extracting and analyzing information from the full text of papers.

Moreover, the R packages developed in this study can be utilized to investigate similar research questions in various fields of medical science, as they have the capability to analyze any publication indexed in PubMed. The packages are open-source and I plan to continue improving these tools to keep them current with the latest developments in the field of text mining and NLP.

COVID Symtpom Study Sweden (CSSS) is one of the largest epidemiologic initiatives ever done in Sweden. Since its launch in April 2020, it has received over 20 million daily reports from approximately 208 000 individuals in Sweden through the COVID Symptom Study (CSS) app. The results of **paper II** showed the potential of app-based syndromic surveillance to monitor the biggest public health crisis of our time, thus highlighting the potential of such frameworks for pandemic preparedness. The rapid development and implementation of this data collection and reporting structure, facilitated by advances in data mining and AI, is unprecedented and can serve as a model for future epidemiological surveillance efforts.

Moreover, the results from CSSS demonstrated the value of app-based surveillance in countries with well-functioning healthcare system and widespread access to PCR testing. While further research is needed to confirm this, app-based surveillance may be even more crucial in countries where access to healthcare resources is more limited, as it can provide timely and accurate information on the spread of COVID-19.

The results of our latest study, **paper III**, highlight the significant impact of BMI and diabetes on the symptomatology of COVID-19 infection and the duration of long-COVID symptoms. We had also showed that vaccination had a protective effect for both COVID-19 infection and long-COVID symptoms in these at-risk groups. These results highlight the need for further research to understand the mechanisms behind the increased susceptibility and prolonged symptoms in obese and diabetes populations. It is important to investigate the effect of other factors that may influence the susceptibility and duration of symptoms in these populations, such as comorbidities, genetic predisposition, lifestyle, socioeconomic and environment. Furthermore, the findings of this study, adds to the growing evidence that supports future guidelines for COVID-19 (and future pandemics) for these populations.

It is worth noting that our study relied on self-reported data from the CSS app, which may introduce some limitations. Despite these limitations, our study adds to the growing evidence on the importance of app-based syndromic surveillance in providing timely and accurate information on the spread and impact of COVID-19. As we continue to grapple with the ongoing pandemic, app-based surveillance will likely remain a valuable tool for pandemic preparedness and response efforts.

Finally, is worth mentioning the **CSSS dashboard**, which was a valuable tool to share data from the CSSS, helping track the spread of COVID-19 in Sweden. The interactive maps at the regional level, along with the estimated prevalences of national COVID-19 infection, provided daily updates on the CSSS data that was useful for both the public and policy makers. The CSSS dashboard has now been archived, but the codes used to create it are available on GitHub, and the aggregated data is available for download through the *covidsymptom* R package. The package is still available for download on CRAN, and local health authorities have reported using the data and the dashboard in conjunction with other government data to track the spread of the virus in some smaller regions in Sweden. The open-source nature of the CSSS dashboard and *covidsymtpom* package allows for their easy adaptation and customization in future public health crises. The rapid development and deployment of these tools during the COVID-19 pandemic highlights their potential as a valuable asset for pandemic preparedness and response efforts in the future.

Popular science summary

Recent advancements in technology are revolutionizing the way we collect and analyze information in healthcare. The proliferation of new sources of data, such as wearable devices and health-related apps, combined with improvements in data analysis techniques, is leading to a growing amount of information referred to as "big data" in healthcare. Artificial intelligence (AI) and data mining are powerful tools for effectively managing and analyzing this large and complex data. AI-based methods have been successful in analyzing complex healthcare data sets and finding hidden patterns and insights. Data mining, which uses algorithms and statistical models to identify trends, associations, patterns, and features of interest in large data sets, is valuable in providing researchers, data scientists, and public health analysts with insights into factors that affect the health of individuals and communities. Together, these tools can help transform complex healthcare data into useful information and knowledge that can improve healthcare practices.

In this thesis, I built frameworks based on data mining and AI in order to efficiently analyze large health datasets and gain insights into important public health issues.

In the first paper of this thesis, I analyzed the characteristics of researchers and data sources involved in 2300 genome-wide association studies (GWAS) publications for the top-10 non-communicable (NCDs) cases of death using text data mining and AI-based natural language processing. As technology advances, scientists are increasingly using genetic data to understand disease and improve health. The rise of precision medicine, which takes into account individual differences in genes, environment, and lifestyle to provide personalized and effective treatments, is due to the development of new technologies such as GWAS that make it possible to quickly and accurately collect and analyze large amounts of genomic data.

However, the results of our study revealed issues that need to be addressed in order to ensure that the benefits of GWAS research are more evenly distributed and to address the most pressing health needs of the region. Overall, we found that the GWAS were mostly led by male scientists from high-income countries (HIC), particularly the United States (US), and that most of the data used in these studies came from individuals of European ancestry. This is concerning because the lack of diversity in the participants may limit the generalizability of existing genetic discoveries to global populations and perpetuate existing health disparities. Addressing these issues can involve a range of measures, building local capacity

through collaboration with researchers from the underrepresented countries to developing algorithms and techniques that can identify and mitigate bias in the data. Nevertheless, this study highlighted the importance of data mining and AI-based techniques for effectively processing and analyzing large amounts of unstructured data.

The COVID-19 pandemic has highlighted the importance of effective data analysis frameworks. The need for real-time information has led to a surge in data generation and sharing, with frameworks based on data mining and AI techniques having a great potential to be crucial in understanding the spread and impact of the virus. In the second study of this thesis, I presented the COVID Symptom Study Sweden (CSSS), one of the largest epidemiologic initiatives ever done in Sweden.

Using a mobile app (COVID Symptom Study, CSS), participants from Sweden, United Kingdom and US reported daily their symptoms, covid test results and vaccine doses. With data from the Swedish participants, we developed a framework to analyze and predict the spread of COVID-19 at a national and regional level in Sweden. Our framework was able to accurately predict the first two waves of COVID-19 that occurred in the country, highlighting its potential as a valuable tool for monitoring disease trends during times of limited COVID-19 testing capacity. Additionally, we were also able to predict regional levels of COVID-19 hospital admissions seven days in advance, demonstrating the potential of app-based syndromic surveillance as a valuable tool for pandemic preparedness and response efforts.

As the world reaches the third year of the pandemic, new challenges have emerged, one of which is the phenomenon of "long-COVID," where individuals continue to experience symptoms for weeks or even months after their initial infection. This is an important aspect of the ongoing pandemic, as it affects a large number of individuals, and it is critical to understand the causes and consequences of this phenomenon. Additionally, the COVID-19 pandemic is causing a profound impact on populations globally, and certain groups, such as individuals who are overweight or obese, and those with diabetes, have been disproportionately affected by the disease. These groups are at higher risk of severe illness and death, and it is important to continue monitoring and studying the impact of the pandemic on these populations, in order to better understand the underlying causes and develop targeted interventions to address these disparities.

In the third study of this thesis, I used data from the COVID Symptom Study (CSS) app to investigate the symptom patterns of COVID-19 infection and the duration of long-COVID symptoms in relation to body mass index (BMI) and diabetes. Our findings revealed that participants who were obese and those diagnosed with diabetes had a higher prevalence of COVID-19 symptoms during the infection period as well as a longer duration of long-COVID symptoms. Furthermore, we

discovered that vaccination had a protective effect against both COVID-19 infection and long-COVID symptoms in these at-risk groups. These results highlight the importance of addressing the disproportionate impact of COVID-19 on certain populations and the utility of app-based syndromic surveillance in providing timely and accurate information on the spread and impact of the virus.

In conclusion, the aim of this thesis was to investigate the application of data mining and AI-based pipelines to extract insights from large health datasets and generate public health evidence. The research presented here demonstrates the potential of these tools for automating and simplifying the data mining process, and for providing valuable insights into various public health issues. Through the use of these for the development of frameworks to extract and analyze information, it was possible to gain a deeper understanding of the latest two decades of genomic research in NCDs, the early spread of COVID-19 in Sweden, and the impact of BMI and diabetes on COVID-19 symptom presentation and the development of long-COVID.

Populärvetenskaplig sammanfattning

Den senare tidens teknologiska framsteg revolutionerar sättet på vilket vi samlar in och analyserar information inom hälso- och sjukvården. De många nya datakällorna, såsom bärbara enheter och hälsorelaterade appar, leder tillsammans med förbättrade dataanalystekniker till att alltmer ”big data” finns tillgängliga inom vården. Artificiell intelligens (AI) och datautvinning är kraftfulla verktyg för effektiv hantering och analys av dessa omfattande och komplexa data. AI-baserade metoder har använts framgångsrikt för att analysera komplexa hälsorelaterade datauppsättningar och för att hitta dolda mönster och insikter. Datautvinning, som använder algoritmer och statistiska modeller för att identifiera trender, associationer, mönster och intressanta egenskaper i stora datauppsättningar, är ett värdefullt verktyg när det gäller att förse forskare, datavetare och folkhälsoanalytiker med information om faktorer som påverkar hälsan på individ- och samhällsnivå. Tillsammans kan dessa verktyg bidra till att omvandla komplexa hälsodata till användbar information och kunskap som kan förbättra rutinerna inom hälso- och sjukvården.

I den här avhandlingen har jag konstruerat ramverk baserat på datautvinning och AI för att på ett effektivt sätt analysera stora uppsättningar hälsodata och få insikter om viktiga folkhälsoproblem.

I avhandlingens första artikel analyserade jag egenskaperna hos forskare och datakällor involverade i 2 300 GWAS-publikationer (GWAS − genome-wide association studies). Målet var att identifiera de tio främsta dödsorsakerna som inte var relaterade till icke smittsamma sjukdomar med hjälp av textdatautvinning och AI-baserad naturlig språkbearbetning. Allt eftersom teknologin utvecklas använder forskare i allt högre utsträckning genetiska data för att förstå sjukdomar och förbättra folkhälsan. Framväxten av precisionsmedicin, som tar hänsyn till individuella skillnader i gener, miljö och livsstil för att tillhandahålla individanpassade och effektiva behandlingar, har varit möjlig tack vare nya tekniker som GWAS, vilka gör det möjligt att snabbt och noggrant samla in och analysera stora mängder genomiska data.

Resultaten från vår studie avslöjade dock problem som måste åtgärdas för att det ska gå att säkerställa att nyttan av GWAS-forskningen fördelas jämnare och för att de mest akuta vårdbehoven i en region ska tillgodoses. Totalt sett såg vi att GWAS-forskning främst leddes av manliga forskare från höginkomstländer, i synnerhet

USA, och att majoriteten av data som användes i dessa studier kom från individer med europeiskt ursprung. Detta är bekymrande då bristen på mångfald bland deltagarna kan begränsa hur generaliserbara befintliga genetiska upptäcker är på globala populationer och följaktligen kan bibehålla befintliga hälsoskillnader. Att ta itu med dessa frågor kan involvera en rad åtgärder, exempelvis att bygga upp lokal kapacitet genom samarbeten med forskare från underrepresenterade länder för att utveckla algoritmer och tekniker som kan identifiera och mildra bias i data. Denna studie belyste ändå vikten av datautvinning och AI-baserade tekniker för effektiv bearbetning och analys av stora mängder ostrukturerade data.

Covid-19-pandemin har belyst hur viktiga effektiva dataanalysramverk är. Behovet av information i realtid har lett till en stor ökning av datagenerering och -delning. Här har ramverk baserade på datautvinning och AI-tekniker stor potential att vara avgörande för förståelsen av virusets spridning och inverkan. I avhandlingens andra artikel presenterade jag COVID Symptom Study Sweden (CSSS), ett av de största epidemiologiska initiativen någonsin i Sverige.

Med hjälp av appen COVID Symptom Study (CSS) kunde deltagare från Sverige, Storbritannien och USA dagligen rapportera symtom, covidtestresultat och vaccinationer. Vi använde data från deltagarna i Sverige för att utveckla ett ramverk för att analysera och förutspå spridningen av covid-19 på en nationell och regional nivå i Sverige. Detta ramverk kunde korrekt förutspå de första två covid-19-vågorna i landet, vilket visar på dess potential som ett värdefullt verktyg för övervakning av sjukdomstrender i tider med begränsad kapacitet för covid-19-testning. Vi kunde dessutom förutspå regionala nivåer av covid-19-relaterade sjukhusinläggningar sju dagar i förväg, vilket demonstrerar potentialen för appbaserad syndromövervakning som ett värdefullt verktyg för pandemiberedskap och -insatser.

Under det tredje pandemiåret har nya utmaningar dykt upp, varav en består av fenomenet “postcovid”, där individer fortsätter att uppleva symtom veckor eller till och med månader efter den ursprungliga infektionen. Detta är en viktig aspekt av den pågående pandemin då det påverkar ett stort antal individer och det är av yttersta vikt att vi förstår orsakerna för och konsekvenserna av detta fenomen. Dessutom har covid-19-pandemin en djupgående inverkan på populationer globalt och vissa grupper, till exempel individer som är överviktiga eller har fetma och personer med diabetes, har påverkats oproportionerligt mycket av sjukdomen. Dessa grupper löper högre risk att drabbas av allvarlig sjukdom och dödsfall. Det är därför viktigt att vi fortsätter att övervaka och studera hur pandemin påverkar dessa populationer för att bättre förstå de underliggande orsakerna och för att utveckla målriktade interventioner för att komma till rätta med dessa skillnader.

I avhandlingens tredje artikel använde jag data från appen COVID Symptom Study för att undersöka symtommönstren för covid-19-infektion och varaktigheten av postcovidsymtom i relation till BMI (kroppsmasseindex) och diabetes. Våra fynd

visade att deltagare med fetma och personer med en diabetesdiagnos hade en högre prevalens av covid-19-symtom under infektionsperioden samt längre varaktighet av postcovidsymtom. Vidare upptäckte vi att vaccination hade en skyddande effekt mot både covid-19-infektion och postcovidsymtom i dessa riskgrupper. Dessa resultat belyser vikten av att komma till rätta med covid-19:s oproportionerliga inverkan på vissa populationer och användbarheten av appbaserad syndromövervakning för att tillhandahålla läglig och exakt information om virusets spridning och inverkan.

Sammanfattningsvis var syftet med denna avhandling att undersöka tillämpningen av datautvinning och AI-baserade pipelines för att hämta insikter från stora hälsodatauppsättningar och generera folkhälsoevidens. Den forskning som presenteras här demonstrerar den potential som finns hos dessa verktyg för att automatisera och förenkla datautvinningsprocessen och för att tillhandahålla värdefulla insikter om olika folkhälsoproblem. Genom att använda dessa verktyg för att utveckla ramverk för att extrahera och analysera information var det möjligt att få en djupare förståelse för de senaste två årtiondena av genomisk forskning om icke smittsamma sjukdomar, den tidiga spridningen av covid-19 i Sverige och hur BMI och diabetes påverkade presentationen av covid-19-symtom och utvecklingen av postcovid.

Divulgação científica (sumário)

Os avanços tecnológicos têm revolucionado a forma como coletamos e analisamos informações na área da saúde. Ferramentas como "wearables" e aplicativos móveis relacionados à saúde, juntamente com melhorias na metodologia de análise de dados, têm aumentado exponencialmente a quantidade de dados disponíveis na área da saúde, conhecida como "big data". A inteligência artificial (IA) e a mineração de dados são ferramentas valiosas para gerenciar e analisar essa grande quantidade de dados eficazmente. Os métodos baseados em IA têm se mostrado bem-sucedidos na análise de conjuntos complexos de dados de saúde e na descoberta de padrões e insights ocultos. A mineração de dados, que utiliza algoritmos e modelos estatísticos para identificar tendências, associações, padrões e características relevantes em grandes conjuntos de dados, é valiosa para fornecer insights sobre fatores que afetam a saúde de indivíduos e comunidades para pesquisadores, cientistas de dados e analistas de saúde pública. Juntas, essas ferramentas podem ajudar a transformar dados complexos de saúde em informações e conhecimentos úteis que podem melhorar a prática de saúde.

Nesta tese, construí estruturas baseadas em mineração de dados e IA para analisar eficientemente grandes conjuntos de dados de saúde e obter insights sobre importantes questões de saúde pública.

No primeiro artigo desta tese, utilizei técnicas de mineração de dados de texto e processamento de linguagem natural para analisar as características dos pesquisadores e fontes de dados de 2300 publicações de estudos de associação genômica ampla (GWAS) rna área das 10 principais doenças crônicas não transmissíveis.

Com o avanço da tecnologia, os cientistas têm se utilizado cada vez mais de dados genéticos para compreender as doenças e melhorar a prática médica. A medicina de precisão, que leva em conta as diferenças individuais de genes, fatores ambientais e estilos de vida para fornecer tratamentos personalizados e eficazes, tem se beneficiado muito do desenvolvimento de novas tecnologias, como o GWAS, que permitem a coleta e análise rápida e precisa de grandes quantidades de dados genômicos. No entanto, os resultados do nosso estudo revelaram questões que precisam ser abordadas para garantir que os benefícios da pesquisa GWAS sejam distribuídos de forma equitativa e atendam às necessidades de saúde de cada região.

Em geral, descobrimos que a maioria dos estudos de GWAS foi liderada por cientistas do sexo masculino, de países de alta renda, especialmente dos Estados Unidos, e que a maioria dos dados usados nos estudos veio de indivíduos de ascendência europeia. Isso é preocupante, pois a falta de diversidade nos participantes pode limitar a generalização das descobertas genéticas existentes para outras populações, perpetuando assim disparidades de saúde já existentes. Abordar questões como essa pode envolver medidas como o incentivo à colaboração com pesquisadores de países pouco representados na nossa amostra, como também o desenvolvimento de algoritmos e técnicas que possam identificar e mitigar os viéses nos dados de GWAS. De qualquer forma, nosso estudo destacou a importância da mineração de dados e das técnicas baseadas em IA para processar e analisar eficazmente grandes quantidades de dados não estruturados.

A pandemia de COVID-19 também destacou a importância de estruturas de análise de dados eficazes, já que a necessidade de informações em tempo real causou uma explosão na quantidade de dados gerados. No segundo estudo desta tese, apresentei o COVID Symptom Study Sweden (CSSS), uma das maiores iniciativas epidemiológicas já realizadas na Suécia.

Usando um aplicativo móvel (COVID Symptom Study, CSS), participantes da Suécia, Reino Unido e EUA relataram diariamente sintomas, resultados de testes de COVID e doses de vacina. Utilizamos os dados dos participantes suecos provenientos do CSS para desenvolver uma estrutura de análise de dados e prever a disseminação do COVID-19 em nível nacional e regional na Suécia. Nosso método foi capaz de prever com precisão as duas primeiras ondas de COVID-19 que ocorreram no país, destacando seu potencial como ferramenta valiosa para monitorar tendências de doenças durante períodos de capacidade limitada de teste de COVID-19. Além disso, também fomos capazes de prever os níveis regionais de internações hospitalares por COVID-19 com sete dias de antecedência, demonstrando o potencial da vigilância baseada em aplicativo como uma ferramenta valiosa para os esforços de preparação e resposta à pandemias.

À medida que o mundo entra no terceiro ano da pandemia de COVID-19, novos desafios surgiram, como o fenômeno do COVID longo, em que indivíduos continuam apresentando sintomas por semanas ou até meses após a infecção. A COVID longa têm afetado um grande número de pessoas e por isso é fundamental compreender as causas e consequências deste fenômeno. Além disso, não é novidade que o COVID-19 tem um impacto mais significativo em certos grupos de pessoas (chamods grupos de risco). Por exemplo, estudos têm mostrado como indivíduos com sobrepeso ou obesidade e indivíduos diabéticos, apresentam um maior risco de desenvolver sintomas graves de COVID e tem uma maior chance de morte. É, portanto, essencial continuar monitorando e estudando o impacto da

pandemia nessas populações para entender melhor as causas subjacentes e desenvolver intervenções direcionadas para lidar com essas disparidades.

No terceiro estudo desta tese, utilizei dados do aplicativo COVID Symptom Study (CSS) para investigar os padrões de sintomas da infecção por COVID-19 e a duração dos sintomas longos de COVID em relação ao índice de massa corporal (IMC) e diabetes. Os resultados revelaram que os participantes obesos e os participantes diabéticos apresentaram uma maior prevalência de sintomas de COVID-19 durante o período de infecção, além de uma duração mais longa de sintomas desintomas longo de COVID comparados aos outros grupos. Além disso, descobrimos que a vacinação teve um efeito protetor contra a infecção por COVID-19 e sintomas prolongados de COVID nesses grupos de risco. Esses resultados destacam a importância de abordar o impacto desproporcional do COVID-19 em certas populações e a utilidade da vigilância de saúde baseada em aplicativos como forma de fornecer informações relevantes e precisas sobre a disseminação e o impacto do vírus.

Em conclusão, essa tese teve como objetivo investigar a aplicação de técnicas de mineração de dados e de pipelines baseados em IA para extrair informacões relevantes de grandes conjuntos de dados de saúde e gerar evidências para a saúde pública. Os resultados apresentados destacam o potencial dessas ferramentas para automatizar e simplificar o processo de mineração de dados, e fornecer informações valiosas sobre vários problemas de saúde pública. Através da utilizaçnao dessas ferramentas, pudemos obter uma maior compreensão acerca da pesquisa genômica na área de doenças crônicas não transmissíveis, da disseminação inicial do COVID-19 na Suécia e sobre grupos de risco para COVID-19 e COVID longa.

Acknowledgements

I would like to express my sincere gratitude to my two supervisors, **Paul W. Franks** and **Maria F. Gomez** for their invaluable guidance, support, and encouragement throughout my research journey. **Paul**, your belief in my potential has allowed me to grow as a professional and a person. Your mentorship has been a constant source of motivation and inspiration, and I am deeply grateful for the positive impact you have had on my life and career. Your support and friendship have been priceless over the years, and I am truly thankful for everything you have done for me. Thank you. **Maria**, I am grateful for the opportunity to work with you during the COVID Symptom Study, and for your willingness to take me on as a PhD student after Paul's partial departure from Lund University. Your expertise, leadership, and limitless support have been instrumental in my development as a researcher. Thank you.

I would also like to thank my colleagues and members (past and present) from the Genetic and Molecular Epidemiology (GAME) unit. I joined GAME for a short internship during the second year of my masters program and I was fortunate enough to have the opportunity to extend my stay for longer than what was originally planned. The intellectual and professional growth I have experienced while working there was inestimable, and the friendships I have made over the years are invaluable. **Giuseppe** (Nick) **Giordano**, I am grateful for your attentiveness and willingness to provide me with the right support I needed during my PhD. Your dedication and commitment to the GAME unit and its members were evident as we navigated the transition period, and I am thankful for the positive and collaborative atmosphere you have fostered within our group. **Hugo Pomares-Millan**, I am happy to have had the opportunity to share this journey with you since the MPH program. I admire your professionalism and kindness, and I am thankful for our friendship. **Daniel Coral,** I am very glad you have joined the unit and by doing so balancing the Latin American share of the office in our favor! You are a brilliant professional and a kind friend, and I am grateful for the opportunity to have shared some moments of the Friday morning supervision meetings with you and, at the same time, benefit from your thoughtful insights. **Sebastian Kalamajski,** Seb, thank you for your friendship and all the fun in our almost-weekly lunch at Kontrast. I enjoyed our conversations about machine learning and bioinformatics, but not so much losing to you on Duna. **Mi Huang**, my good friend, your exceptional dedication to advancing science has truly inspired me. I have no doubt that you will continue to excel in your work.

Pascal Mutie, 'Pasquale', thank you for laughing at my jokes and for engaging in such interesting discussions with me about programming, statistics, and science in general. **Naeimeh Atabaki-Pasdar, my deep learning book club peer, it was a pleasure to learn with you. You are not only an exceptional researcher but also a kind and generous friend. You have truly paved the way for all of us and I am deeply grateful for your guidance and support. **Neli Tsereteli,** I could never imagine finding someone as geeky about data science as I am. I am incredibly grateful for the time we spent working together, especially during the CSS. You are a highly talented data scientist and I have learned so much from you - more than you might realize. **Pernilla Siming,** you were an integral part of GAME, and your absence has been deeply felt since you left LU. Thank you for your hard work and assistance in navigating the administrative world of LU and for the Swedish lessons during lunchtime. **Marketa Sjögren,** thank you for your constant help and willingness to support me during the final years of my PhD. You are a very kind and resourceful colleague. I also want to express my gratitude to **Juan Fernandez Tajes** and **Ewan Pearson** for their expertise and inputs in my scientific projects. I extend my thanks also to **Tibor V. Varga, Angela Estampador, Alaitz Poveda,** and **Robert Koivula**, former members of GAME who were incredibly helpful to me during the initial phase of my PhD.

I would like to express my deepest gratitude to the participants and teams behind the **COVID Symptom Study (CSS)**, without whom a significant portion of this work would not have been possible. I am thankful for the contributions of **Jordi Merino, Liane dos Santos Canes**, **Carole H. Sudre**, **Sajaysurya Ganesh**, **Claire J. Steves**, **Jonathan Wolf**, **Tim D. Spector**, and all the other team members from **ZOE Limited**, **King's College London**, and **Massachusetts General Hospital** who were involved in the CSS and who collaborated with my research projects through our COVIDX meetings.

I am deeply grateful to the amazing team of the **COVID Symptom Study Sweden (CSSS)** for their contributions to this work. I would like to specifically thank **Paul W. Franks**, **Maria F. Gomez**, and **Tove Fall** for their guidance and leadership during the project and for the valuable input they provided to my projects. I also want to thank **Beatrice Kennedy**, with whom I had the privilege of co-authoring the first CSSS publication, for her tireless work and valuable contributions to the projects I was involved with. **Nikolay Oskolkov**, who was also my mentor at the NBIS Swedish Bioinformatics Advisory Program, I am grateful for all the brainstorming, inputs, and methodological discussions we had related to my projects. To the CSSS analyst team - **Marlena Maziarz**, **Ulf Ummar**, **Georgios Varotsis**, **Neli Tsereteli**, and **Lampros Spiliopoulos** - I thank you for the discussions about statistics, programming, and data visualization and for your inputs on the projects I had the opportunity to lead in CSSS. **Camilla Selberg**, you were a key pillar in this project, and I am thankful for your problem-solving approach,

willingness to help, and the great work we developed together during the project. I would also like to extend my thanks to **Diem Nguyen** and all the other contributors of the CSSS for their efforts and support. Thank you to all of you for your dedication and support.

I would like to express my gratitude to all the Department faculty members of the LUDC, for their help and support during my PhD. In particular, I want to express my heartfelt gratitude to **Mattias Borell**, **Ulrika Blom-Nilsson,** and **Jonathan Esguerra,** for their support. I also extend my gratitude to my colleagues from the wet lab **Claire Lyons**, **Elaine Cowan**, **Alexander Hamilton**, **Klinsmann Carolo,** and **João Paulo Cunha** for all your support and fun parties.

I would like to express my deepest gratitude to my friends and family who have supported me throughout this journey. Here, I will bounce between English and Portuguese.

Rodrigo Souza Ramos and **Dimitrios Floudas**, Rodri and Dimi, we were fortunate to meet you and even luckier to call you family here in Malmö. Thank you for all your support during these years and for all the fun we have together, especially the crazy nights at Azalee. Your friendship means the world to me, and I am grateful to have you in my life.

Sara and **Daniel Johari**, I am thankful for you, for your friendship, for our dinners, board game nights, and deep conversations about life. You have always been amazing listeners to our crazy life stories from across the ocean. South America tour needs to happen!

Esther González-Padilla, I feel very lucky to have shared part of this journey with you. You are an incredibly smart professional and an amazing friend. Thank you for letting the MPH group work more bearable, for all our MCU discussions, and for always being there for us.

I would also like to extend my gratitude and express my deepest thanks to the friends I had the pleasure to meet during my MPH program, and who have remained friends throughout the years: **Noel & Caro, Moe & Jana, Filipa, Alexis & Joakim, Georgia (and family), and Ayza.**

Gelson, **Lorena** e **Nina**, obrigado por todo o apoio durante meu PhD, mas também pelas nossas tardes de raclette, noites de boardgames, idas à praia e a inesquecível eleição de 2022.

Meu querido amigo-irmão **Ivan Luís**, **Laura** e **Ernesto**: obrigado por todo apoio e amizade durante todos esses anos. Vocês sempre estiveram presentes, mesmo que à distância, durante toda a minha jornada na Suécia. Nunca vou me esquecer da nossa viagem ao Porto, do show de Duda Beat e da fatídica noite que comemoramos Lula livre!

Aos meus amigos do Tapajós – **Sam**, **Renan**, **Leo**, **Filipe**, **Lucinho** e **Vitão** – pelos longos anos de amizade e pelo apoio durante esse período.

Ao meu grande amigo e parceiro de longa data, **Victor Monteiro Ramos (e família)**, um dos grandes responsáveis pelo meu desenvolvimento professional como fisioterapeuta. Sou imensamente grato ao apoio e suporte desde o início dessa jornada.

Dirijo um agradecimento especial aos meus pais, **Francisco** e **Ilze**, por serem modelos de integridade e superação, pelo seu amor, incentivo e total ajuda na superação dos obstáculos que surgiram ao longo desta longa caminhada. Ao meu irmão e melhor amigo, **Daniel Fitipaldi**, por todo apoio, carinho, amizade, fraternidade, além de todas as nossas recentes partilhas sobre ciência e engenharia de dados. Aos meus queridos avós, **Terezinha** e **Geraldo** (do qual sinto eternas saudades), meus tios **Junior**, **Yana, Izaura**, **Joana D`arc**, **Tonho,** minhas primas **Camila** e **Marília** e **toda a nossa família**: agradeço pelo seu amor incondicional. Sem o apoio de vocês, eu não seria quem eu sou e nem estaria onde eu estou. Amo vocês.

Finalmente, um agredecimento especial para minha melhor amiga, minha parceira **Camila**. Obrigado por trilhar toda essa longa jornada ao meu lado, por acreditar em mim quando eu não acreditei, por me dar suporte quando precisei, por lutar as minhas lutas, por sempre me incentivar a ser a minha melhor versão e, principalmente, por todo o amor. Espero poder retribuir à altura. Com muito amor, obrigado.

References

1. Jiawei Han MK, Jian Pei. Data Mining: Concepts and Techniques. 3rd ed2012.

2. Dinov ID. Volume and Value of Big Healthcare Data. J Med Stat Inform. 2016;4.

3. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J. 2019;6(2):94-8.

4. Bajwa J, Munir U, Nori A, Williams B. Artificial intelligence in healthcare: transforming the practice of medicine. Future Healthc J. 2021;8(2):e188-e94.

5. Velmovitsky PE, Bevilacqua T, Alencar P, Cowan D, Morita PP. Convergence of Precision Medicine and Public Health Into Precision Public Health: Toward a Big Data Perspective. Front Public Health. 2021;9:561873.

6. Saberi-Karimian M, Khorasanchi Z, Ghazizadeh H, Tayefi M, Saffar S, Ferns GA, et al. Potential value and impact of data mining and machine learning in clinical diagnostics. Crit Rev Clin Lab Sci. 2021;58(4):275-96.

7. Safdari R, Rezayi S, Saeedi S, Tanhapour M, Gholamzadeh M. Using data mining techniques to fight and control epidemics: A scoping review. Health Technol (Berl). 2021;11(4):759-71.

8. Ian J. Goodfellow and Yoshua Bengio AC. Deep Learning: MIT Press; 2016.

9. Topol E. Deep Learning: How Artificial Intelligence can make healthcare human again: Basic Books; 2019.

10. Chang V, Bailey J, Xu QA, Sun Z. Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. Neural Comput Appl. 2022:1-17.

11. Shao W, Luo X, Zhang Z, Han Z, Chandrasekaran V, Turzhitsky V, et al. Application of unsupervised deep learning algorithms for identification of specific clusters of chronic cough patients from EMR data. BMC Bioinformatics. 2022;23(Suppl 3):140.

12. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2014;42(Database issue):D1001-6.

13. Fitipaldi H, McCarthy MI, Florez JC, Franks PW. A Global Overview of Precision Medicine in Type 2 Diabetes. Diabetes. 2018;67(10):1911-22.

14. Slieker RC, Donnelly LA, Fitipaldi H, Bouland GA, Giordano GN, Akerlund M, et al. Replication and cross-validation of type 2 diabetes subtypes based on clinical variables: an IMI-RHAPSODY study. Diabetologia. 2021;64(9):1982-9.

15. Petrovski S, Goldstein DB. Unequal representation of genetic variation across ancestry groups creates healthcare inequality in the application of precision medicine. Genome Biol. 2016;17(1):157.

16. Need AC, Goldstein DB. Next generation disparities in human genomics: concerns and remedies. Trends Genet. 2009;25(11):489-94.

17. Genevieve LD, Martani A, Shaw D, Elger BS, Wangmo T. Structural racism in precision medicine: leaving no one behind. BMC Med Ethics. 2020;21(1):17.

18. Tedros Adhanom Ghebreyesus SS. Get ready for AI in pandemic response and healthcare The BMJ; 2021

19. Albahri AS, Hamid RA, Alwan JK, Al-Qays ZT, Zaidan AA, Zaidan BB, et al. Role of biological Data Mining and Machine Learning Techniques in Detecting and Diagnosing the Novel Coronavirus (COVID-19): A Systematic Review. J Med Syst. 2020;44(7):122.

20. Menni C, Valdes AM, Freidin MB, Sudre CH, Nguyen LH, Drew DA, et al. Real-time tracking of self-reported symptoms to predict potential COVID-19. Nat Med. 2020;26(7):1037-40.

21. Drew DA, Nguyen LH, Steves CJ, Menni C, Freydin M, Varsavsky T, et al. Rapid implementation of mobile technology for real-time epidemiology of COVID-19. Science. 2020;368(6497):1362-7.

22. Dantas LF, Peres IT, Bastos LSL, Marchesi JF, de Souza GFG, Gelli JGM, et al. App-based symptom tracking to optimize SARS-CoV-2 testing strategy using machine learning. PLoS One. 2021;16(3):e0248920.

23. Suri A, Askari M, Calder J, Branas C, Rundle A. A real-time COVID-19 surveillance dashboard to support epidemic response in Connecticut: lessons from an academic-health department partnership. J Am Med Inform Assoc. 2022;29(5):958-63.

24. Gardner L. The COVID-19 Dashboard for Real-time Tracking of the Pandemic: The Lasker-Bloomberg Public Service Award. JAMA. 2022;328(13):1295-6.

25. Chollet F. Deep Learning with Python. Third Edition ed: Manning; 2021.

26. Jeremy Howard SG. Deep Learning for coders with fastai and PyTorch. First Edition ed: O'Reilly Media; 2020.

27. Christian Janiesch PZ, Kai Heinrich Machine learning and deep learning. Electron Markets. 2021;31:685–95.

28. Wengang Zhang HL, Yongqin Li, Hanlong Liu, Yumin Chen, Xuanming Ding. Application of deep learning algorithms in geotechnical engineering: a short critical review. Artif Intell Rev. 2021;54:5633–73.

29. van Dijk WB, Fiolet ATL, Schuit E, Sammani A, Groenhof TKJ, van der Graaf R, et al. Text-mining in electronic healthcare records can be used as efficient tool for screening and data collection in cardiovascular trials: a multicenter validation study. J Clin Epidemiol. 2021;132:97-105.

30. Malden DE, Tartof SY, Ackerson BK, Hong V, Skarbinski J, Yau V, et al. Natural Language Processing for Improved Characterization of COVID-19 Symptoms: Observational Study of 350,000 Patients in a Large Integrated Health Care System. JMIR Public Health Surveill. 2022;8(12):e41529.

31. BuHamra SS, Almutairi AN, Buhamrah AK, Almadani SH, Alibrahim YA. An NLP tool for data extraction from electronic health records: COVID-19 mortalities and comorbidities. Front Public Health. 2022;10:1070870.

32. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 2017;45(D1):D896-D901.

33. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019;47(D1):D1005-D12.

34. Sollis E, Mosaku A, Abid A, Buniello A, Cerezo M, Gil L, et al. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. Nucleic Acids Res. 2023;51(D1):D977-D85.

35. Wang Y, Tsuo K, Kanai M, Neale BM, Martin AR. Challenges and Opportunities for Developing More Generalizable Polygenic Risk Scores. Annu Rev Biomed Data Sci. 2022;5:293-320.

36. Ozaki K, Ohnishi Y, Iida A, Sekine A, Yamada R, Tsunoda T, et al. Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. Nat Genet. 2002;32(4):650-4.

37. Ikegawa S. A short history of the genome-wide association study: where we were and where we are going. Genomics Inform. 2012;10(4):220-5.

38. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, et al. Complement factor H polymorphism in age-related macular degeneration. Science. 2005;308(5720):385-9.

39. Dewan A, Liu M, Hartman S, Zhang SS, Liu DT, Zhao C, et al. HTRA1 promoter polymorphism in wet age-related macular degeneration. Science. 2006;314(5801):989-92.

40. Wellcome Trust Case Control C. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007;447(7145):661-78.

41. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. Am J Hum Genet. 2012;90(1):7-24.

42. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. Am J Hum Genet. 2017;101(1):5-22.

43. Loos RJF. 15 years of genome-wide association studies and no signs of slowing down. Nat Commun. 2020;11(1):5900.

44. Mahajan A, Taliun D, Thurner M, Robertson NR, Torres JM, Rayner NW, et al. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. Nat Genet. 2018;50(11):1505-13.

45. Michailidou K, Beesley J, Lindstrom S, Canisius S, Dennis J, Lush MJ, et al. Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. Nat Genet. 2015;47(4):373-80.

46. Emil Uffelmann QQH, Nchangwi Syntia Munung, Jantina de Vries, Yukinori Okada, Alicia R. Martin, Hilary C. Martin, Tuuli Lappalainen, Danielle Posthuma. Genome-wide association studies. Nat Rev Methods Primers. 2021;1(59).

47. Barroso I. The importance of increasing population diversity in genetic studies of type 2 diabetes and related glycaemic traits. Diabetologia. 2021;64(12):2653-64.

48. Bentley AR, Callier S, Rotimi CN. Diversity and inclusion in genomic research: why the uneven progress? J Community Genet. 2017;8(4):255-66.

49. Gurdasani D, Barroso I, Zeggini E, Sandhu MS. Genomics of disease risk in globally diverse populations. Nat Rev Genet. 2019;20(9):520-35.

50. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. Nat Genet. 2019;51(4):584-91.

51. Mills MC, Rahal C. A scientometric review of genome-wide association studies. Commun Biol. 2019;2:9.

52. Popejoy AB, Fullerton SM. Genomics is failing on diversity. Nature. 2016;538(7624):161-4.

53. Sirugo G, Williams SM, Tishkoff SA. The Missing Diversity in Human Genetic Studies. Cell. 2019;177(1):26-31.

54. Zhang H, De T, Zhong Y, Perera MA. The Advantages and Challenges of Diversity in Pharmacogenomics: Can Minority Populations Bring Us Closer to Implementation? Clin Pharmacol Ther. 2019;106(2):338-49.

55. Takeuchi F, McGinnis R, Bourgeois S, Barnes C, Eriksson N, Soranzo N, et al. A genome-wide association study confirms VKORC1, CYP2C9, and CYP4F2 as principal genetic determinants of warfarin dose. PLoS Genet. 2009;5(3):e1000433.

56. Fung E, Patsopoulos NA, Belknap SM, O'Rourke DJ, Robb JF, Anderson JL, et al. Effect of genetic variants, especially CYP2C9 and VKORC1, on the pharmacology of warfarin. Semin Thromb Hemost. 2012;38(8):893-904.

57. Franks PW. The complex interplay of genetic and lifestyle risk factors in type 2 diabetes: an overview. Scientifica (Cairo). 2012;2012:482186.

58. Bookman EB, McAllister K, Gillanders E, Wanke K, Balshaw D, Rutter J, et al. Gene-environment interplay in common complex diseases: forging an integrative model-recommendations from an NIH workshop. Genet Epidemiol. 2011;35(4):217-25.

59. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, et al. Addendum: A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020;588(7836):E6.

60. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, et al. A new coronavirus associated with human respiratory disease in China. Nature. 2020;579(7798):265-9.

61. Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. The proximal origin of SARS-CoV-2. Nat Med. 2020;26(4):450-2.

62. Cucinotta D, Vanelli M. WHO Declares COVID-19 a Pandemic. Acta Biomed. 2020;91(1):157-60.

63. (JHU) Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU). COVID-19 Dashboard, 2022 [Available from: https://coronavirus.jhu.edu/map.html]

64. Zhang R, Li Y, Zhang AL, Wang Y, Molina MJ. Identifying airborne transmission as the dominant route for the spread of COVID-19. Proc Natl Acad Sci U S A. 2020;117(26):14857-63.

65. Tan ST, Kwan AT, Rodriguez-Barraquer I, Singer BJ, Park HJ, Lewnard JA, et al. Infectiousness of SARS-CoV-2 breakthrough infections and reinfections during the Omicron wave. medRxiv. 2022.

66. WHO. Advice for the public: Coronavirus disease (COVID-19) 2022 [Available from: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public]

67. Han E, Tan MMJ, Turk E, Sridhar D, Leung GM, Shibuya K, et al. Lessons learnt from easing COVID-19 restrictions: an analysis of countries and regions in Asia Pacific and Europe. Lancet. 2020;396(10261):1525-34.

68. Brown A, Flint SW, Kalea AZ, O'Kane M, Williams S, Batterham RL. Negative impact of the first COVID-19 lockdown upon health-related behaviours and psychological wellbeing in people living with severe and complex obesity in the UK. EClinicalMedicine. 2021;34:100796.

69. Lauer SA, Grantz KH, Bi Q, Jones FK, Zheng Q, Meredith HR, et al. The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application. Ann Intern Med. 2020;172(9):577-82.

70. CDC. People with Certain Medical Conditions 2022 [Available from: https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/people-with-medical-conditions.html]

71. Tejada-Vera B, Kramarow EA. COVID-19 Mortality in Adults Aged 65 and Over: United States, 2020. NCHS Data Brief. 2022(446):1-8.

72. Watson C. Rise of the preprint: how rapid data sharing during COVID-19 has changed science forever. Nat Med. 2022;28(1):2-5.

73. Bok K, Sitar S, Graham BS, Mascola JR. Accelerated COVID-19 vaccine development: milestones, lessons, and prospects. Immunity. 2021;54(8):1636-51.

74. Dron L, Kalatharan V, Gupta A, Haggstrom J, Zariffa N, Morris AD, et al. Data capture and sharing in the COVID-19 pandemic: a cause for concern. Lancet Digit Health. 2022;4(10):e748-e56.

75. European Comission. Covid-19: How unprecedented data sharing has led to faster-than-ever outbreak research 2020 [Available from: https://ec.europa.eu/research-and-innovation/en/horizon-magazine/covid-19-how-unprecedented-data-sharing-has-led-faster-ever-outbreak-research]

76. Schmeelk S, Davis A, Li Q, Shippey C, Utah M, Myers A, et al. Monitoring Symptoms of COVID-19: Review of Mobile Apps. JMIR Mhealth Uhealth. 2022;10(6):e36065.

77. Bonander C, Stranges D, Gustavsson J, Almgren M, Inghammar M, Moghaddassi M, et al. A regression discontinuity analysis of the social distancing recommendations for older adults in Sweden during COVID-19. Eur J Public Health. 2022;32(5):799-806.

78. Arjun MC, Singh AK, Pal D, Das K, G A, Venkateshan M, et al. Characteristics and predictors of Long COVID among diagnosed cases of COVID-19. PLoS One. 2022;17(12):e0278825.

79. Moy FM, Hairi NN, Lim ERJ, Bulgiba A. Long COVID and its associated factors among COVID survivors in the community from a middle-income country-An online cross-sectional study. PLoS One. 2022;17(8):e0273364.

80. Notarte KI, Catahay JA, Velasco JV, Pastrana A, Ver AT, Pangilinan FC, et al. Impact of COVID-19 vaccination on the risk of developing long-COVID and on existing long-COVID symptoms: A systematic review. EClinicalMedicine. 2022;53:101624.

81. Sudre CH, Murray B, Varsavsky T, Graham MS, Penfold RS, Bowyer RC, et al. Attributes and predictors of long COVID. Nat Med. 2021;27(4):626-31.

82. Vimercati L, De Maria L, Quarato M, Caputi A, Gesualdo L, Migliore G, et al. Association between Long COVID and Overweight/Obesity. J Clin Med. 2021;10(18).

83. Liane S. Canas EM, Jie Deng, Carole H. Sudre, Benjamin Murray, Eric Kerfoot, Michela Antonelli, Liyuan Chen, Khaled Rjoob, Joan Capdevila Pujol, Lorenzo Polidori, Anna May, Marc F. Österdahl, Ronan Whiston, Nathan J. Cheetham, Vicky Bowyer, Tim D. Spector, Alexander Hammers, Emma L. Duncan, Sebastien Ourselin, Claire J. Steves, Marc Modat. Profiling post-COVID syndrome across different variants of SARS-CoV-2. medRxiv. 2022.

84. Chi WY, Li YD, Huang HC, Chan TEH, Chow SY, Su JH, et al. COVID-19 vaccine update: vaccine effectiveness, SARS-CoV-2 variants, boosters, adverse effects, and immune correlates of protection. J Biomed Sci. 2022;29(1):82.

85. Morales J, Welter D, Bowler EH, Cerezo M, Harris LW, McMahon AC, et al. A standardized framework for representation of ancestry data in genomics studies, with application to the NHGRI-EBI GWAS Catalog. Genome Biol. 2018;19(1):21.

86. Lu Z. PubMed and beyond: a survey of web tools for searching biomedical literature. Database (Oxford). 2011;2011:baq036.

87. Williams FMK, Freidin MB, Mangino M, Couvreur S, Visconti A, Bowyer RCE, et al. Self-Reported Symptoms of COVID-19, Including Symptoms Most Predictive of SARS-CoV-2 Infection, Are Heritable. Twin Res Hum Genet. 2020;23(6):316-21.

88. Antonelli M, Penfold RS, Merino J, Sudre CH, Molteni E, Berry S, et al. Risk factors and disease profile of post-vaccination SARS-CoV-2 infection in UK users of the COVID Symptom Study app: a prospective, community-based, nested, case-control study. Lancet Infect Dis. 2022;22(1):43-55.

89. SmiNet. SmiNet: Elektronisk anmälan av smittsamma sjukdomar 2021 [Available from: https://www.folkhalsomyndigheten.se/sminet/]

90. Novus. Novus Sverigepanel 2022 [Available from: https://novus.se/metoder/sverigepanel/]

91. CRUSH. CRUSH Covid Uppsala 2022 [Available from: https://www.uu.se/forskning/projekt/crush-covid/]

92. Socialstyrelsen. Patientregistret 2019 [Available from: https://www.socialstyrelsen.se/statistik-och-data/register/patientregistret/]

93. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. 2021.

94. Vikas Yadav SB. A Survey on Recent Advances in Named Entity Recognition from Deep Learning models. arXiv. 2019.

95. Kenneth Benoit AM. spacyr: Wrapper to the 'spaCy' 'NLP' Library. R package version 121. 2021.

96. Eyre H, Chapman AB, Peterson KS, Shi J, Alba PR, Jones MM, et al. Launching into clinical space with medspaCy: a new clinical text processing toolkit in Python. AMIA Annu Symp Proc. 2021;2021:438-47.

97. Santamaria L, Mihaljevic H. Comparison and benchmark of name-to-gender inference services. PeerJ Comput Sci. 2018;4:e156.

98. Sebo P. Performance of gender detection tools: a comparative study of name-to-gender inference services. J Med Libr Assoc. 2021;109(3):414-21.

99. Reza N, Tahhan AS, Mahmud N, DeFilippis EM, Alrohaibani A, Vaduganathan M, et al. Representation of Women Authors in International Heart Failure Guidelines and Contemporary Clinical Trials. Circ Heart Fail. 2020;13(8):e006605.

100. Gareth James DW, Trevor Hasti, Robert Tibshirani. An Introduction to Statistical Learning: with Applications in R. 2nd ed: Springer; 2021.

101. Agresti A. Categorical Data Analysis. 2nd ed: John Wiley & Sons, Inc.; 2002.

102. Signorell A. DescTools: Tools for descriptive statistics. CRAN. 2022;R package version 0.99.47.

103. Schimek MG. Penalized Logistic Regression in Gene Expression Analysis. Proc The Art of Semiparametrics Conference. 2003.

104. Antonio Olmos PG. A Practical Guide for Using Propensity Score Weighting in R Practical Assessment, Research & Evaluation. 2015;20(13).

105. Shenyang Guo MWF. Propensity score analysis: Statistical methods and applications. 2nd ed: SAGE; 2015.

106. Hadley Wickham JB. R Packages: O'Reilly Media, Inc.; 2015.

107. Fitipaldi H. affiliation: R package to work with the affiliation field from PubMed publications. R package version 009. 2021.

108. Fitipaldi H. genderAPI: Functions for interacting directly with the Gender API. R package version 001. 2021.

109. Fitipaldi H. covidsymptom: COVID Symptom Study Sweden Open Dataset. R package version 093. 2021.

110. Simo Goshev SW. Package Development in R. Bookdown, editor2019.

111. Diseases GBD, Injuries C. Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet. 2020;396(10258):1204-22.

112. Wickham H. Mastering Shiny: Build Interactive Apps, Reports, and Dashboards Powered by R: O'Reilly Media, Inc; 2021.

113. Sievert C. Interactive Web-Based Data Visualization with R, plotly, and shiny: CRC Press; 2020.

114. Lanera C, Azzolina D, Pirotti F, Prosepe I, Lorenzoni G, Berchialla P, et al. A Web-Based Application to Monitor and Inform about the COVID-19 Outbreak in Italy: The COVID-19ita Initiative. Healthcare (Basel). 2022;10(3).

115. Salehi M, Arashi M, Bekker A, Ferreira J, Chen DG, Esmaeili F, et al. A Synergetic R-Shiny Portal for Modeling and Tracking of COVID-19 Data. Front Public Health. 2020;8:623624.

116. Tebe C, Valls J, Satorra P, Tobias A. COVID19-world: a shiny application to perform comprehensive country-specific data visualization for SARS-CoV-2 epidemic. BMC Med Res Methodol. 2020;20(1):235.

117. Valls J, Tobias A, Satorra P, Tebe C. [COVID19-Tracker: a shiny app to analise data on SARS-CoV-2 epidemic in Spain]. Gac Sanit. 2021;35(1):99-101.

118. Winston Chang BR. shinydashboard: Create Dashboards with 'Shiny'. R package version 072. 2021.

119. Joe Cheng BK, Yihui Xie. leaflet: Create Interactive Web Maps with the JavaScript 'Leaflet' Library. R package version 210. 2022.

120. Statistikmyndigheten. Digitala gränser 2022 [Available from: https://www.scb.se/hitta-statistik/regional-statistik-och-kartor/regionala-indelningar/digitala-granser/]

121. Postnummerservice. Postnummerytor, 2-siffrig nivå 2020 [Available from: https://www.postnummerservice.se/utbud/referensdata/geodata/postnummerytor-2-siffriga]

122. Yihui Xie JC, Xianying Tan. DT: A Wrapper of the JavaScript Library 'DataTables'. R package version 020. 2021.

123. Hadley Wickham DN, Thomas Lin Pedersen. ggplot2: Elegant Graphics for Data Analysis: Springer-Verlag New York; 2016.

124. Liu C, Zeinomar N, Chung WK, Kiryluk K, Gharavi AG, Hripcsak G, et al. Generalizability of Polygenic Risk Scores for Breast Cancer Among Women With European, African, and Latinx Ancestry. JAMA Netw Open. 2021;4(8):e2119084.

125. Duncan L, Shen H, Gelaye B, Meijsen J, Ressler K, Feldman M, et al. Analysis of polygenic risk score usage and performance in diverse human populations. Nat Commun. 2019;10(1):3328.

126. Crew B. The top 10 countries for scientific research in 2018. Nature index. 2019.

127. Fontelo P, Liu F. A review of recent publication trends from top publishing countries. Syst Rev. 2018;7(1):147.

128. Rotimi SO, Rotimi OA, Salhia B. Authorship Patterns in Cancer Genomics Publications Across Africa. JCO Glob Oncol. 2021;7:747-55.

129. Adedokun BO, Olopade CO, Olopade OI. Building local capacity for genomics research in Africa: recommendations from analysis of publications in Sub-Saharan Africa from 2004 to 2013. Glob Health Action. 2016;9:31026.

130. Ross MB, Glennon BM, Murciano-Goroff R, Berkes EG, Weinberg BA, Lane JI. Women are credited less in science than men. Nature. 2022;608(7921):135-45.

131. Valerie K. Bostwick BAW. Nevertheless She Persisted? Gender Peer Effects in Doctoral STEM Programs. J Lab Econ. 2018;40:397–436.

132. Catherine Buffington BC, Christina Jones, Bruce A. Weinberg. STEM Training and Early Career Outcomes of Female and Male Graduate Students: Evidence from UMETRICS Data Linked to the 2010 Census. American Economic Review. 2016;106:333–8.

133. Rossman H, Keshet A, Shilo S, Gavrieli A, Bauman T, Cohen O, et al. A framework for identifying regional outbreak and spread of COVID-19 from one-minute population-wide surveys. Nat Med. 2020;26(5):634-8.

134. Delphi Epidata API. COVID-19 Trends and Impact Survey 2022 [Available from: https://cmu-delphi.github.io/delphi-epidata/symptom-survey/]

135. Gao F, Zheng KI, Wang XB, Sun QF, Pan KH, Wang TY, et al. Obesity Is a Risk Factor for Greater COVID-19 Severity. Diabetes Care. 2020;43(7):e72-e4.

136. Simonnet A, Chetboun M, Poissy J, Raverdy V, Noulette J, Duhamel A, et al. High Prevalence of Obesity in Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2) Requiring Invasive Mechanical Ventilation. Obesity (Silver Spring). 2020;28(7):1195-9.

137. Piernas C, Patone M, Astbury NM, Gao M, Sheikh A, Khunti K, et al. Associations of BMI with COVID-19 vaccine uptake, vaccine effectiveness, and risk of severe COVID-19 outcomes after vaccination in England: a population-based cohort study. Lancet Diabetes Endocrinol. 2022;10(8):571-80.

138. Rawshani A, Kjolhede EA, Rawshani A, Sattar N, Eeg-Olofsson K, Adiels M, et al. Severe COVID-19 in people with type 1 and type 2 diabetes in Sweden: A nationwide retrospective cohort study. Lancet Reg Health Eur. 2021;4:100105.

139. Ohno M, Dzurova D. Body Mass Index and Risk for COVID-19-Related Hospitalization in Adults Aged 50 and Older in Europe. Nutrients. 2022;14(19).

140. Davis HE, McCorkell L, Vogel JM, Topol EJ. Long COVID: major findings, mechanisms and recommendations. Nat Rev Microbiol. 2023.

141. Berbudi A, Rahmadika N, Tjahjadi AI, Ruslami R. Type 2 Diabetes and its Impact on the Immune System. Curr Diabetes Rev. 2020;16(5):442-9.

142. Folkhalsomyndigheten. Statistik och analyser om covid-19 inklusive vaccinationer 2022 [https://experience.arcgis.com/experience/09f821667ce64bf7be6f9f87457ed9aa]

Printed by Media-Tryck, Lund 2023 NORDIC SWAN ECOLABEL 3041 0903

HUGO FITIPALDI completed his BSc in Physiotherapy at Universidade Federal de Pernambuco (UFPE), Brazil. In 2018, he graduated with a MSc in Public Health (MPH) from Lund University. Hugo has completed his PhD at the Genetic and Molecular Epidemiology unit at the Lund University Diabetes Centre. Hugo's thesis focuses on the application of data mining and AI-based frameworks to extract insights from large health datasets and generate public health evidence.

LUND UNIVERSITY

FACULTY OF MEDICINE

Department of Clinical Research

Lund University, Faculty of Medicine
Doctoral Dissertation Series 2022:24
ISBN 978-91-8021-363-9
ISSN 1652-8220