universität wien

MASTERARBEIT | MASTER'S THESIS

Titel | Title

Time Series Comparison of Global Infectious Disease Cases and Online Symptom-Checker Assessments
A comparison of time series resulting from confirmed cases of infectious diseases and online symptom-checker assessment.

verfasst von | submitted by
Marc David Zobel BSc

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of
Master of Science (MSc)

Wien | Vienna, 2025

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 921
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Informatik
Betreut von | Supervisor:	Assoz. Prof. Dipl.-Inform.Univ. Dr. Christian Böhm

Contents

Abstract

This work explores the similarity between test data generated via the online symptom checker *Symptoma* and confirmed global COVID-19 cases aggregated by Johns Hopkins University (JHU). It contributes to the large body of work concerned with the surveillance of epidemics under the umbrella term Infodemiology[11]. We show that *Symptoma* produces unique time series for diseases with different symptoms. We selected the top 40 returned diseases in Austria and produced time series for these by aggregating the amount of users that had the respective disease in the top 30 of possible causes returned by *Symptoma*. The Pearson correlation coefficient (PCC) between those time series and COVID-19 as well as the overlap coefficient between symptoms were calculated. By comparing those two we found that *Symptoma* produces distinct time series for diseases that are dissimilar in their set of symptoms. Additionally data from Symptoma GmbH and Johns Hopkins University was collected and preprocessed between 1st of May 2020 and 1st of May 2021. The PCC was calculated after introducing optimal lags between -30 and +30 days for 18 out of 84 countries that showed 366 consecutive days of usage. The countries show a median lag of +10 days and a median PCC of 0.55. Those results are confirmed by the shifted euclidean distance (L_2-$norm$), dynamic time warping (DTW) and time warped edit distance ($TWED$). The positive median lag suggests that cases show up earlier in *Symptoma* and that the data produced by the online symptom checker could be used as an early warning signal. We further discuss the limitations of such usage. The similarity between real world case counts and *Symptoma* data decreased in most countries after the second wave of infections. We attribute this to media fatigue within the population with regards to the COVID-19 pandemic as others have suggested in their work as well[36].

1 Introduction

When looking at the global population history over the last 2.000 years we can see a significantly increased growth rate starting from roughly the end of the 17th century. A period of rapid progress in the areas of manufacturing and transportation started around this time. We call this period the Industrial Revolution. In just 200 years humanity developed and improved the steam engine, built large networks of railway tracks and developed automatic manufacturing methods. Telegraphs were built in 1844, which in 1860 already connected the east coast of the US with areas west of the Mississippi. In 1870 Louis Pasteur developed the first vaccines, leading to an improvement of peoples lifespans. The patent for the telephone dates back to 1879. Not even 30 years later, in 1903, the first engine powered plane built by the Wright brothers stays in the air for 12 seconds and only five years later the first car produced in an assembly line[25, 5]. How was such a rapid development of technology and world population possible?

One of the reasons given by Nate Silver [52] is the development of technologies, which facilitate communication and cooperation. We find that the

invention of the printing press in 1440 by Johannes Gutenberg coincides with the start of said development. Before there was no standardized and reproducible method to share knowledge with a broad audience. Information needed to be copied by hand, which introduced errors and was quite costly [10, 41]. Besides that, information exchange happened orally, which introduces even bigger discrepancies between knowledge sent and the one received[10]. "In Gutenberg's day, Europe was already on the brink of a new age. Exploration, scientific discoveries, and the growth of towns, industry, and commerce were changing the political and religious structure of medieval society" [5]. Gutenberg's invention allowed nearly identical copies and drastically increased affordability. Additionally reprinting led to increased longevity of information. Though most of the early publications were of religious nature, this development had undoubtedly a positive influence on the proliferation of technological and scientific knowledge [10, 52].

One of Gutenbergs first large printing projects on which he worked on between 1452 and 1455 is, what we call today the Gutenberg-Bible [5]. Another is the printing of the Catholicon, a detailed latin dictionary, including grammar as well, compiled in 1286 by Johann Balbus. Gutenberg finished this project 1461. A Bishop in 1468 commented: "volumes which in former times could scarce be bought for a hundred gold pieces are today to be had for reading in good versions and free of faults troughout for twenty"[10]. Price and quality of work led to printing becoming very popular. Until 1476 several print shops were opened in Naples, Nuremberg, Paris and Westminster. Many more would be opened in the following years. In the year 1500 about 2.000 different book editions, pamphlets and broadsides were printed, next to calendars, charts, proclamations and advertisements [5]. Before the end of the 1500s nearly 30.000 works were printed. Books, being the most dense source of information were still largely in latin, which meant that in the beginning almost only the latin elite had access to scientific relevant information. Elizabeth Eisenstein writes: "[...] the end of the fifteenth century still seems much too early to think of scientific knowledge being 'rapidly disseminated to whole new classes'", page 691 [10]. Those elites saw an opportunity to spread their influence and knowledge, which led to further investments in printing. Protestant theologists especially encouraged lay reading to spread their belief in light of the Luther Reformation 1517. Translation workflows for religious texts would later be used to translate scientific texts as well. Between 1583 and 1712 the Elzevier printing firm produced around 1.600 volumes currently known as the Elzevier Collection, containing mostly scientific works in latin in the beginning. A large number of them were small in size and easy to carry. They expanded their market in the second quater of the 17th century to include non-scientific work and published in, for the time, a large number of languages. These include German, Italian, Dutch, French, Greek, Hebrew, Rabbinic, Samaritanic, Armenic, Arabic, Persian, Ethiopia, making print media available for a broad audience. 1690 the first american newspaper was printed in Boston [6], which improved the spread of information to the common folk. At the same time "[...] a majority of well-read Europeans had been persuaded by the ideas of Copernicus, Vesalius, Galileo,

Newton, and others. A scientific revolution had been accomplished through the power of individuals' ideas and the spread of those ideas by tracts and books produced with the printing press." Samuel Crompton writes [6]. This suggests that printing influenced the spread of scientific information, which in turn supported major advancements in engineering. One of those advancements feeding back into the popularity and spread of printed media was rotary printing, patented in 1790 by William Nicholson as the rotary press and further developed into the rotary drum printing in 1843 by Richard March Hoe [22].

Just a year after, in 1844 the Telegraph was invented, reigning in the age of near instant communication over large distances. The invention of the Telephone followed in 1876 and radio broadcasts started in 1922 [15]. Communication methods kept improving, which in turn helped to improve communication methods by connecting clever minds for better cooperation. 1945 saw the first modern computers, a historic event that contributed greatly to how society, science and communication would in the years to follow. During the later years of the 1970s Robert E. Kahn introduced the idea of open-architecture networking. With the help of Vinton G. Cerf, those ideas would later be refined and published as the TCP protocol [28]. Before there were two main ways to communicate over long distances. Either asynchronous via written letters or synchronous via telephone. Information exchange via print media is slow since the physical medium has to be transported to the destination. Communication via telephone emulates a conversation. The Internet provided humanity with the means to communicate directly in written form, over great distances. Originally developed in 1968 the small, military backed, university connecting network ARPANET adopted TCP routing in 1983. This decision made it the first subdivision of the Internet we know today. The TCP protocol provided much needed reliability and data security to network based communications. In 1975 the first personal computer goes into production, providing the ground work for information technology to reach the masses. Fast forward a couple of decades and the Internet is everywhere. It started with static pages, defined using HTML and continued to evolve into what is commonly referred to as Web 2.0, a network providing dynamic pages, services and platforms for communication, information exchange. E-government services, music streaming, video streaming, e-learning, messenger and one of the domains of this work, digital health, are technologies that rely heavily on the aspects of reliable, secure and fast communication via the global network. They all have analogue versions, which in the case of messengers, music and video streaming are in decline. The analogue versions of digital health, e-government and e-learning will most likely remain in our society for quite some time since their underlying systems are greatly interconnected with structures which are known to change at a very slow pace. Innovation in those fields can still improve accessibility and decrease bureaucratic overhead.

Digital Health

The development of the printing press naturally had an effect on the distribution of medical literature and thus knowledge. "doctors, surgeons, and apothecaries were usually literate, and doctors and leading surgeons were usually literate in Latin as well as in their own language. They were also, on the whole, wealthy, able to buy books out of their own resources and to build up their own libraries." V. Nutton reports [41]. The documentation of knowledge has always been a big part of medicine and the innovation from handwritten manuscripts to the printed book improved distribution, amount and variety. Being a printer meant to look at the quality of the original text and decide if publishing it would be worth the time and effort. The use of high quality materials still translated to a high priced product, but there are examples of low quality prints, designed for less wealthy people. One of those instances was the "[...] Epitome, deliberately aimed at a less wealthy market, printed on poorer paper, with few pictures and with a poorer typeface, but published at a price that more could afford."[41]. Another are anatomical fugitive sheets, which were sold before 1543 around Europe for just a few pence. They used "[...] flaps to reveal the insides of the body, and were accompanied by a brief text, sometimes in latin, sometimes in vernacular."[41].

The Internet and digital storage media can be seen as the next step of innovation in regards to the printing press for medical texts. Not only does the digital nature of data storage allow for free copies, but the global reach of the internet further improves the availability, amount and variety of data. Technologies in this field are grouped under the umbrella term digital health. According to F. Seth, the Internet performs five critical functions for the patient in the medical domain: "(1)disseminate information, (2)aid informed decision making, (3)promote health, (4)provide a means for information exchange and support, (5)Increase self-care and manage demand for health services, lowering direct medical costs."[13]. These map well to the positive effects of improved knowledge exchange that were outlined before. One of the first concepts under discussion were electronic health records (EHRs). The goal was a specification to store data related to the patient, treatment and medical history in a structured format. Multiple versions were designed and implemented. But due to local differences in storage and interpretation of the data types, to date no global standard has emerged [4]. In the last decade solutions that can be classified as assistants gained popularity. Those include algorithms and products, which assist practitioners during various tasks like diagnosis, tests and anamnesis [17]. Another sub group consists of software and wearable hardware, which assists patients in measuring health related factors. The last subgroup contains systems, which accept text input, e.g. symptoms, as query to perform a risk assessment and guide the patient towards relief. Systems of the latter subgroup can be called symptom-to-disease health assistants, also known as symptom checkers.

COVID-19 Pandemic

Epidemics and pandemics were always present in human history and changed societies in their wake. Plague, smallpox and measles reduced European, Middle Eastern and Asian populations greatly. The later two were carried over by the spanish conquistadores during exploration and conquest of the american continent at the end of the fifteenth century. Lacking natural resistance, the high mortality rate of the native population led to conversion to christianity and facilitated said conquest. [46] [42]

In the twentieth century the AIDS pandemic broke out and quickly spread on all continents. The countries of the african continent suffered significant shortening of life expectancy, massive destruction of family units, and orphanhood. Societies had and to some degree still have to deal with the damage dealt to the economy. Due to its immune suppression AIDS led to increased cases of other illnesses, leaving the healthcare system under great pressure. [46] [42] [63]

The spread of the novel coronavirus SARS-CoV-2 in late 2019, early 2020 is the first global health crisis in the digital age. Originating in Wuhan, China, the virus spread quickly to other countries because of its highly infectious nature. On the 11th of February 2020 the disease caused by the SARS-CoV-2 was named COVID-19 by the WHO. Only one month later, on the 11th of March it was categorized as a pandemic. In the following months governments implemented policies to prevent the spread of the disease, enable research and minimize the damage to the economy with mixed results. A crucial part during all the stages of the pandemic was the tracking of new infections to identify emerging clusters. Tracking of epidemics[16],[37],[26] is a well studied field whose research provides a base for research targetting tracking in a pandemic scenario. The state of digital communication ensured fast propagation of current data, but the count prone to underestimation. Reporting of new cases started slow, since it took a while for test kits to become readily available [61]. During that time most cases were reported by doctors and hospitals. Governmental health departments would aggregate and release those case numbers, sometimes supported by corporations or research institutions. Johns Hopkins University became one of the institutions to aggregate and release a dataset of global scope, containing new infections counted per day, per country. Data like this is valuable for policy makers to measure the current situation of a pandemic and use those measurements to make decisions. Low testing capacity because of resource shortage or missing research in the field can skew those measurements and reduce their usefulness.

Symptom Checkers

Before the proliferation of health services via the world wide web patients could only rely on doctors and to some extent medical hotlines for relief. This could in certain cases be a barrier preventing patients to seek help. Interactive web sources developed to provide web-based diagnoses are sometimes referred to as symptom checkers. These systems return a list of potential diseases based on a

list of symptoms [39]. Some symptom checkers assist the user in the process of communicating the relevant symptoms. Some provide a risk assessment and/or triaging advice. The symptoms provided by the patient can be compiled into a structured report. The printed version of said document can speed up the process of anamnesis and ordination during a following doctor visit. The resulting digital health assistant can guide the patient through the steps towards relief based on the risk assessment. Munsch et al. conducted a study which compared the accuracy of 10 symptom checkers [39]. They used Sensitivity, Specificity, F1 score and Matthew's Correlation Coefficient as measures of comparison. The 10 symptom checkers, including their respective F1 score are: Docyet (0.27), Ada (0.42), Apple (0.7), Babylon (0.7), CDC (0.71), Your.MD (0.72), Providence (0.75), Cleveland Clinic (0.76), Infermedica (0.8), Symptoma (0.91).

Symptoma and COVID-19

Previous studies have shown that Symptoma performance is above other symptom checkers [34] [39]. Symptoma GmbH implemented three strategies to achieve that state. *Strategy one* consists of "a large text corpus that includes scientific publications, medical textbooks, patient self-reports, and Electronic Health Records" [34] mapping user input to possible causes. *Strategy two* is a proprietary symptom-disease database built from the aforementioned corpus and curated by medical doctors. *Strategy three* is a combination of factors, "including, but not limited to, symptom occurrence frequency rates, country-specific disease incidences and feedback loops from specific user sessions" [34]. All three strategies factor in the assessment performed by Symptoma. Martin et al.[34] tested the performance of symptoma using medical cases from the British Medical Journal (BMJ) were sourced and transcribed by medical experts into positive and negative symptoms, risk factors and patient related metadata. A subset with high similarity to COVID-19 was selected. A complementary set of COVID-19 cases were compiled from literature and computer generated by permutating a list of COVID-19 symptoms. The analysis of Martin et al. based on this testset showed an accuracy of 96.36%.

2 Motivation

Importance of Tracking COVID-19 Cases

The global outbreak of the COVID-19 pandemic has presented unprecedented challenges to public health systems worldwide [34]. Tracking the spread of the virus and accurately identifying infected individuals in a timely manner is crucial for effective containment and mitigation strategies[1]. In this thesis, we aim to explore the potential of utilizing user data obtained from Symptoma, a widely used symptom checker, to represent the transmission dynamics of COVID-19 by comparing it with the time series of confirmed COVID-19 cases.

Leveraging Symptoma's User Data

Symptoma, a widely used symptom checker application, collects data from users worldwide who voluntarily provide symptoms that could indicate a possible COVID-19 infection. This data collection provides a unique opportunity to analyze the symptoms reported by individuals and explore potential associations with the virus. Symptoma's backend, curated by medical professionals, generates a risk assessment based on the provided symptoms. By comparing this user data with the time series of confirmed COVID-19 cases, we can gain insights into the usefulness of Symptoma's symptom-based approach and assess its alignment with conventional surveillance methods.

Comparative Analysis of Time Series Methods

To achieve our research objectives, we will employ various methods for comparing time series data, including the Pearson correlation coefficient (PCC), dynamic time warping (DTW), time warped edit distance ($TWED$), and the Euclidean distance (L_2 $norm$). These techniques allow us to quantify the similarity or dissimilarity between the time series of confirmed COVID-19 cases and the data obtained from Symptoma. By applying these methods, we aim to uncover patterns, trends, and potential connections between the Symptoma user data and confirmed COVID-19 cases, as well as potential limitations of the approach.

2.4 Research Contributions

This study contributes in two main ways: First, it examines the alignment between the Symptoma users and the confirmed COVID-19 cases captured by traditional surveillance systems. This analysis can shed light on the effectiveness and challenges of symptom-based screening approaches and their potential integration into existing surveillance strategies.

Second, we explore and compare different time series comparison methods, namely the Pearson correlation coefficient (PCC), dynamic time warping (DTW), time warped edit distance ($TWED$), and the Euclidean distance (L_2 $norm$). By evaluating these techniques, we can determine their suitability for analyzing COVID-19-related time series data and identify which method(s) yield the most meaningful insights.

Through these contributions, this thesis aims to provide practical insights into the challenges of using readily available online systems to inform decision-making processes e.g. for early detection efforts.

3 Related Work

The use of web based resources for disease surveillance is not a new concept. This chapter highlights a selection of work related to ours, published in the last 20 years between 2004 and 2024.

In their study named "Analysis of Web access logs for surveillance of influenza", published in 2004, the authors Johnson et al.[23] performed a correlation analysis between the number of users visiting the website Healthlink and traditional surveillance data from the Centers for Disease Control and Prevention (*CDC*). They reported strong correlations between 0.70 and 0.80, but a large variation on timeliness, which is essentially a time series lag.

In 2006 Gunther Eysenbach published his work about tracking influenza cases using a Google Adsense campaign with keywords like "flu" and "flu symptoms" [11]. He reported a *PCC* of $r = 0.91$ between the data he collected and new influenza case counts the week after. His method showed a higher correlation than the traditional method of influenza like illness reports from sentinel physicians (*ILI-SPR*) (0.75). According to our knowledge he coined the term "Infodemiology" as a framework for data collection in context of digital pandemic, epidemic and disease surveillance. Some of the following studies classify themselves under the term Infodemiology.

Polgreen et al. published their study "Using Internet Searches for Influenza Surveillance" [47] in 2008 where they compared 2 years worth of search data from Yahoo search query logs with two different sets of traditional surveillance data. The first set contained reports of clinical laboratories that were members of World Health Organization (*WHO*) Collaborating Laboratories or the National Respiratory and Enteric Virus Surveillance System. The second set contained weekly mortality reports attributed to Pneumonia and Influenza from the 122 cities mortality reporting system operated by the *CDC*. The authors fit a linear model to both datasets and found a peak in searches 4-6 weeks before an increase in mortality. They reported an average R^2 of 0.3041 with a lag of 5 weeks between the time series and a variation between 0.4250 and 0.1227.

In 2014 Lazer et al. published their article "The Parable of Google Flu: Traps in Big Data Analysis"[27] in Science. The authors gave an overview over the limitations of influenza surveillance using the software Google Flu Trends (*GFT*). They highlighted the fact that GFT tended to overestimate flu epidemics, at worst by more than 50% and required additional input. An improved model was suggested that combined GFT and CDC data, which reduced the Mean absolute error (*MAE*) from 0.486 (*GFT*) to 0.232 (GFT + CDC). In their discussion they addressed further limitations of such a system. External manipulation via bot nets could lead to false predictions. Internal manipulation via changes to the search algorithm could lead to a change in search volume for related terms. They argued that the GFT model would be impacted by such changes and attacks since it was based on the relative prevalence of search terms. For us this work provided a great source for insights into the limitations of digital surveillance systems and and how to overcome them.

"Global Disease Monitoring and Forecasting with Wikipedia" by Generous et al. [14] was published at the end of 2014. Due to limitations of their data they selected 7 diseases and 9 countries to form 14 disease-location pairs. Each of these pairs corresponds to a disease related Wikipedia page in the language spoken in that location. Wikipedia access logs were collected for these pages and used to build linear models and tested their effectiveness in nowcasting,

forecasting and anti-forecasting. The latter was included because predicting a close past can sometimes be faster than publication of actual results. Their models show a large variance between $r^2 = 0.22$ and $r^2 = 0.92$. The results were divided into three classes. 8 cases were classified as successful. The r^2 of these results varies between 0.69 and 0.92 with a mean of 0.83. The timing of the best forecast varies between -9 days and +28 days. Failure of the remaining 6 cases were attributed mostly towards a subtle signal-to-noise ratio in the Wikipedia data or patterns in the official data being too subtle to capture.

Published in January 2020, the authors Joshi et al. used the social media platform Twitter to built an early detection system for acute disease outbreaks [24]. They proposed a system with 4 steps. The first step pre-filters the Twitter posts based on the time period and location. Additionally a keyword based filter was applied, which yielded three datasets: posts that contain "cough", "breath" and name of the illnesses under research: thunderstorm asthma and hay fever. These datasets were labeled as "cough", "breath" and "other". Step 2 was designed to filter out messages that could not be classified as personal health report. Two strategies were employed. (1) The heuristic and keyword rule-based approach which determines personal health mention based on certain phrases in the context of illness names, "cough" and "breath". (2) A statistical approach that uses an embedding for each tweet obtained by averaging over the GloVe vectors of words contained in the tweet. These tweet vectors were used to train two support vector machine (SVM) based classifiers. During step 3 everything but the first positively classified self health report per day per user is removed. This step operates under the assumption that users issue meaningful self health reports only once per day. For step 4 the authors implemented a time-between-events based monitoring algorithm for syndromic surveillance to flag outbreaks. Their results showed that the heuristic classifier didn't return any alerts. The "other" dataset only produced alerts that were late, while "breath" produced more relevant alerts when compared to "cough". Their work highlights valuable insights into building a digital outbreak surveillance system based on free text data.

A. Mavragani published their work "Tracking COVID-19 in Europe: Infodemiology Approach"[36]. In their work they use the PCC to show the correlation between country-level COVID-19 data obtained from Worldometer for 5 European countries and Google Trends data for the topic "Coronavirus". For Italy they looked at the county level data as well. They report a decrease in correlation when the disease was widespread already and suggest news and media fatigue as a possible explanation. A suggested solution is to only use their method of monitoring until the correlation declines. Conceptually this work shows great similarities to our work[64].

Similarly Higgins et al. published their study "Correlations of Online Search Engine Trends With Coronavirus Disease (COVID-19) Incidence: Infodemiology Study"[21] where they researched similarities between Google Trends data, Baidu Search Index data and real world COVID-19 case data obtained from the WHO dashboard. The authors used the PCC to show correlations between search engine data and real world data. The former was partitioned

into searches for COVID-19 related keywords, including symptoms. Their methods for time series analysis include fitting auto-regressive integrated moving average (*ARIMA*) models to the individual search volumes and real world data, assessing "the correlation between explanatory and dependent time series"[21] with sample cross-correlation functions (*CCF*) and lastly the calculation of lags "by comparing asynchronous [...] and synchronous cross-correlations"[21]. The authors report strong correlations ($r > 0.50$) between most of the search terms and real world data. They show that search engine data predates real world cases and suggest reporting bias as one of the possible causes.

In June 2022 Li et al. published their study[30] on correlations between data from the Chinese Center for Disease Control (*China CDC*) and online media sources, like search data, news articles and microblogs. Different combinations of these resources were used to build models. They used a lagged *PCC* to measure the similarity between real world cases and the models with a maximum lag of 20 days. The best performing model showed a correlation of 0.960.

Most of the aforementioned studies investigated data directly related to the pandemic. In their 2022 study "Estimating time-series changes in social sentiment @ Twitter in U.S. metropolises during the COVID-19 pandemic"[48] the authors R. Saito and S. Haruyama used a slightly different approach. They collected tweets that were posted in a 25-mile radius around the town halls of the cities New York City, Los Angeles and Chicago. The transformer models GPT-3 and BERT were fine-tuned using labeled tweets from the Sentiment 140 dataset. Sentiment was modeled via an index of -1, 0 and 1 representing negative, neutral and positive emotions in context of a set of keywords. These words were specifically chosen to not relate to COVID-19 to capture a more general sentiment. During validation BERT and GPT-3 achieved 72.4% and 81.0% accuracy on the collected dataset. The authors decided to set all tweets with a probability estimation result lower than 0.70 to neutral because of this validation result. They obtained COVID-19 case counts from local news and calculated the correlation between spread of infection and sentiment index. For New York City and Los Angeles their models show a correlation greater than 0.70 and up to 0.97 with a lag of 2 weeks. The models underperformed on Chicago data where a strong correlation could not be shown. A decline in negative sentiment could be observed as a long-term trend. The authors argued that relaxed restrictions, vaccinations and a decreased lethality of mutant strains could be a cause.

2023 Stolerman et al. published their work on building a real-time, county-level early warning system for COVID-19 outbreaks in the United States[54]. The authors collected real world case data from a database maintained by Johns Hopkins University (*JHU*). As digital data-streams they selected the following systems: (1) Google Trends with keywords like "Covid", "Covid19", "Covid 19 WHO", among others; (2) posts obtained from Twitter API filtered by location and keywords; (3) Apple Mobility requests for directions in areas of interest and (4) UpToDate, a medical resource with curated information for physicians. The authors narrowed the number of counties from 3006 down to 97 based on population. They developed three main classes of models: (1) the Naive model issued an alarm whenever the reproduction number (R_t) increases; (2)

Single Source models, which track one data-stream (i.e. the keyword "Covid19 from Google Trends"); (3) one Multiple Source model which combines different sources and issues a warning based on a threshold. The Naive model produced a large number of false positive alarms. Using Single Source models nearly cut the number of false positive results in half. The Multiple Source model showed a false discovery rate of 0.28 and displayed successful early and real-time warnings in 78% of cases.

Deiner et al. published their Infodemiology study[8] in March 2024. They used the Large Language Models GPT-3.5 Turbo and GPT-4 to assess tweets for signs of Conjunctivitis. Those automated assessments where compared to two human raters. Each tweet got assigned a percentage value, representing the likelihood of it being a Conjunctivitis outbreak. Both models showed a PCC between 0.53 and 0.77 when compared to the human raters. When compared to real world data the tweet volume assessed as Conjunctivitis related did not show a correlation. The authors note that reporting on that disease did change, which could explain those results.

In their 2024 study "Infodemiology of Influenza-like Illness: Utilizing Google Trends Big Data for Epidemic Surveillance"[51] the authors Shih et al. compared the models $ARIMA$, Long short-term memory ($LSTM$) and multiple linear regression (MLR) for forecasting Influenza like illness (ILI) outbreaks using keyword search data from Google Trends, reported cases of infection by the Taiwan Center for Disease Control and climate data from the Environmental Protection Agency of the Executive Yuan. They report that the MLR model using Google Trends data, temperature and humidity, performed best overall based on root mean square error ($RMSE$). The $ARIMA$ model achieved the highest R^2 with the same input. The $LSTM$ model outperformed the other models in cases of sudden and sharp increases in infection counts.

4 Time Series

Hamilton and Douglas describe time series, the basis of their book "Time Series Analysis"[19] as dynamic consequences of events over time. They provide a linear first-order difference equation (12) to model such a dynamic system.

$$y_t = \phi y_{t-1} + w_t \tag{1}$$

Equation (1) defines the sequence of values y at time t as being dependent on its predecessor y_{t-1}, scaled by a constant factor ϕ. w_t is a time dependant component, which further influences the signal. Serra et al. [50] simply describe a time series as "Observations that unfold over time [...]" and lists "financial data [...], medical data [...], computer data [...], or motion data [...]" as examples. Another perspective is provided by Wang et al., who describe time series as "a sequence of pairs $T = [(p_1, t_1), (p_2, t_2), ..., (p_i, t_i), ..., (p_n, t_n)]$, where $(t_1 < t_2 < ... < t_i < ... < t_n)$, where each p_i is a data point in a d-dimensional data space, and each t_i is the time stamp at which the corresponding p_i occurs" [59].

A time series can generally be defined as a time ordered sequence of quantities. Each element is usually, but not necessarily equidistant from its neighbours. This work focuses on discrete time series, which are usually obtained by sampling an analogue signal at fixed or variable intervals. In contrast, continuous time series are represented as functions with at least one parameter being related to time. Sampling can mean to either take the exact quantity at the current point in time or to apply aggregation to all values within the time step. If the intervals of sampling are fixed and equidistant, then the series is sampled using a constant sampling rate.

4.1 Time Series preprocessing

Preprocessing can be achieved in either the time domain or the frequency domain. Time series preprocessing in the time domain tries to adjust one or multiple of {1} quantity value, {2} sampling rate, {3} series length so that the data fits the parameters of an analysis or model. As an example, in case of log-normal data the log transformation can be used to adjust quantity values {1} to closer resemble a normal distribution. This transformation is performed by applying the logarithm to each quantity, resulting in a new time series of equal length, but with lower amplitudes.

In the time domain, each value is directly related to a point in time. For example the time of measurement. Time domain adjustments of quantity values {1} generally involve normalization or standardization methods. Those methods are used to scale the data so that the values fall into a common range, usually $[0, 1]$ or center and scale the data to achieve a mean of 0 and a standard deviation of 1. Paparrizo et al. [45] gave a good overview of commonly used methods. A subset of those methods was employed in this work and will be listed here. A popular normalization method is the min-max normalization[45] specified in equation (2).

$$\vec{x}' = \frac{\vec{x} - min(\vec{x})}{max(\vec{x}) - min(\vec{x})} \tag{2}$$

The functions $min()$ and $max()$ are defined as follows:

$$min(\vec{x}) = x_i;\ x_i \leq x_j\ \forall j \neq i;\ j, i \in [1, ..., n] \tag{3}$$

$$max(\vec{x}) = x_i;\ x_i \geq x_j\ \forall j \neq i;\ j, i \in [1, ..., n] \tag{4}$$

The *min-max normalization* maps the lowest value of a series to 0 and the highest value to 1. Values in between are scaled accordingly. Depending on the analysis it might be preferable for the series to not contain any zeros. In that case one can adjust equation (2) so that the range, values will be mapped to, can be chosen freely:

$$\vec{x}' = a + \frac{\vec{x} - min(\vec{x}) \cdot (b - a)}{max(\vec{x}) - min(\vec{x})} \tag{5}$$

Equation (5) leads to the values x_i of $\vec{x}$ to be adjusted as follows, $a \leq x_i \leq b\ \forall$ $x_i \in \vec{x}'$.

The most used method according to Paparrizo *et al.* [45] is the z-normalization:

$$\vec{x} = \frac{\vec{x} - mean(\vec{x})}{std(\vec{x})} \tag{6}$$

where functions $mean()$ and $std()$ are defined as follows:

$$mean(\vec{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{7}$$

$$std(\vec{x}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - mean(\vec{x}))^2} \tag{8}$$

Applying z-normalization leads to the values over a given time series being centred around its average value. Everything is then shifted so that the mean is zero. The *mean normalization* is the result of combining parts of other methods. As shown in (9) the numerator is based on the z-normalization (6), while the denominator is based on the min-max normalization (2)

$$\vec{x}' = \frac{\vec{x} - mean(\vec{x})}{max(\vec{x}) - min(\vec{x})} \tag{9}$$

Figure (2) showcases the methods listed above when applied to the Symptoma high risk COVID-19 cases and the newly infected cases of the Johns Hopkins University for the country Czechia.

To achieve a uniform sampling rate $\{2\}$ either aggregation or resampling methods are used. As stated before, aggregation involves combining values inside a range via a function to get one value for that range. Some of these methods have a smoothing effect on the data and are used to remove higher frequencies.

A popular aggregation method is the *centered mean* defined as:

$$centered\_mean(x_i) = \frac{1}{2r+1}\sum_{k=-r}^{r} x_{i+k} \tag{10}$$

where r describes the radius of the aggregation window. Similarly one can extend the window only in one direction:

$$rolling\_mean(x_i) = \frac{1}{w+1}\sum_{k=0}^{w} x_{i+k} \tag{11}$$

where the parameter w specifies the number of time steps included in the calculation. Both methods have to account for undefined values of x_{i+k} at the edges of the time series. In such cases parameters r and w can be reduced at both ends so that the window only includes defined values. Another strategy is to temporary append values to both ends. The quantity of those values depends on the data. The number 0 is a popular choice.

Resampling methods usually fall into one of the following two categories. Increasing the sampling frequency is called upsampling, while decreasing it is called downsampling. The usefulness of either method depends on the domain of the analysis. Downsampling simply removes values from the time series and sometimes shifts the time index to preserve a uniform distance between samples. If enough data is available, one can simply resample the dataset with increased, equal spacing between points. Upsampling follows the same process, if the conditions similarly allow for it, but with reduced equal spacing instead.

Preprocessing in the frequency domain uses the discrete Fourier transform (DFT) (12) to determine the frequency spectrum $D_k(z)$ of a discrete signal z with length n [2].

$$D_k(z) = \sum_{j=0}^{n-1} z_j e^{\frac{-2\pi ijk}{m}} \tag{12}$$

"In the early 1900s, the French mathematician Jean-Baptiste Fourier showed that any composite signal is actually a combination of simple sine waves with different frequencies, amplitudes, and phases."[12]. Those components can also be represented by complex values, instead of sine waves in the form of $a + bj$, where a is the real part and bj is the complex part and j is defined as $j = \sqrt{-1}$. The DFT enables us to extract the coefficients of those components, which is the frequency spectrum. For example "A complete sine wave in the time domain can be represented by one single spike in the frequenc domain." [12]. This process is reversible using the inverse of the discrete Fourier transform (iDFT) (13).

$$D_k^{-1}(z) = \frac{1}{m} \sum_{j=0}^{n-1} z_j e^{\frac{2\pi ijk}{m}} \tag{13}$$

Preprocessing in the frequency domain entails converting the time series into its components, operating on the resulting coefficients and using the iDFT to convert the result back into a time series. One example of preprocessing in the frequency domain are frequency based filters. In their simplest form, one can set the coefficients for the corresponding frequencies to 0. If high frequency coefficients are changed that way a smoothing effect is achieved.

4.2 Distance measures

Measuring distances between time series is a way to represent similarity.

Euclidean Distance One of the most common and simple methods used is the *euclidean distance*. To calculate it the L_n norm, denoted by d_{L_n} in the following definition (14) can be used. Specifically, the *euclidean distance* is equivalent to L_2.

$$d_{L_n}(x, y) = \left(\sum_{i=1}^{M} (x_i - y_i)^n \right)^{\frac{1}{n}} \tag{14}$$

x and y denote two time series of length M, whose respective value at index i is denoted by x_i and y_i [50]. The *euclidean distance* can be calculated in the frequency domain as well:

$$d_{FC}(x,y) = \left(\sum_{i=1}^{\theta} (\hat{x}_i - \hat{y}_i)^2 \right)^{\frac{1}{2}} \tag{15}$$

where $\hat{x}_i$ and $\hat{y}_i$ represent the complex value pairs of the i-th Fourier coefficient. θ can be set to $\theta = M/2$ because of the symmetrical nature of the Fourier coefficients. Since θ acts like a window, values smaller than $M/2$ can be used to exclude high-frequency components [50]. Simply put, the euclidean distance in the time domain incorporates a smoothing effect that can be adjusted using θ.

Formula (14) and (15) show that values corresponding to the same index are used for comparison. Methods like that are called *lock-step measures* [59]. These measures work well in the case of events influencing both time series at the same time. If one series lags significantly behind the other, lock-step methods would fail to discover similarities. To construct a toy example consider a one-way train track with two sensors designed to detect passing trains. If the sensors are far enough apart so that a train triggers one significantly later than the other then lock-step measures would consider the resulting time series as different, they measure the same events just with a bit of lag.

Sliding Distances A category of distance measures with the ability to detect such lagged similarities in time series are so called *sliding distances* [45]. They operate by shifting, an operation $\vec{x_{(s)}}$ which rearranges the data points of series $\vec{x}$ by moving all points by $|s|$ positions[59].

$$\vec{x}_{(s)} = \begin{cases} (0, ..., 0, x_1, x_2, ...x_{m-s}), & s \geq 0 \\ (x_{1-s}, ..., x_{m-1}, 0, ..., 0), & s < 0 \end{cases}$$

$s \geq 0$ indicates a shift to the right, meaning data points are shifted towards the future, while $s < 0$ indicates a shift to the left.

Cross-Correlation All shifts $s \in [-m, m]$, where m is the length of both series ($\vec{x}$ and $\vec{y}$), produce the *cross-correlation sequence* $CC_w(\vec{x}, \vec{y})$ with $w \in \{1, 2, ..., 2m-1\}$ containing the inner product of the two time series in every possible shift. This sequence contains the best shift at position $i \in \{1, 2, ..., 2m-1\}$, which fulfills $CC_i(\vec{x}, \vec{y}) = max(CC_w(\vec{x}, \vec{y}))$. *Cross-Correlation* in the time domain is an expensive operation. Fortunately due to the Fast Fourier Transform (FFT) [2] and its inverse ($F()$, $F^{-1}()$) we can perform this operation in $\mathcal{O}(m \cdot \log(m))$ instead of $\mathcal{O}(m^2)$ [59] with equation (16).

$$CC_w(\vec{x}, \vec{y}) = F^{-1}\left\{F(\vec{x}) * F(\vec{y})\right\} \tag{16}$$

normalized versions of equation (16) exist as normalized cross-correlation (NCC) assuming some underlying time series normalization [44, 45]:

$$NCC_b(\vec{x}, \vec{y}) = max\left(\frac{CC_w(\vec{x}, \vec{y})}{m}\right) \tag{17}$$

$$NCC_u(\vec{x}, \vec{y}) = max\left(\frac{CC_w(\vec{x}, \vec{y})}{m - |w - m|}\right) \tag{18}$$

$$NCC_c(\vec{x}, \vec{y}) = max\left(\frac{CC_w(\vec{x}, \vec{y})}{||\vec{x}|| \cdot ||\vec{y}||}\right) \tag{19}$$

NCC_b from equation (17) is the biased estimator, where the result vector of the cross-correlation is normalized by m, the length of both time series $\vec{x}$ and $\vec{y}$. The unbiased estimator NCC_u (eq. (18)) normalizes the result vector by the difference of series length m and $|w - m|$, which is essentially the absolute value of the respective shift $|s|$. This introduces a penalty for shifts close to 0. Lastly there is the coefficient normalization NCC_c from equation (19)), which normalizes by the product of the norms of the two time series. The normalized versions of the *cross-correlation* makes it easier to compare time series with different magnitudes.

Dynamic Time Warping In contrast to *sliding distances* and *lock-step measures*, *elastic similarity measures* [59] try to optimize an alignment between time series based on the cost of said alignment [50]. This flexible mapping between indices of time series is an effective way to account for gaps and unevenly sized data. *Dynamic time warping* (DWT) [49] is a classic elastic distance measure. Its methods are comparable to the algorithms called Needleman-Wunsch [40] and Smith-Waterman [53], used to compute global and local sequence alignments in biochemical sequences using dynamic programming. These algorithms work on string sequences and find the best mapping between two sequences or the best placement for subsequences. In the case of *elastic similarity measures* the result is a warping path $W = \{w_1, ..., w_k\}$ with $k \geq m$ [45]. First a matrix D is constructed with dimensions $M \times N$ equal to the length of each time series. Initialization of the matrix adheres to the following rules $D_{i,j} = \infty$ for $0 \leq i \leq M; 0 \leq j \leq N$ and $D_{0,0} = 0$. The cost of the alignment at each position can be obtained by recursively applying:

$$D_{i,j} = f(x_i, y_j) + min(D_{i,j-1}, D_{i-1,j}, D_{i-1,j-1}) \tag{20}$$

$f(x_i, y_j)$ is the local cost function of the alignment. In case of one-dimensional time series it can simply be defined as:

$$f(x_i, y_j) = \sqrt{(x_i - y_j)^2} \tag{21}$$

as described in [50], [3]. Zhang et al. [62] proposed a grid based cost function, called lower bound distance (LBD) in their work on the topic of vegetation

recovery after earthquakes. Values of time series X and Y are placed on a grid. If the two values x_i, y_j are in the same cell, then the distance is considered as zero. Otherwise the result of the cost function is based on the distance between the two cells. Maus et al. [35] incorporated the difference of the time index into their cost function to account for seasonal overlap. Their time-weighted extension to DTW ($TWDTW$) prevents sequences in originating from different seasons to being matched.

$$f(x_i, y_j) = |x_i - y_j| + w_{i,j} \tag{22}$$

$w_{i,j}$ denotes the weighted time difference between the indices i and j. Maus et al. give two versions for w. One is linear:

$$w_{i,j} = g(i, j) \tag{23}$$

The other one is a logistic model:

$$w_{i,j} = \frac{1}{1 + e^{-\alpha(g(i,j)-\beta)}} \tag{24}$$

where the function $g(i, j)$ returns the difference between the indices i and j in number of days. α is the steepness and β the midpoint to calculate the weighted time difference. It is possible to include other information into that calculation. This becomes useful for multi dimensional time series, where each dimension is a distinct feature of the series. An example would be time series collected from different reference points on the globe. Depending on the research question, the inclusion of geographical proximity by incorporating it into the cost function can lead to further insights.

The matrix resulting from the process described by (20) and (21) is shown in Figure (3). In a process called backtracking the best alignment in the context of the cost function (20) is determined. Starting from the bottom right most field $D_{M,N}$ a path is generated by selecting the adjacent field with the lowest score. Coordinates i, j of every field are stored. The process is continued until $D_{0,0}$ is selected. The resulting path $(D_{0,0}, ..., D_{i,j}, ..., D_{M,N})$ marks the optimal mapping of the respective elements x_i, y_j between time series x and y, the so called warping path W.

One quality of distance metrics is the triangle inequality eq. (25), which assures that results are comparable.

$$d(x, y) \leq d(x, z) + d(y, z) \tag{25}$$

x,y,z are real valued time series and $d(x, y)$ denotes the distance between them. DTW does not satisfy the triangle inequality. The following example eq. (26), provided by Marteau [33], proves this. Given time series A,B,C:

$$\begin{aligned} A = [1];\ B = [1, 2];\ C = [1, 2, 2] \\ D_{A,B} = 1;\ D_{B,C} = 0;\ D_{A,C} = 2 \\ D_{A,C} > D_{A,B} + D_{B,C} \end{aligned} \tag{26}$$

Other Elastic Measures Next to *DTW* other methods exist in this class. The Longest Common Sub-sequences (*LCSS*) measures distances between time series via the length of sub-sequences deemed as similar. This is done by introducing a parameter ϵ as a threshold to determine matching points and an additional parameter ω to constraint the warping window [45]. Selecting the right value for ϵ can be hard since this ultimately determines the outcome of the binary decision: [similar, not similar]. Originally proposed in 1966 as the Levenshtein Distance (*LD*) [29], the Edit Distance (*ED*) is used as a dissimilarity measure for character sequences. It is "[...] the smallest number of insertions, deletions, and substitutions required to change one string into another" [33]. *ED* builds on the concept of binary similarity. This works well with nominal data, were binary decisions of equality can be modeled by group theory. The similarity of the elements of real valued time series can be difficult to measure in a binary decision space. Similar to *LCSS*, the Edit Distance on Real sequence (*EDR*) uses threshold determined point wise binary similarity classification, while Edit Distance with Real Penalty (*ERP*) introduces a gap penalty based on the real valued distance between the values. Where *LCSS* determines only the longest match, *EDR* takes the whole sequence into account and introduces fixed gap penalties between matching sub-sequences. These penalties causes *EDR* and *ERP* to be similar to *DTW*.

Time Warped Edit Distance Both *DTW* and *LCSS* are not considered as metrics because they do not satisfy the triangle equality [33]. Time warped edit distance *TWED* was developed to fill the niche of a metric to measure time series dissimilarity. Another quality of *TWED* is that it fulfills the triangle inequality specified in eq. (25). Marteau describes the process of *TWED* as a time series matching game with three possible operations. Let $x'_i = (x_i, t_i)$ and $y'_j = (y_j, t_j)$ where t_i is the time stamp of position i in the respective time series. The following operations work on so called elements and sections. An element is the vector at any position $i; j$ in x'_i and y'_j. A segment is the connection between two elements, characterized by $(x'_(i-1), x'_i)$ or $(y'_(j-1), y'_j)$. Operations *delete-A* and *delete-B* remove an element from the respective time series. The cost of these operations is the length of the vectors $(x'_i - x'_{i-1})$ or $(y'_j - y'_{j-1})$ plus a constant $\lambda \geq 0$. The *match* operation moves a segment of A so that it overlaps with a segment of B. Marteau suggests that the cost associated with the *match* operation is proportional to the sum of lenghts of the vectors $(y'_j - x'_i)$ and $(y'_{j-1} - x'_{i-1})$. Magdy *et. al* describes it as "the sum of lengths of two vectors connecting the start and end of both segments" [31]. In the context of dynamic programming this translates to the following recursive function:

$$D_{i,j} = min \begin{cases} D_{i-1,j} + d(x'_i, x'_{i-1}) + \lambda & delete\text{-}A \\ D_{i-1,j-1} + d(x'_i, y'_j) + d(x'_{i-1}, y'_{j-1}) & match \\ D_{i,j-1} + d(y'_j, y'_{j-1}) + \lambda & delete\text{-}B \end{cases} \tag{27}$$

$d(\vec{a}, \vec{b})$ is the length of the two vectors $\vec{a}$ and $\vec{b}$. One can further parameterize eq. (27) with a separate factor γ to control the stiffness in the time domain.

The constraint $\gamma > 0$ assures that this version is still considered a distance. Both parameters are dependent on the dataset and can be learned using cross validation[33, 31].

$$D_{i,j} = min \begin{cases} D_{i-1,j} + d_{LP}(x_i, x_{i-1}) + \gamma \cdot d_{LP}(t_i, t_{i-1}) + \lambda & \textit{delete-A} \\ D_{i-1,j-1} + d_{LP}(x_i, y_j) + \gamma \cdot d_{LP}(t_i, t_j) + & \\ d_{LP}(x_{i-1}, y_{j-1}) + \gamma \cdot d_{LP}(t_{i-1}, t_{j-1}) & \textit{match} \\ D_{i,j-1} + d_{LP}(y_j, y_{j-1}) + \gamma \cdot d_{LP}(t_j, t_{j-1}) + \lambda & \textit{delete-B} \end{cases} \tag{28}$$

$d_{LP}(\vec{a}, \vec{b})$ is the LP-norm of vectors $\vec{a}$ and $\vec{b}$. Both versions eq. (27) and eq. (28) contrast the standard version of DTW by accounting for temporal differences as well. The following example repeats eq. (26) in the context of $TWED$ with equally spaced time indices.

$$\begin{aligned} A = [(1,1)];\ B = [(1,1),(2,2)];\ C = [(1,1),(2,2),(2,3)] \\ \lambda = 1.0;\ \gamma = 0.001 \\ D_{A,B} \sim 2;\ D_{B,C} \sim 3;\ D_{A,C} \sim 1 \\ D_{A,C} \leq D_{A,B} + D_{B,C} \end{aligned} \tag{29}$$

This shows that at least for this specific case the triangle inequality holds. Marteau provides the complete proof in [32].

5 Data Description

The COVID-19 pandemic led to a worldwide effort in tracing newly infected individuals. The resulting country wise time series are the basis for this work. We chose a period of one year starting from 1st of May 2020 to use the most recent data at the time and to cover a large period. In the following subsections of this chapter, we will present the two main data sources for this thesis. Additionally the steps of and reasoning for the preprocessing will be given.

5.1 Johns Hopkins University

COVID-19 infections get aggregated by the respective administrative regions and usually released via public channels. The Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (*JHU*) created a Github Repository [1] [56, 9] in February 2020 to aggregate COVID-19 cases globally [56]. Figure (4) illustrates part of the content of that file, while Table 1 gives an overview of the columns. As requested by the terms of use of this data set it will hereby be called *JHU CSSE COVID-19 Data*. For readability reasons we will use the abbreviations *JH* or *JH counts* in plots to save space.

[1] https://github.com/CSSEGISandData/COVID-19

Field	Format	Description
Province/State	String	Subdivision of the country. For example States. Not available for all Countries.
Country/Region	String	Name of the country.
Lat	Float	Latitude. Part of the geo coordinate for the center of the country.
Long	Float	Longitude. Part of the geo coordinate for the center of the country.
$m+/d+/yy$	Integer	Following columns contain the cumulative COVID-19 cases up to the date denoted by the column name, starting from 22^{th} of January 2020. $m+$: month without leading zero. $d+$: day without leading zero. yy: last two digits of the year.

Table 1: Columns of *"time_series_covid19_confirmed_global.csv"*

The full list of sources for *JH* can be found in the associated repository [2]. Though it has to be stated that the list is incomplete. At least the source for Austria is missing. The degree of overlap between the sources and the countries included in the data set has not been investigated in detail. Another problem with the data is modifications. The CSSE gives a detailed list of all corrections in the repository [3]. For example delayed reporting and errors in the source data.

On the 1st of March 2020 the passengers of the cruise ship Diamond Princess were counted as a separate entity despite having US citizenship. Between April 4th and 11th only confirmed cases were counted for France. This was changed so that starting from April 12th 2020 possible cases are included as well. On May 27th 2020 the Netherlands stopped reporting recovered cases, which led to the removal of all recovered cases from the Netherlands dataset. The case data for Belarus for April 18th and 19th 2020 had to be updated on June 9th due to late reporting. On the 21st of january 2021 the deaths of Sweden had to be altered to be consistently "death by date of report" in contrast to "by date of death". On the 5th of November 2021 aggregation of japanese cases stopped in favor of the cases reported by NHK due to their reporting standards. This is but a fraction of the corrections that were applied to the dataset. In some cases historical data is changed, in other cases reporting policy and standards change. One must keep these changes in mind when conducting experiments using this data and assure reproducibility of results and comparability of time series produced with different underlying rules.

[2] https://github.com/CSSEGISandData/COVID-19

[3] https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data

5.2 Symptoma

Symptoma GmbH is a digital healthcare company. One of its products is the online symptom checker *Symptoma* [34]. Users accessing *Symptoma* are presented with a field for free form text input of their symptoms, as can be seen in Figure 5. They can be provided directly as a comma separated list or via answering the questions *Symptoma* presents. The column *possible causes* shows the medical concepts, i.e. diseases, strongly associated with the provided symptoms. Figure 6 shows the session overview. On the left side one can see the provided input and answered questions. The middle highlights the reported symptoms by category symptom, negated and unsure. The right side is reserved for possible causes identified by the symptoms provided. Symptoms and possible causes are part of the broad category medical concept, which we shorten to concept. For this work the top 5 possible causes are considered as "*high risk*". This means that the risk of having the disease given the provided list of symptoms is assessed as *high*. User sessions are persisted in our database and contain a *session_id* uniquely identifying the User, up to 30 medical concepts associated with the symptoms and the origin of the request on a country level, determined via the service *geo-ip*. The *session_id* is only recorded if the user agrees to be tracked. Neither ip-address nor uniquely identifiable browser features are collected. Thus it is impossible to trace input back to the user, which classifies this data set called *Symptoma data* as anonymous user data.

Aggregation of the *Symptoma data* based on the date of interaction with the system (*date*), the iso-2 code identifying the country origin (*countryCode*) and the internal identifier of the medical concept in question (*concept_id*) compile to daily records for each country. Using the *concept_id* for COVID-19 as a filter returns a statistic of daily sessions per country containing COVID-19 in the list of possible causes of that session. The number of sessions assessed as having a *high risk* for COVID-19 is determined by filtering for sessions that have COVID-19 in the top 5 possible causes. The end result is a list of records containing the fields summarized in Table 2. Filtering by *countryCode* returns the time series

Field	Format	Description
date	DateTime	date associated with the record
countryCode	String	iso-2 code of the country associated with the record e.g.: at for austria.
count	Integer	Number of requests containing COVID-19 as possible cause
uniqueUsers	Integer	Number of unique user sessions containing COVID-19 as possible cause.
highRisk	Integer	Number of requests containing COVID-19 in the top 5 of possible causes.

Table 2: Fields of the records in the aggregated *Symptoma data*

for each respective country. Certain combinations of *date* and *countryCode* are

missing in this data set. This happens if during that specific day no user request originating from that country was recorded. Occurrences like that are referred to as gaps in the data and are dealt with during preprocessing. From here on the time series resulting from the mapping of the fields *date* and *highRisk* are referred to as *Symptoma data*.

5.3 Preprocessing

JH contains cumulative cases of COVID-19 infections. *Symptoma data* is more comparable to daily new infections since user sessions can be seen as a self report test. To make both data sets comparable we computed the daily difference of the cumulative cases in *JH* for each country separately. Figure (7) shows the time series for *Germany*, *Switzerland*, *United Kingdom* and *Austria*. As described in Chapter 5.2, there are gaps in the *Symptoma data*. They were filled in with zeros as part of the preprocessing to represent the fact that no user of *Symptoma* from the relevant country produced a *high risk* result with respect to COVID-19 on that day. This leads to both data sets having the same sampling rate and type of data. The magnitude of the two time series are not comparable. For most countries the number of users cannot be directly compared to the amount of COVID-19 cases. Z-normalisation was used to account for the difference and to make countries with drastically different numbers of users comparable. The added benefit is that raw user counts, which are considered sensitive company data, are encoded in a meaningful way. Figure (8) shows the result of the z-normalisation (eq. 6) on both time series for Austria. As illustrated by Figures (8) and (7) *JH* shows a periodic pattern repeating roughly every 7 days. Additionally daily variations introduce a lot of noise in both time series. A centered mean with window size $w = 7$ was used to smooth out the variations and the repeating weekly pattern, concluding the preprocessing. The resulting two time series *Symptoma data* and *JH* for Austria are shown in Figure (9).

6 Methods

In this section we discuss the methods used for our analysis. All of our chosen methods in regards to the time series analysis are either well established or based on well studied concepts. We give more specific reasons for each method below. This study is centered around the comparison between the COVID19 data set provided by Johns Hopkins University (*JH*) and the user data of Symptoma related to COVID-19 (*Symptoma data*). The former is well understood and published, while the latter requires analysis to ascertain its value in the domain of pandemic data collection and infectious disease monitoring. We chose well understood methods to eliminate as much uncertainty as possible from our study and establish a baseline for Symptoma.

6.1 Concept independence analysis

The aim of this study is to investigate if online symptom checkers are able to capture disease related data via user interaction with the system. In theory, if the amount of users is high enough, then the system should capture enough information to be comparable to a testing strategy. The current situation provides a good opportunity to test this theory. Because of the pandemic, digital health systems gained a lot of traction, which is true for Symptoma as well.

Others have already assessed the accuracy of Symptoma in regards to classifying reported symptoms being caused by COVID-19 [34]. An important point, which needs to be addressed is that the number of user sessions assessed to have a high risk in regards to COVID-19 might correlate only with the total number of users of the system. If this is the case then different diseases should show the same pattern. Were this true, then the answer to our question: "Is Symptoma able to capture the spread of the pandemic?", would have to be a clear no. To strengthen the results of this study, a comparison to non-COVID-19 related diseases is needed. The aim of the comparison is to show the independence of the *Symptoma data* from media exposure. If diseases with different underlying causes produce different time series then it can be reasoned that our COVID19 statistics are independent of increased media exposure and temporary increased user counts.

To ensure comparability of the selected diseases with COVID-19 the following constraints need to be taken into consideration: (1) small difference in user session counts between the two concepts, (2) large difference in symptoms. This analysis would be biased greatly and thus worthless if the magnitude of user sessions containing the diseases differ too much. Reason (2) stems from the fact that Symptoma uses symptoms to return possible causes to the user. Thus, if the list of symptoms overlaps too much, then the time series will have a small distance as well, since those concepts will be associated with the same sessions in a lot of cases. COVID-19 and Influenza are both viral diseases with symptoms related to the respiratory tract, they appear together often. The following methods were used to select diseases for comparison with COVID-19:

A list of top 40 returned diseases are collected for the country Austria, from 1^{st} of May 2020 till 1^{st} of May 2021. Because of the aforementioned symptom based similarity of other respiratory diseases to COVID-19, this list needs to be filtered. The overlap coefficient [38] defined in equation (30)

$$overlap(X, Y) = \frac{|X \cap Y|}{min(|X|, |Y|)} \tag{30}$$

is a common technique to determine the similarity of sets. A list of possible causes or symptoms can be classified as a set, since they only appear once and their order does not matter. We used the overlap coefficient to determine the symptom based distance of each disease to COVID-19. For each disease we created a set of the respective symptoms as they appear in our database and calculated the overlap score for those sets and the set of symptoms associated with COVID-19. Because of the way *Symptoma* works it is ensured that diseases

with a high overlap will have similar time series. Of the top 40 concepts we selected those that show a low overlap.

6.2 Cross-Correlation

It has been mentioned before, how lock-step measures are based on the distance between single elements of two time series. The goal of this work is to discover similarities between *Symptoma data* and *JH*. Since some of the distance measures used in this work are lock-step measures, it is important to determine the lag between the two time series for each country separately. To achieve this efficiently the cross-correlation will be used to compute optimal lags. As the task is to look for overlaps in epidemiological data, we decided that a lag of up to 30 days is acceptable. It is unreasonable to assume that Symptoma would be able to model the pandemic more than 30 days in advance and lags of more than 30 days behind *JH* are equally meaningless. Others have found similar values for lags between digital monitoring systems and real world case counts[14]. Additionally, to recreate Figure 2 of our previous work [64] we calculated the PCC for all lags between $[-30, 30]$. Ideally the curve produced by all the shifts for the respective country shows a clear global maximum and a steep slope to be considered of significant importance.

6.3 Distance measures

Lock-step measures employ an element wise approach to determine dissimilarities between sequential data. The *Symptoma data* set stems from a purely digital system, while the *JH* data set stems from multiple sources and sometimes long reporting chains. Because of the different modes of data collection element wise distances of these time series cannot be expected to be optimal. To find the optimal alignment within a reasonable window the same method discussed for the cross-correlation was used. Similarly the constraint of $[-30, 30]$ was applied for the shifts. Resulting undefined regions were filled with zeros.

The following lock-step dissimilarity measures were calculated for each shift, to determine the optimal alignment and best score between *Symptoma data* and *JH*: *pearson correlation coefficient* (PCC), *euclidean distance* (L_2*-norm*). Additionally the dynamic measures *dynamic time warping* (DTW) and *time warped edit distance* ($TWED$) were used to calculate dissimilarity scores independently from the series alignment. The results of L_2*-norm*, DTW and $TWED$ were further normalized by the time series length. These values can be understood as mean distance per day. We argue that this increases their interpret-ability.

To show the agreement between the different distance measures the PCC between their respective results was calculated. This was done with the expectation that the PCC should show a strong negative correlation towards all other measures. The L_2*-norm*, DTW and $TWED$ measure dissimilarity via the increased distance from zero, while correlation traditionally interprets this as a strong positive or negative relationship between the datasets. Further it can be

expected that the dynamic measures show a strong positive correlation, because both are designed to account for lagged features in the time series.

6.4 Software Packages and Libraries

Tools for data extraction and manipulation were implemented using the programming language Python version 3 [18]. For easier data handling the libraries Numpy [20] and Pandas [55] were used. Table 3 shows the packages used for transformations and distance measures.

Package	Usage
Pandas [55]	*PCC*
Scipy [58]	L_2-*norm* and *z-score*
pytwed[4]	function *twed()* to calculate *TWED*
edist [43]	function *dtw_numeric* to calculate *DTW*

Table 3: Software packages used for each part of this work

7 Results

7.1 Concept independence analysis

The following table shows the results of the comparison between different medical concepts and COVID-19.

disease	pcc	interval	p-value	overlap
Migraine	0.195	(0.094, 0.292)	0.000	0.080
Chronic Fatigue Syndrome	0.242	(0.143, 0.336)	0.000	0.114
Bacterial Meningitis	0.160	(0.058, 0.258)	0.002	0.136
Anxiety Disorder	0.244	(0.145, 0.338)	0.000	0.136
Meningitis	0.261	(0.163, 0.354)	0.000	0.177
Scarlet Fever	0.239	(0.14, 0.333)	0.000	0.179
Rubella	0.168	(0.067, 0.266)	0.001	0.180
Mycoplasma Pneumonia	0.041	(-0.062, 0.143)	0.434	0.227
Tuberculosis	0.076	(-0.026, 0.177)	0.145	0.239
Granulomatosis with Polyangiitis	0.204	(0.104, 0.3)	0.000	0.261
Bacterial Pneumonia	0.328	(0.233, 0.416)	0.000	0.295
Viral Lower Respiratory Infection	0.440	(0.353, 0.519)	0.000	0.300

[4] https://github.com/jzumer/pytwed

Table 4: 12 out of the top 50 medical concepts compared with COVID-19. Selected via threshold $pcc < 0.5$ to prove that the *Symptoma Pulse Data* contains unique time series for sufficiently different concepts (low *overlap* coefficient). *disease*: Name of the disease *pcc*: The pearson correlation coefficient comparing the time series of the disease to the time series of COVID-19. *interval*: The upper and lower bound of the confidence interval for the PCC. *p-value*: p-value for the PCC, $\alpha = 0.05$. *overlap*: overlap coefficient (eq. (30)) between the symptoms of the medical concept compared with the symptoms of COVID-19. The table is ordered by *overlap*.

Table 4 lists 12 concepts that show a weak correlation ($PCC < 0.5$) and additionally have a small overlap in symptoms ($overlap \leq 0.3$). 5 of these 12 concepts show a correlation smaller than 0.2, suggesting a very weak relationship at best.

7.2 Distances

Figures 10 and 11 illustrate the differences for the PCC after shifting the time series by up to 30 days. Countries like the United States (US) and Great Britain (GB) improved from a strong negative correlation to showing a slight negative correlation between the time series. Other countries show a slight improvement towards positive correlation, which leads to the median of the shifted values being 0.083 above the one without a shift.

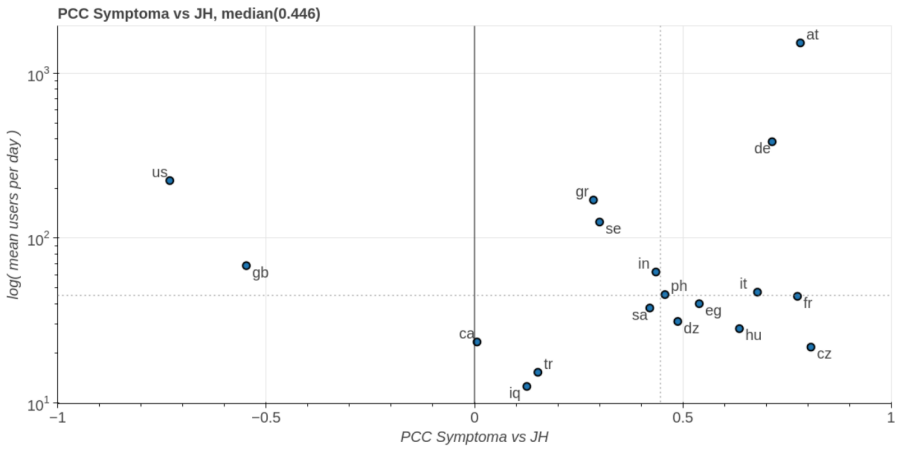


Figure 10: Pearson Correlation Coefficient

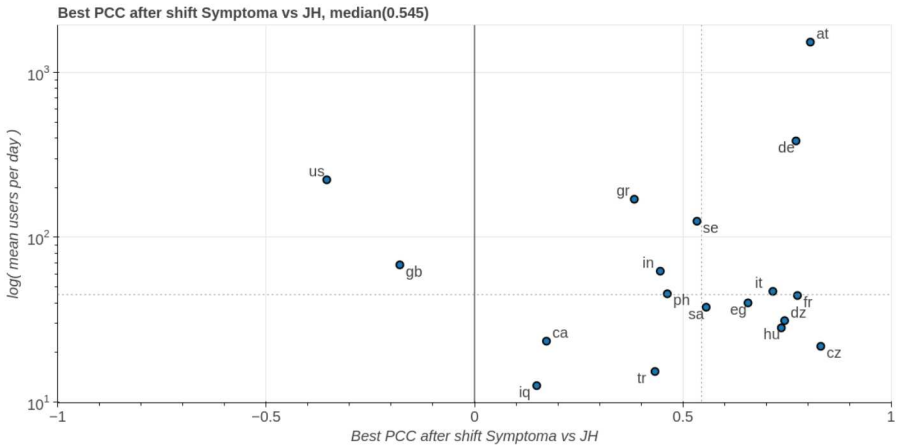


Figure 11: Shifted Pearson Correlation Coefficient

Figures 12 and 13 illustrate the differences for the L_2-*norm* after shifting the time series by up to 30 days.

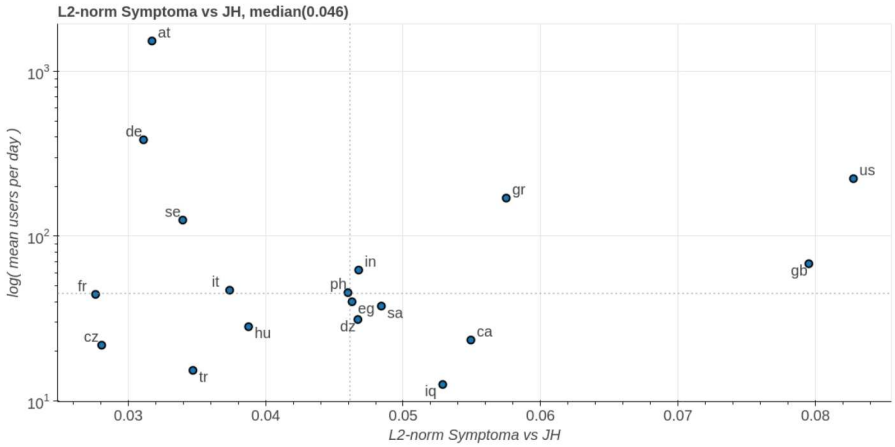


Figure 12: L_2-norm normalized by time series length (366). The dotted lines show the median for each axis respectively.

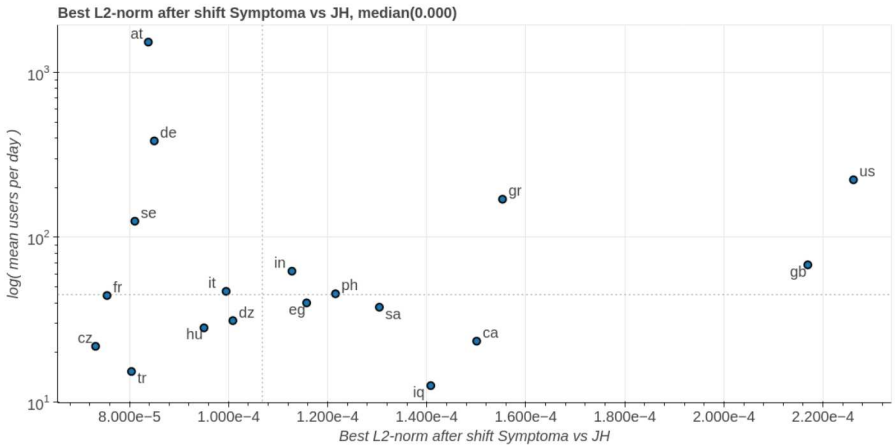


Figure 13: Shifted Euclidean Distance normalized by time series length (366). The dotted lines show the median for each axis respectively.

Figures 14 and 15 illustrate the dissimilarity scores for the dynamic distance measures *DTW* and *TWED* calculated based on the *Symptoma data* and *JH* time series.

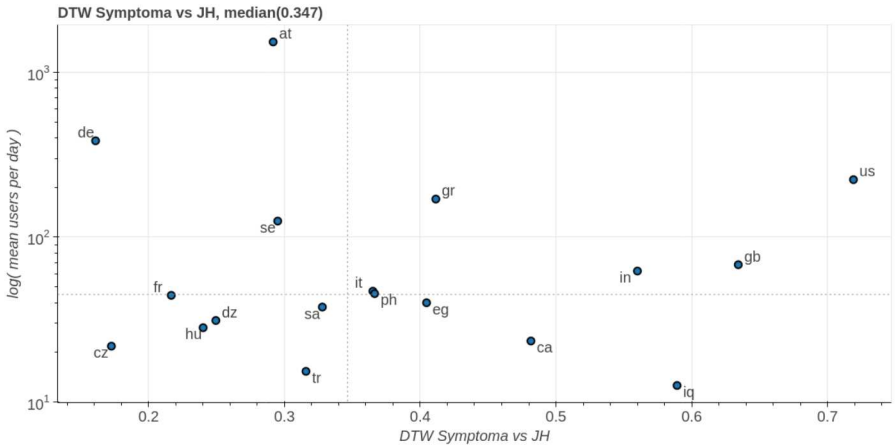


Figure 14: Dynamic Time Warping normalized by time series length (366). The dotted lines show the median for each axis respectively.

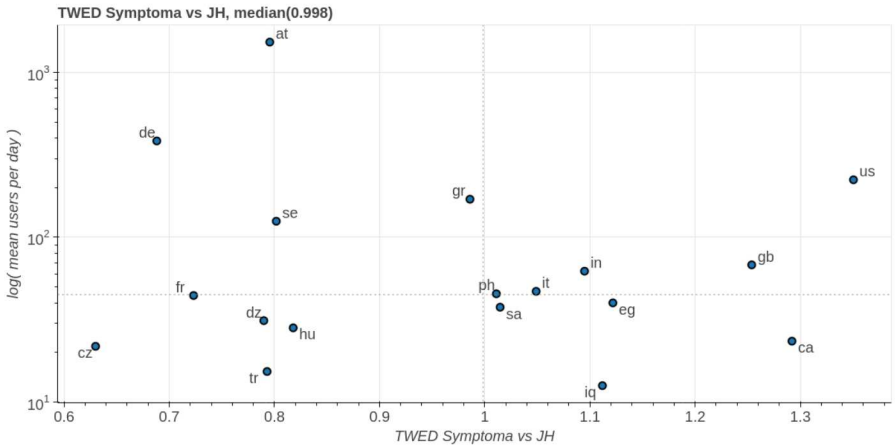


Figure 15: Time Warped Edit Distance normalized by time series length (366). The dotted lines show the median for each axis respectively.

Figure 16 show the significance of the shifts for PCC and L_2norm respectively. Both show a median shift of 4 days towards the future for the optimal alignment of *Symptoma data* and *JH*. Small shifts close to zero are always found at the positive side of the scale. The United States, Canada and Great Britain can be identified as clear outliers in both subplots due to their proximity to the border of the plot and their low score when compared to the median. Regarding only the L_2 *norm*, the Philippines and Iraq can be considered as outliers as well. The rest of the countries form a relatively dense cluster around the 4 day mark, spanning roughly two weeks across.

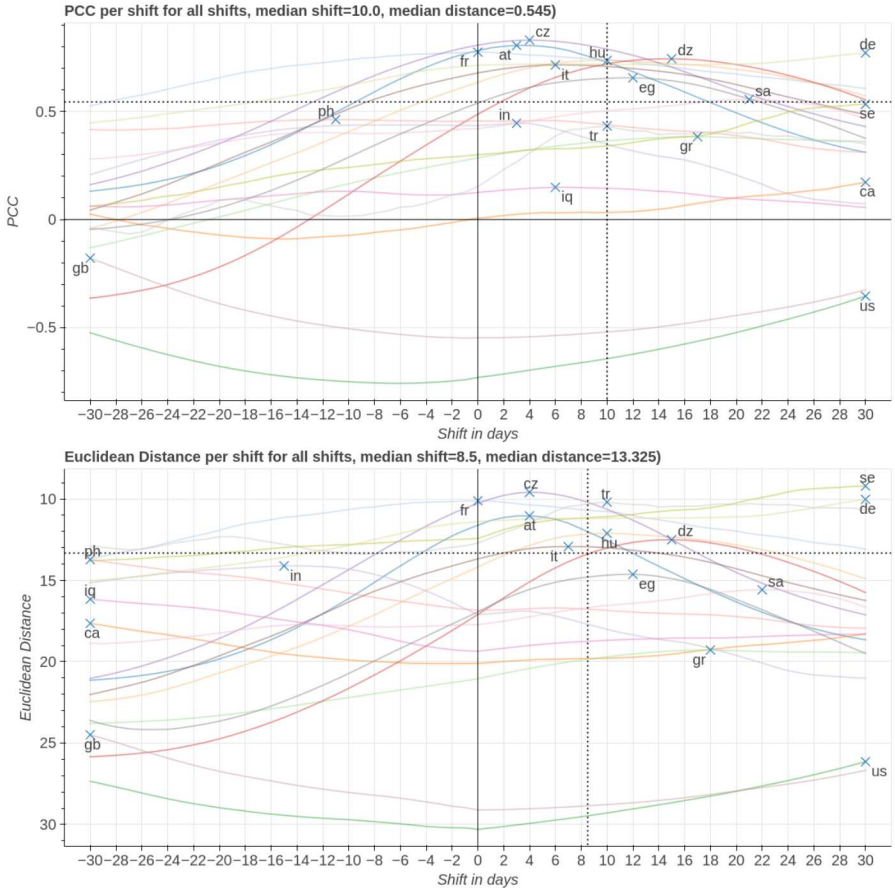


Figure 16: *PCC* and L_2 *norm* for all shifts between $[-30, 30]$ based on *Pulse* and *JH* in 18 different countries. The dotted lines represent the median of the respective axis.

Results of the distance measure comparison is shown by Table 5. The shifted lock-step measures show a high correlation of around 0.8 when compared with their non-shifted counterpart. The dynamic measures show an equally high correlation, when compared to each other. Overall the group of *shifted* L_2, *DTW* and *TWED* show high correlation between 0.73 and 0.83. The highest correlation is between the lock-step measure *shifted* L_2 *norm* and the dynamic measure *TWED*.

Measures	PCC	Shifted PCC	L_2	Shifted L_2	DTW	TWED
PCC	1	0.789	-0.775	-0.496	-0.637	-0.49
Shifted PCC	0.789	1	-0.664	-0.671	-0.709	-0.644

L_2	-0.775	-0.664	1	0.825	0.786	0.73
Shifted L_2	-0.496	-0.671	0.825	1	0.737	0.834
DTW	-0.637	-0.709	0.786	0.737	1	0.804
TWED	-0.49	-0.644	0.73	0.834	0.804	1

Table 5: Correlation between the result sets of the different distance measures.

Table 6 shows the ranking of the selected 18 countries separated by distance measure. *TWED* and *DTW* ranks are relatively close with a maximum distance of 4 for the countries Hungary and Greece. Sweden and Turkey are large outliers for the two lock-step measures *PCC* and L_2 showing a difference in rank of 9 and 8 respectively.

country	pcc	l2	dtw	twed	span	mean
cz	1	1	2	1	1	1.25
fr	3	3	3	3	0	3.00
de	4	5	1	2	4	3.00
at	2	6	6	6	4	5.00
dz	5	8	5	4	4	5.50
hu	6	7	4	8	4	6.25
se	10	2	7	7	8	6.50
tr	13	4	8	5	9	7.50
it	7	9	10	12	5	9.50
sa	9	13	9	11	4	10.50
ph	11	11	11	10	1	10.75
eg	8	12	12	15	7	11.75
in	12	10	15	13	5	12.50
gr	14	16	13	9	7	13.00
iq	16	14	16	14	2	15.00
ca	15	15	14	17	3	15.25
gb	17	17	17	16	1	16.75
us	18	18	18	18	0	18.00

Table 6: Ranks of the countries by distance measure as numerical values. The column *span* is calculated via the difference between the maximum and minimum rank. The table is sorted by the mean rank.

7.3 Country specific results

We selected 7 countries to be included in this result section for the following reasons: Czechia shows the highest similarity between *JH* and *Symptoma data*, while Sweden shows the highest variance in rank. Being an Austrian company, Symptoma GmbH has the largest number of users and more wide spread media coverage in Austria. We included Germany as well, since it has the second most Symptoma users and good media coverage due to language. Greece was included

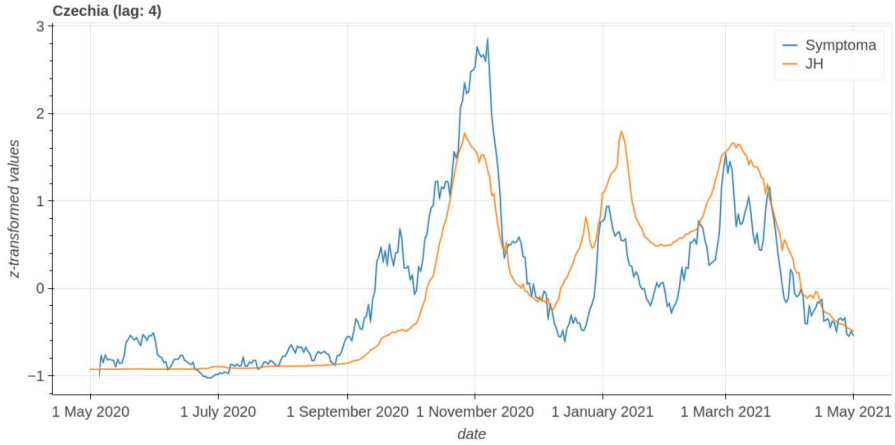
because of its large amount of users and the contrast of the high correlation in our previous work [64] (PCC: 0.94) against the relatively low correltation found by work (PCC: 0.29). Lastly, the United States and the United Kingdom were included as examples for a very low degree of similarity between the time series.

Czechia Out of all countries included in this work, Czechia overall shows the best results in terms of similarity between *JH* and *Symptoma data*. Figure 17a illustrates the high level of overlap of the two time series. The local optimum for the *PCC* has been found by introducing a lag of 4 days to *JH*. Features appear first in *Symptoma data*, before they show up in Czechia, which indicates that Symptoma is slightly ahead of *JH* in Czechia.

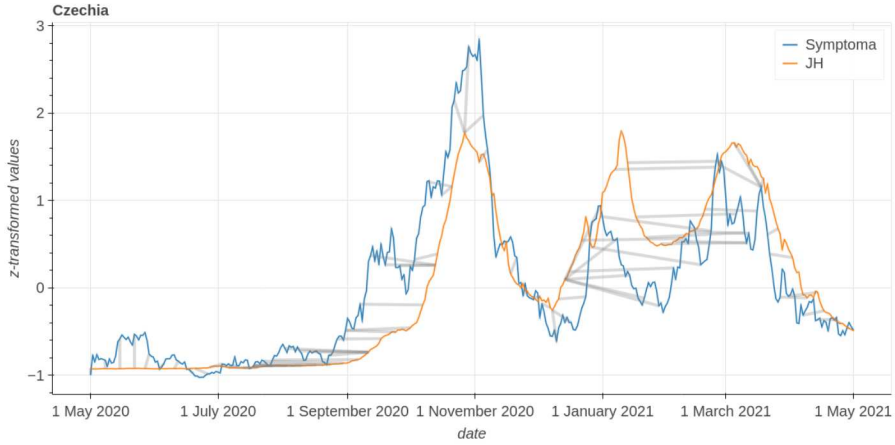
Figure 17b shows the feature mapping obtained from *DTW* on the two time series. The mapping of the first peak spans up to 25 days and maps points that are overall relatively close. Starting from December 2020 the mappings starts to span longer periods. Datapoints from *JH* in December 2020 are mapped to *Symptoma* data in February 2021. The second peak of *JH* got mapped to the third peak of *Symptoma*, spanning roughly 45 days.

Sweden The results of our last work [64], (Figure 24), presented Sweden as uncorrelated, with a *PCC* slightly below 0 for *Symptoma data* and *JH*. In this work we showed a significantly higher *PCC* of 0.3 (+/- 0.1). In Sweden, *Symptoma data* experiences a high amount of noise when compared to the other countries. The lag introduced in JH with 30 days is equal to the maximum allowed lag.

For *DTW* in Figure 18b we find a noisy but relatively close match until the 1st of September. After this point the mapping spans two months between the first large increase in cases for both time series. A highly questionable mapping is the local minimum in mid February. Nearly all points starting from the 1st of January till 1st of May of the *Symptoma data* time series got mapped to that point. The distance in this area varies greatly from a few days up to two months.

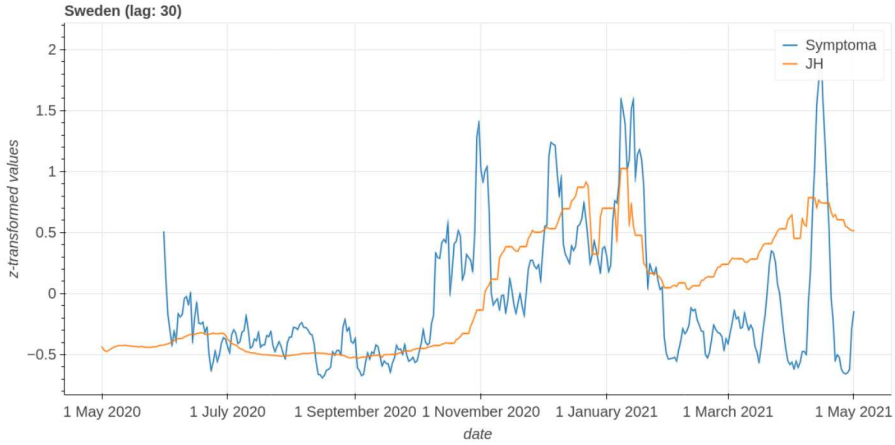


(a) Comparison of time series Symptoma and JH. A lag of 4 days has been applied to JH.

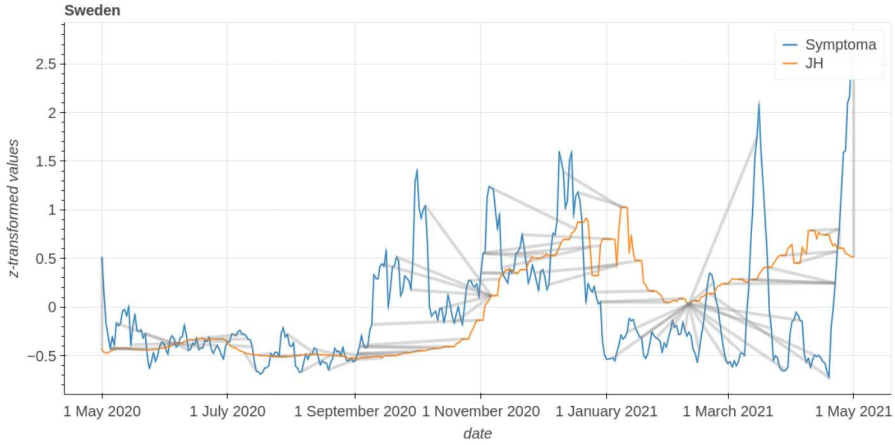


(b) Mapping between Symptoma and JH for Czechia obtained from DTW

Figure 17: Visualizations for Czechia

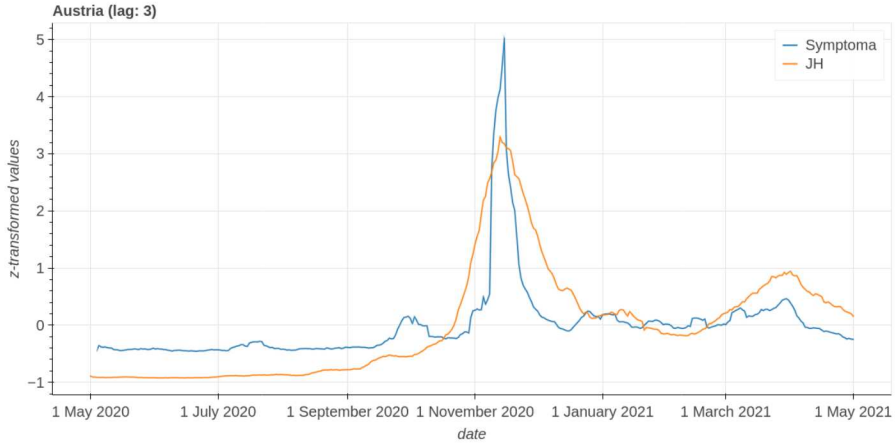


(a) Comparison of time series Symptoma and JH. The maximum lag of 30 days has been applied to JH.

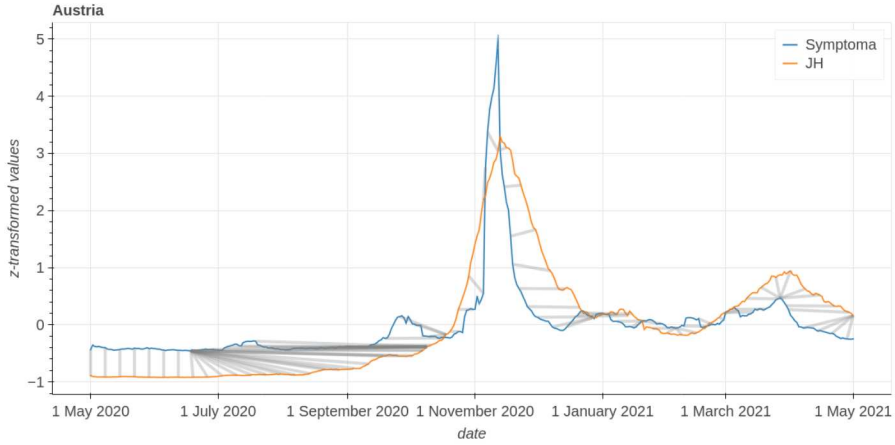


(b) Mapping between Symptoma and JH for Sweden obtained from DTW

Austria In Figure 19a we can observe *Symptoma data* and *JH* shown with an optimal lag of 3 days. *JH* has a slight upward trend, while *Symptoma* remains stationary until November 2020. The first value of JH is close to -1, while the last value is slightly above 0. In contrast the first and last value of *Symptoma data* is close to -0.3. This indicates a slight upward trend for *JH* in general, while *Symptoma data* remains stationary overall. The period between 1st of January 2021 and 1st of March visually overlap quite well. As does the global maximum at the end of November 2020.



(a) Comparison of time series Symptoma and JH. A lag of 3 days has been applied to JH.

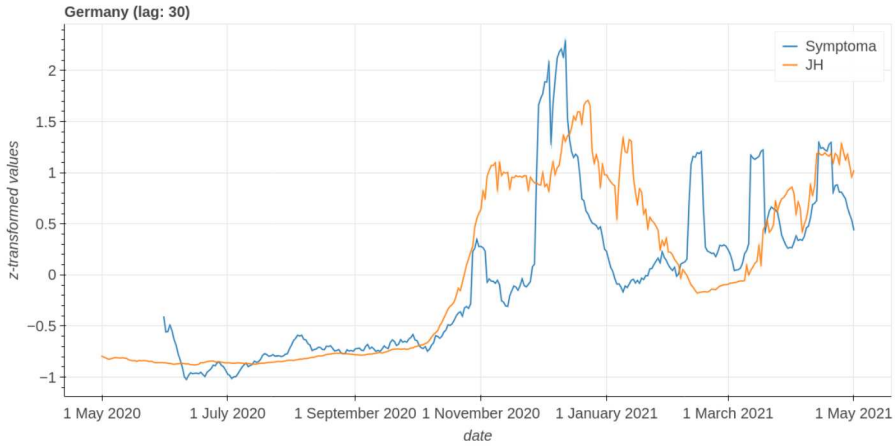


(b) Mapping between Symptoma and JH for Austria obtained from DTW

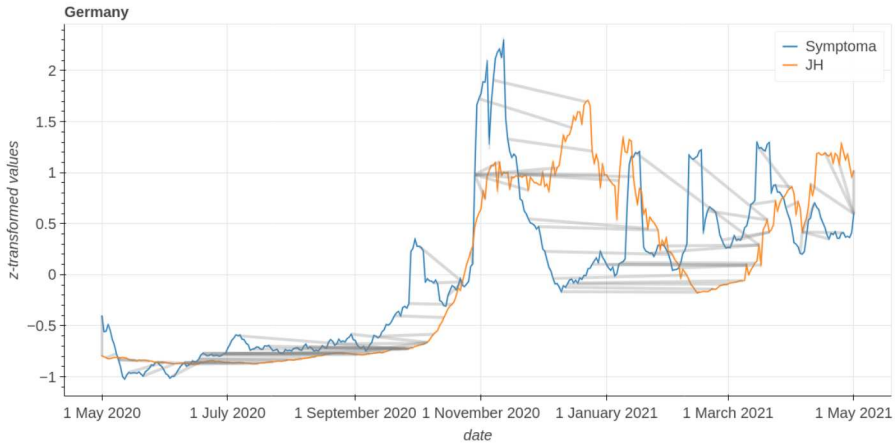
The time periods from 1st of May 2020 till 15th of June, 1st of November till 1st of January 2021 and from 1st of March 2021 till 1st of May 2021 are examples for good mappings produced by *DTW*. The points are relatively close and the shapes appear similar, just scaled up a bit. In contrast, the time period from 1st of July till 1st of October shows mappings which increase in distance up to 70 days. Other mappings remain at a distance smaller than a month.

Germany Figure 20a shows the overlap of *Symptoma data* and *JH* with a lag of 30 days. The shape of the time series until February 2021 overlap rather well. The exception being November 2020, where *Symptoma data* shows a small local minimum while *JH* continues with an upward trend and levels out. One set of interesting features of *Symptoma data* are the three peaks in mid February

2020, mid March 2020 and late April 2020, spaced roughly one month apart.



(a) Comparison of time series Symptoma and JH with a lag of 30 days applied to Symptoma.

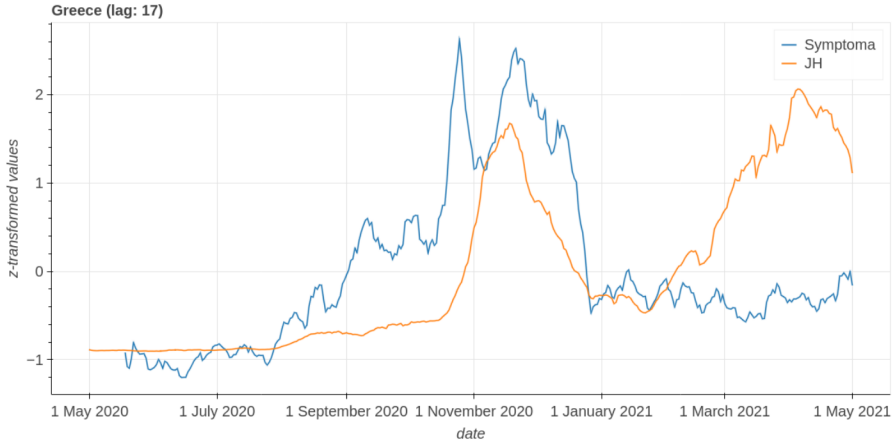


(b) Mapping between Symptoma and JH for Germany obtained from DTW

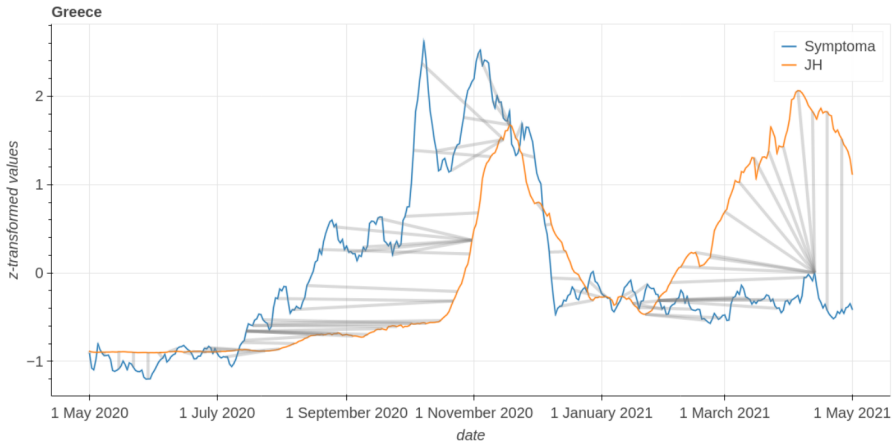
Although sub-optimal, the mapping in Figure 20b shows that some structures get mapped rather well. One example of this is the local minimum of *Symptoma data* at the 10th of December, which gets mapped to the local minimum of *JH* at the 14th of February. The same can be said about the two local maxima at the 10th of November and the 22nd of December. From November 2020 onwards *DTW* produced longer mappings spanning up to 3 months.

Greece In our previous work Greece showed the second highest user base combined with a very high *PCC* of 0.94. The results of this work show a very low *PCC*. The distance measures L_2*-norm*, *DTW* and *TWED* show a

relatively high distance between *JH* and *Symptoma data* when compared to other countries. Figure 21a shows the time series of *JH* and *Symptoma data* for Greece with a lag of 17 days. During the second wave, starting from August 2020 till February 2021 the time series show a high correlation. *Symptoma data* did not capture the third wave and remains relatively level, while *JH* shows an upward trend.



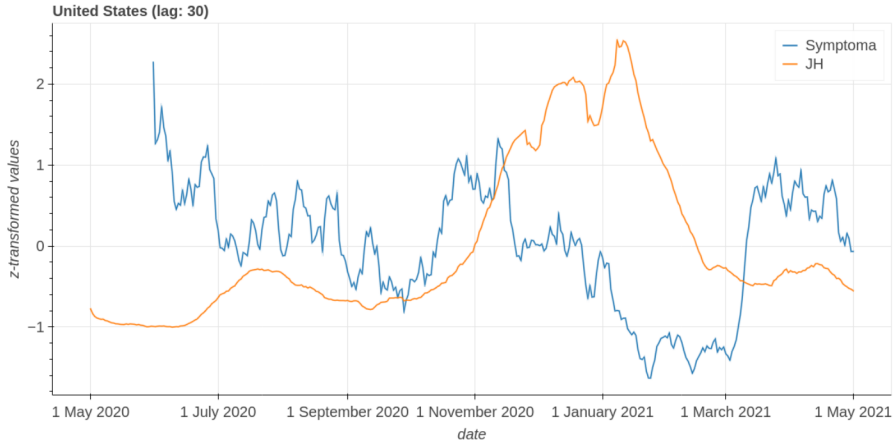
(a) Comparison of time series Symptoma and JH with a lag of 17 days applied to Symptoma.



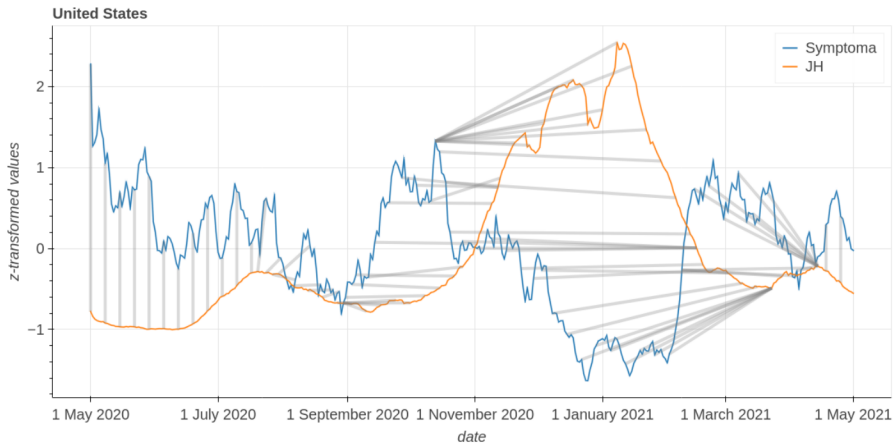
(b) Mapping between Symptoma and JH for Greece obtained from DTW

Figure 21b shows long distance mappings in the period from mid July 2020 till December 2020 having at least a length of 45 days. The period from February 2021 till April 2021 shows some more mappings of similar length. Other periods show relatively short mappings.

United States In our previous work we reported a *PCC* of −0.342. In this work we found a *PCC* of −0.74 with an interval of roughly +/- 0.045. Regardless of distance measure, we cannot argue for a close relationship between *JH* and *Symptoma data* since the USA received the lowest rank on all of them. Figure 22a shows the overlap between *JH* and *Symptoma data* at a lag of 30 days. The period of July 1st 2020 until December 2020 shows a good overlap, while the rest of the time series shows no obvious similarities.



(a) Comparison of time series Symptoma and JH with a lag of 30 days applied to Symptoma.



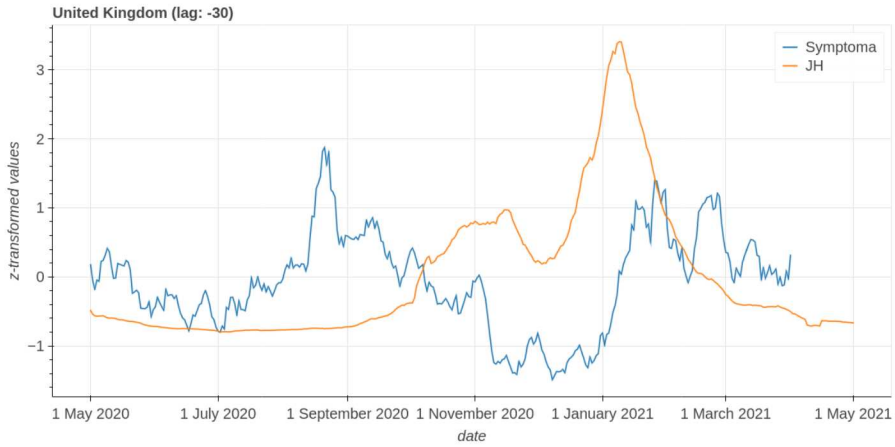
(b) Mapping between Symptoma and JH for the United States of America obtained from DTW

Figure 22b highlights four interesting features. First is the period of May until August 2020 shows a linear mapping. Second is the period of September until November of *Symptoma data* that gets mapped to the period of September

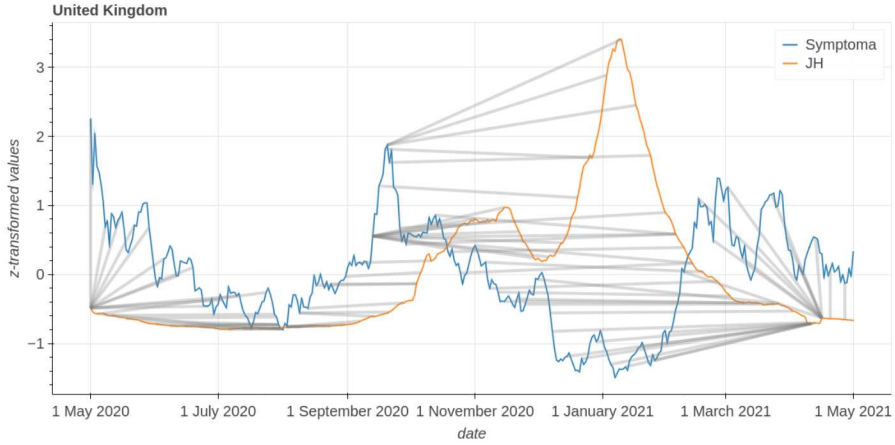
2020 till January 2021 of *JH*. The local maximum in this period of *Symptoma data* at $y = 1.4$ is mapped to all points of the corresponding period of *JH* at $y \geq 1.4$. The third interesting feature is the downward slope of *Symptoma data* starting from November 2020 till February 2021, that is mapped to the slope of *JH* starting in mid of February 2021, ending in April 2021. The last interesting feature is the period from late February 2021 till late April 2021 of *Symptoma* that gets mapped to a singular point, a local maximum of *JH* in mid April 2021. Most of these mappings span multiple months.

United Kingdom Like the USA, the United Kingdom shows a strong negative correlation and a high distance between *Symptoma data* and *JH*, when compared to other countries. Figure 23a illustrates the low similarity of *Symptoma data* and *JH* for the United Kingdom with a lag of -30 days. Both time series have a local maximum at end of August 2020 (*Symptoma data*) and at the start of January 2021 (*JH*) respectively, roughly four months apart.

The mapping obtained by DTW in Figure 23b shows some good results. The slope from 1st of August till mid September maps well to the slope from 1st of September till 1st of November. The local maximum in mid September was mapped well to the global maximum of *JH* at the 5th of January. All of these mappings span at least one month. The whole period from May 1st 2020 until 1st of August 2020 of *Symptoma data* got mapped to the 1st of May of *JH*. A similar concentration happened with the periods from December 2020 till February 2021 and March 2021 till April 2021 of *Symptoma data* which got mapped to two, relatively close points of*JH* in mid April 2021.



(a) Comparison of time series Symptoma and JH with a lag of -30 days applied to Symptoma.



(b) Mapping between Symptoma and JH for the United Kingdom obtained from DTW

8 Discussion

8.1 Concept independence analysis

The analysis aimed to investigate the independence of time series for different medical concepts in the context of the COVID-19 *Symptoma data*, specifically focusing on the comparison with the medical concept COVID-19 from Symptoma's database. The results presented in Table 4 provide valuable insights into the relationship between various diseases and COVID-19, while also shedding light on the limitations of the analysis.

The findings indicate that out of the top 50 medical concepts compared with COVID-19, 12 concepts exhibited a weak correlation (PCC < 0.5) and had a low

overlap in symptoms (overlap ≤ 0.3). This suggests that these concepts have distinct time series compared to COVID-19 and provide evidence for the independence of the *Symptoma data* from media exposure and temporary increases in user counts. The rest of the top 50 concepts contain diseases related to the respiratory tract with similar symptoms to COVID-19. Since symptoms are one of the factors the algorithm of Symptoma considers, correlation between diseases with similar symptoms is the expected result. To prove that Symptoma is able to produce different time series for medical concepts with different underlying symptoms the subset of 12 concepts suffices.

However, it is important to interpret these results within the context of the analysis and acknowledge its limitations. First, the selection of the top 50 returned diseases for Austria may introduce bias, as it assumes that these diseases are representative of the broader population and healthcare-seeking behavior. The generalizability of the findings to other regions or populations should be approached with caution. Austria was chosen since it shows significantly higher usage. A weighted mean over all the countries based on the user statistics might remove the bias. Alternatively the top concepts could be drawn from user sessions of all countries in our database. We argue that the method of selection is of minor importance as long as the set of medical concepts shows a large enough variance and represent a major user base of Symptoma.

Another limitation lies in the use of the overlap coefficient as a measure of symptom-based similarity. While the overlap coefficient provides a simple and intuitive approach, it does not capture the complexity of disease symptomatology. Different diseases may share common symptoms, leading to potential confounding factors in the analysis. Additionally, the overlap coefficient only considers the presence or absence of symptoms and does not account for their severity or frequency. Further research using more sophisticated techniques, such as semantic similarity measures or clinical knowledge-based approaches, could provide a more nuanced understanding of symptom similarity. For this analysis the overlap coefficient was deemed enough. Symptoma does not consider severity or frequency in its algorithm, which means that the overlap coefficient provides a sufficient abstraction of medical concept similarity.

Furthermore, the analysis solely focused on the comparison of time series between diseases and COVID-19. While this approach helps evaluate the independence of *Symptoma data*, it does not directly address the accuracy or reliability of Symptoma in identifying and classifying COVID-19 cases. Symptoma GmbH studied the accuracy of its algorithm in a past study[39], but more work is needed to validate those results in the future.

Additionally, the reliance on the *Symptoma data* introduces its own set of limitations. The accuracy and completeness of the data depends on user interactions with the system and the quality of symptom reporting. Users may exhibit variations in symptom reporting behavior, potentially impacting the reliability of the time series analysis. Moreover, the availability and adoption of Symptoma may vary across different regions, introducing geographical biases. Both points can be addressed via a detailed analysis of user interaction with the chatbot. Such an analysis was out of scope for this work.

Despite these limitations, the results of this analysis provide valuable insights into the relationship between COVID-19 and other medical concepts within the *Symptoma data*. The identification of 12 concepts with weak correlations and low symptom overlap suggests that Symptoma captures sufficiently unique patterns for different diseases. This finding supports the potential of online symptom checkers, such as Symptoma, to provide valuable data for disease surveillance and monitoring, particularly in the context of a pandemic.

8.2 Distance Measures

Comparison to past results. In our previous work [64] we analyzed the time period starting from 8th of April 2020 and ending at the 1st of October 2020. At the time that was the longest period data was available for analysis. In this work we used data starting from 1st of May 2020 till 30th of April 2021 to cover a whole year. Figure 1 (Figure 24) and Figure 2 (Figure 25) from [64] were included to ease the visual comparison between current and former results. The results of this work do not match the results obtained previously. We believe difference in time period to be the major reason for these mismatched results. This work covers roughly 6 month more and leaves out the month of April 2020. This was done to get the most recent data that still overlaps with the time period of the previous study. Most countries experienced a second or even a third wave of infections at the end of 2020 and during the early months of 2021. Others[48] found that the populations perception and thus actions related to the pandemic changed over time. Symptoma is a system that gathers self-reported symptom data. We believe people used Symptoma more during the first wave and moved on in later waves due to fatigue. Those later waves were not covered by our previous work.

8.3 Country Distances

The following section highlights our general findings as well as a sample of 6 out of 18 countries included in the results of this work. We present a visual comparison of the time series. For each country this includes one graph, where the time series of JH is shifted by the lag, which leads to the optimal score for *PCC*. Another graph illustrates the DTW mapping between the time series. To improve readability the mapping is only shown for every 7th day. The *PCC* was chosen for comparison with our past work. We chose DTW as the second metric for visualization because it is one of the elastic distance metrics we chose for this study. Our second reason is that the software package we chose to calculate $TWED$ does not provide a way to retrieve the mappings for visualization, but the package for DTW provides a way to extract the mappings.

General Findings: Our work is part of of a large body of literature that used the *PCC* to study time series similarities between digital media and infectious diseases[23, 11, 14, 36, 30, 8]. We found high correlations of up to 0.81 (Czechia) with a median of 0.545. *Symptoma data* shows high correlation with *JH* in most

of the 18 countries included in this work. For the other distance measures we found low distances for those that showed high correlations. This is supported by our correlation analysis on those measures. As expected, every distance measure shows a strong negative correlation with *PCC* ($r_{L2shifted} = -0.671$, $r_{DTW} = -0.709$, $r_{TWED} = -0.644$, Table 5). Countries where *Symptoma data* showed an anti-correlation with *JH* mostly included those with a small user base relative to the total population (US, GB). Data from other countries showed higher similarity despite their relatively low user count. We suggest a more detailed, international study of Symptoma users to investigate. Possible reasons include differences in digital literacy and symptom self reporting behavior. In general the result of our last work[64] still stands: most countries with a high user base show close similarity to real world COVID-19 case counts.

We calculated lags for *PCC* and *L_2-norm* specifically. In our previous work we found a median lag of +5 days. In this work the shift is even higher at median +10 days for *PCC* and median +8.5 days for *L_2-norm*. Those results suggest that more work is needed to define when and how Symptoma can be used as an early warning signal for consensus based monitoring systems[54]. A detailed analysis of trends, events, when and how *Symptoma data* precedes these could improve clarity on the usefulness of such a system based on situational parameters. Others have suggested that media and general news fatigue might introduce bias into monitoring systems that use fixed digital data streams for a long time. We found similar tendencies in our data where time series similarity degraded over time (Greece, Figure 21a). However instead of combating the fatigue itself, finding the right situational parameters to use a system is more effective. Another solution that requires studying is to use the fraction of total users from an area that tested positive for a disease. Currently our work only uses the raw test counts.

We applied time series preprocessing techniques to make our data comparable with the data obtained from Johns Hopkins University. Normalization via Z-transformation and a moving average of 7 days was used to map data sets to a comparable range of values. The time series for *Symptoma data* still shows a higher degree of noise. Both systems collect information based on people, which makes one person the smallest unit of change. Johns Hopkins University tracks infections of large quantities in a population. The average user count of Symptoma is lower than that. One person has, relatively speaking, more of an influence on *Symptoma data*. Normalization to the same value range highlights the variance of *Symptoma data*. Further preprocessing steps could smooth out *Symptoma data* but they would also introduce bias if selection of methods and parameters are not studied in depth beforehand. Another way to mitigate the variance is to increase the user base of Symptoma via cooperation with other companies[60] and general awareness campaigns. However, this introduces other limitations that will be addressed later.

A caveat of our chosen normalization is that countries like Austria, the US and UK show a large gap between the first couple of months. Normalization has a large influence on our chosen similarity measures, especially the lock-step measures. Future studies should invest more time into the choice of a

normalization. We chose the z-normalization after testing it with only a small sample set of time series and kept it so to not bias our results on what we wanted to see.

For calculation of the mappings between time series with DTW we used the standard parameters of the package. Every mapping shows at least one instance of a gap greater than one month, sometimes even three months. For the calculation of lags we proposed a cutoff of 30 days. Mappings spanning such magnitudes can only be rejected based on our chosen cutoff. The standard cost function for DTW did only account for differences in values and not for differences in time. Future studies need to devise a cost function that accounts for both and defines a limit to the time a mapping can span.

Czechia: Out of all countries included in this work, Czechia overall shows the best results in terms of similarity between *JH* and *Symptoma data*. Figure 17a illustrates the high level of overlap of the two time series. The local optimum for the *PCC* has been found by introducing a lag of 4 days to JH. This indicates that *Symptoma data* is slightly ahead of *JH* in Czechia and general features appear first in *Symptoma data*, before they show up in Czechia. We argue that the second peak in January 2021 shows a bigger lag (25 days), meaning that it appeared in Symptoma even earlier when compared to the optimal lag of 4 days.

Figure 17b shows the feature mapping obtained from DTW on the two time series. We can see that the mapping works quite well for the first peak. Parts of the second peak were mapped to the third peak, which highlights the limitations of the standard cost function mentioned before.

The *PCC* of 0.807 and optimal lag in this work show comparable results to our previous work [64]. *Symptoma data* models the pandemic in Czechia roughly 4-5 days in advance. Other diseases need to be studied to ensure generalization of our results.

Sweden: The relatively high variance of Swedens ranking in Table 6 makes this country an interesting candidate for a detailed analysis. That table contains the rankings for each of the countries in the rows, achieved through the distance measures in the columns. The rankings represent the distance between *Symptoma* and *JH*, where 1 means that both time series are the closest match for the respective country when compared to the distance between *Symptoma* and *JH* in other countries. The results of our last work [64], (Figure 24), presented *JH* Sweden as uncorrelated, with a *PCC* slightly below 0. In this work we showed a significantly higher *PCC* of 0.3 (+/- 0.1). This slightly positive correlation stems from the fact that the general trend of Swedens *Symptoma data* and lagged (30 days) JH match well until February 2021 (see Table 18a). This explains the relatively low L_2-*norm* as well. In Sweden, *Symptoma data* experiences a high amount of noise when compared to other countries with the same amount of users. Possible explanations include automatic probes or tests of our system, though we found no conclusive evidence for this. We reject red

team attacks as a possibility. Symptoma is currently not in use for large scale monitoring efforts, thus we fail to see a reason for such data manipulation tactics. A more detailed, statistical analysis could shed some light on the question, but such an analysis was out of scope for this work.

The lag introduced in *JH* with 30 days is equal to the maximum allowed lag. This indicates that a more optimal local minimum for the distance between *Symptoma data* and *JH* lies outside our defined scope. The points on the edge of the defined scope can be seen as pseudo optima, since the process of finding a local optimum stops without the assurance that we really found a local optimum. A possible explanation for such a high lag could be that users anticipated a rise in infections earlier in fall, right after summer and tested themselves.

For *DTW* in Figure 18b we find a noisy but relatively close match until the 1st of September. After this point the mapping breaks down because of the limitations of the standard cost function discussed before.

The high amount of noise, the large introduced lag and variance in rank indicate that the similarity between *JH* and *Symptoma data* in Sweden should be investigated more in detail than the scope of this study allows.

Austria: As an Austrian company Symptoma GmbH appears more often in Austrian media and thus is more known. Additionally Symptoma GmbH was working together with Fonds Soziales Wien to provide assistance to the 1450 COVID-19 hot-line [60]. This lead to Symptoma having the highest user count per capita in Austria during that time.

From the data perspective the time series of Austria provides insights into the shortcomings of the *z-normalization* in the context of lockstep distance measures and *DTW*. In Figure 19a we can observe, that *JH* has a slight upward trend, while *Symptoma data* remains stationary. The start and endpoints of the time series are indicators for this. The first value of *JH* is close to -1, while the last value is slightly above 0. In contrast the first and last value of *Symptoma data* is close to -0.3. This leads to a relatively high lockstep distance during the first 4 months when compared to a min-max normalization. The reason for this is that both start points of the time series are either the global minimum or at least very close. A min-max normalization would scale those values to be equal or close to 0. The period between 1st of January 2021 and 1st of March visually overlap quite well. A different normalization would have a large influence lockstep distances.

The time periods from 1st of May 2020 till 15th of June, 1st of November till 1st of January 2021 and from 1st of March 2021 till 1st of May 2021 are examples for good mappings with *DTW*. The points are relatively close and the shapes appear similar, just scaled up or down. In contrast, the time period from 1st of July till 1st of October is a good example for a bad mapping as we discussed in an earlier section.

With regards to the distance measures Austria remains as one of the best performing countries (Table 6). This result is not surprising since Symptoma GmbH operates out of that country and enjoys increased media coverage. It

is important for online media based surveillance tools to have a large and representative user base to cover enough of the population to be a good model for it.

Germany: We presented the lagged time series of *Symptoma data* and *JH* in Germany in Figure 20a. The fall season shows a high degree of similarity. *Symptoma data* tracks the first peak half way, declines and then overshoots the long and relatively flat peak of *JH*. The smaller peak happened at the start of October, while the bigger one happened in early November. The latter one relates well to the time before Christmas, when lots of people wanted to get tested before meeting at Christmas parties. Positively tested user counts decline rapidly during December. We attribute this to media fatigue. In 2021 similarities break down due to three events we discuss later.

Within the period of time selected for this study, Germany is the country with the second most Symptoma users. Out of all countries included in this detailed discussion, Germany shows the potential of a good overlap, but fails mainly due to the three events from 1st of January 2021 till 1st of April. These events each span exactly one week, are roughly one month apart and show values very close together, when compared to the rest of the time series. Because of the anonymity of our data we cannot pinpoint the cause. We did not filter and interpolate those events to highlight what we think is an important limitation. The likely cause is traffic from one of our customers, the FUNKE media group, which hosted a Symptoma symptom checker widget during the time in question on their websites related to medical content. Although this would only hold true if every week roughly the same amount of users were assessed with a high risk in regards to COVID-19. The fact that these events occur over the span of a week rules out automated tests or bots. These would usually occur during a short period of time. Reported symptoms and time differences between consecutive request also show no indication of automated activity. We propose a change in the API that enables tracking and filtering of such events.

Greece: In our previous work Greece showed the second highest user base combined with a very high *PCC* of 0.94. The results of this work show a low *PCC* of 0.285 before shifting to the optimal lag (Appendix Table 7). The distance measures L_2*-norm*, *DTW* and *TWED* show a relatively high distance between *JH* and *Symptoma data* when compared to other countries. Figure 21a shows the time series of *JH* and *Symptoma data* for Greece. During the second wave the time series show a high correlation. But Symptoma did not capture the third wave, which is the reason for the calculated similarity being lower when compared to our past work. The cause for this decline in usage of Symptoma can be attributed to the population being fatigued of the pandemic and not using the symptom checker anymore. An analysis of changes in test frequencies compared to positive test results could shed light on how to handle this limitation.

United States: In our last work we showed reported a *PCC* of -0.342 for the United States. In this work we found a *PCC* of -0.74 with an interval of roughly +/- 0.045. Regardless of distance measure, we cannot argue for a close relationship between *JH* and *Symptoma data* since the USA received the lowest rank on all of them. Although Symptoma gets accessed by above average user per population, we argue that our service is not used homogeneously enough trough out the country to adequately capture the flow of the pandemic. Large portions of the country are rural areas, whose inhabitants we suppose either do not know of or use our system. The USA is a good example for a country, where a detailed analysis of spatial user distribution could lead to insights why we perform badly.

Another angle for future work is to view the COVID-19 statistic of each state separately. For this the API we used to retrieve *Symptoma data* needs to be extended in a way, that allows us to relate users to sub country level administrative structures. One concern of this approach is user privacy. The users willingness to provide private health information is based on their trust that we respect their privacy. If we were to store fine grained location data we increase the risk of de-anonymization which in turn can erode the trust placed in us. Sensitive information which is not stored cannot be leaked. To avoid the risk of de-anonymization after data leaks we need to keep the level of spatial data mining relatively high. The state and region levels should provide enough anonymity, but still give insights into the distribution of users. This in turn can be used to determine the population coverage of our system, which has the potential to be a good indicator for the power of Symptoma user data as a model for healthcare related events like pandemics and epidemics.

United Kingdom: Like the USA, the United Kingdom (GB) shows a strong negative correlation and a high distance between *Symptoma data* and *JH*, when compared to other countries. Both USA and GB experienced a rise of COVID-19 cases from 1st of December 2020 till 1st of January 2021. During the same period Symptoma user sessions assessed as having a high risk in regards to COVID-19 experienced a global minimum. Once the COVID-19 cases started to decrease, Symptoma high risk sessions began to rise again.

Though we have no strong argument for this behaviour, it is a fact that Babylon Health, one of our competitors operates in the United Kingdom. Their increased media presence in the country could lead to users prioritizing their service over ours during times of heightened pandemic awareness in the population.

The mapping obtained by *DTW* in Figure 23b shows some good results for shapes that cannot be considered as a good mapping because the distance exceeds the one month constraint we set. The slope from 1st of August till mid September maps well to the slope from 1st of September till 1st of November. The local maximum a roughly mid September was mapped well to the global maximum of JH at the 5th of January. This shows again that the mapping of shapes works well in general and simply requires a temporal constraint

to prevent mappings with undesired distances in the temporal domain as was discussed before.

9 Conclusion

In this work we present how the exchange of information has been one of the major drivers for humanity to arrive at the world we know today. We further describe how information exchange led to improvements in the medical domain. The Internet and online media residing in it provide a large body of literature, services and a platform for information exchange. The digital health assistant Symptoma already provides value in guiding patients towards the right diagnosis [34]. It is our believe that the user statistics of Symptoma have the potential to support policy makers in healthcare. In our last work[64] we proposed to use statistics of the risk assessment of user sessions to forecast large scale health events, like pandemics. Our results from that paper showed a positive correlation between the data from Symptoma and the data provided by the CSSE at Johns Hopkins University (JH). Further we came to the conclusion that Symptoma is predictive with a median of $+5$ days. In this work we repeated those experiments using an up to date data set, expanded to contain one full year of data. This excludes the first wave of COVID-19 cases for most countries, but still includes wave 2 and 3. We found that some countries experienced a severe drop in PCC with the expanded data set. Greece is a good example for countries where *Symptoma data* and JH overlapped well initially. Our updated data set showed that same overlap in the beginning but revealed that *Symptoma data* does not capture the third wave. We share the hypothesis of Mavragani et al.[36] that this is caused by a decline in interest in COVID-19 related information and systems. In the countries United Kingdom and United States we failed to model the pandemic adequately. In Czechia, Austria and France we managed to model the pandemic well. In other countries we managed to model parts of it. In regards to DTW we found the mapping returned by the standard cost function akin to the L_2-*norm* unsuited to our data. The difference in the time domain has to be accounted for as well to return a sensible mapping in the context of this work. The high overlap in some countries shows that Symptoma has the potential to model large scale health events, if the conditions are right. The system requires a constant stream of users to model reality properly and provides an additional point of reference for an informed decision making process.

10 Future work

Our work showed that user statistics of the online health assistant Symptoma can model real world infectious disease events, if certain conditions are met. The scope of this work only includes COVID-19 between 2020 and 2021. A comparison of *Symptoma data* related to other diseases with real world data

over larger time spans should provide valuable insights into the conditions that guarantee high model performance.

For our calculations we used the full one year span of our data. Visually we found that similarities between time series decline after some time and pointed to fatigue related to COVID-19 like others did[36]. A detailed, step by step analysis of this decline should provide useful parameters to assess the reliability of a system like Symptoma in different stages of a pandemic or epidemic like event. For that one needs to take media presence and type of usage into account. Future work needs to differentiate between self-reports of users, usage through a partner like we had with *FSW*[60] and mandatory usage, when a company mandates the usage of such a system. It is reasonable to assume that fatigue would show up differently depending of the context of use.

We used the measures *PCC*, L_2*-norm*, *DTW* and *TWED*. The two elastic measures need to be revisited with a cost function that incorporates the time domain to better map similar structures in a meaningful way. Finding a *TWED* package which allows extraction of the mapping would enable us to compare it to *DTW* visually and help to better determine the quality of its mapping. A custom implementation of *TWED* is also an option. Others used linear regression models[47, 14], multiple linear regression *MLR* models[51], *ARIMA* models[21, 51] and *LSTM* models[51]. Future work should include such models as well to compare to the large body of past research.

In hindsight we think that our conditions for including a country in the analysis was too strict. Future work will include more countries with the number of gaps being an additional feature instead of a filter criteria.

Similarly the hard cut-off of +/- 30 days for the lags need to be revisited. Future work needs to revisit the question of local maxima on the cut-off threshold. A simple solution could be to calculate optimal lags without a cut-off and filter afterwards.

11 Acknowledgement

We want to thank the development team of Symptoma GmbH for providing the APIs we used to create the *Symptoma data* set. Additionally we want to give our thanks to the management of Symptoma GmbH for allowing us the opportunity to work with this data and to our supervisors from the University of Vienna for their continued support and guidance.

References

[1] Alwashmi, M. F. The use of digital health in the detection and management of COVID-19. *International Journal of Environmental Research and Public Health 17*, 8 (2020), 2906. Number: 8 Publisher: Multidisciplinary Digital Publishing Institute.

[2] Bailey, D. H., and Swarztrauber, P. N. A fast method for the numerical evaluation of continuous fourier and laplace transforms. *SIAM Journal on Scientific Computing 15*, 5 (1994), 1105–1110. Publisher: Society for Industrial and Applied Mathematics.

[3] Berndt, D. J., and Clifford, J. Using dynamic time warping to find patterns in time series. In *Proceedings of the 3rd international conference on knowledge discovery and data mining* (1994), pp. 359–370.

[4] Brown, J. S., Bastarache, L., and Weiner, M. G. Aggregating electronic health record data for COVID-19 research—caveat emptor. *JAMA Network Open 4*, 7 (2021), e2117175.

[5] Childress, D. *Johannes Gutenberg and the printing press*. Twenty-First Century Books, 2007. OCLC: 248029067.

[6] Crompton, S. W. *The printing press: transforming power of technology*. Transforming power of technology. Chelsea House Publishers, 2003.

[7] De Long, J. B. Estimates of world GDP, one million b.c. - present.

[8] Deiner, M. S., Deiner, N. A., Hristidis, V., McLeod, S. D., Doan, T., Lietman, T. M., and Porco, T. C. Use of large language models to assess the likelihood of epidemics from the content of tweets: Infodemiology study. *J Med Internet Res 26* (Mar 2024), e49139.

[9] Dong, E., Du, H., and Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases 20*, 5 (2020), 533–534. Publisher: Elsevier.

[10] Eisenstein, E. L. *The Printing Press as an Agent of Change*. Cambridge University Press, 1980. Google-Books-ID: WR1eajpBG9cC.

[11] Eysenbach, G. Infodemiology: tracking flu-related searches on the web for syndromic surveillance. In *AMIA annual symposium proceedings* (2006), vol. 2006, American Medical Informatics Association, p. 244.

[12] Forouzan, B. A. *Data communications and networking*, 5th ed ed. Huga Media, 2007.

[13] Frank, S. R. Digital health care—the convergence of health care and the internet. *The Journal of Ambulatory Care Management 23*, 2 (2000), 8–17.

[14] Generous, N., Fairchild, G., Deshpande, A., Del Valle, S. Y., and Priedhorsky, R. Global disease monitoring and forecasting with wikipedia. *PLoS computational biology 10*, 11 (2014), e1003892.

[15] Gillespie, G. Technology of communication timeline – gary gillespie - eagle. https://eagle.northwestu.edu/faculty/gary-gillespie/technology-of-communication-timeline/, 2022. accessed: 2022-05-10.

[16] Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., and Brilliant, L. Detecting influenza epidemics using search engine query data. *Nature 457*, 7232 (2009), 1012–1014. Number: 7232 Publisher: Nature Publishing Group.

[17] Golinelli, D., Boetto, E., Carullo, G., Nuzzolese, A. G., Landini, M. P., and Fantini, M. P. How the COVID-19 pandemic is favoring the adoption of digital technologies in healthcare: a literature review. *MedRxiv* (2020), 2020–04. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Type: article.

[18] Guido, V. R., and Drake Jr, F. *Python 3 reference manual*. CreateSpace, 2009.

[19] Hamilton, J. D. *Time Series Analysis*. Princeton University Press, 2020. Publication Title: Time Series Analysis.

[20] Harris, C. R., Millman, K. J., Walt, S. J. v. d., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., Kerkwijk, M. H. v., Brett, M., Haldane, A., Río, J. F. d., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. Array programming with NumPy. *Nature 585*, 7825 (2020), 357–362. Publisher: Springer Science and Business Media LLC.

[21] Higgins, T. S., Wu, A. W., Sharma, D., Illing, E. A., Rubel, K., and Ting, J. Y. Correlations of online search engine trends with coronavirus disease (covid-19) incidence: Infodemiology study. *JMIR Public Health Surveill 6*, 2 (May 2020), e19702.

[22] Infogalactic. Printing - infogalactic: the planetary knowledge core. https://infogalactic.com/info/Printing, 2022. accessed: 2022-04-09.

[23] Johnson, H. A., Wagner, M. M., Hogan, W. R., Chapman, W., Olszewski, R. T., Dowling, J., and Barnas, G. Analysis of web access logs for surveillance of influenza. In *MEDINFO 2004* (2004), IOS Press, pp. 1202–1206.

[24] Joshi, A., Sparks, R., McHugh, J., Karimi, S., Paris, C., and MacIntyre, C. R. Harnessing tweets for early detection of an acute disease event. *Epidemiology 31*, 1 (2020), 90–97.

[25] Kreis, S. The printing press. http://www.historyguide.org/intellect/press.html, march. accessed: 2022-03-01 10:35:15.

[26] Lampos, V., and Cristianini, N. Tracking the flu pandemic by monitoring the social web. In *2010 2nd International Workshop on Cognitive Information Processing* (2010), pp. 411–416. ISSN: 2327-1698.

[27] Lazer, D., Kennedy, R., King, G., and Vespignani, A. The parable of google flu: Traps in big data analysis. *Science 343*, 6176 (2014), 1203–1205. Publisher: American Association for the Advancement of Science.

[28] Leiner, B., Cerf, V., Clark, D., Kahn, R., Kleinrock, L., Lynch, D., Postel, J., Roberts, L., and Wolff, S. A brief history of the internet. *Computer Communication Review 39* (2009), 22–31.

[29] Levenshtein, V. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady 10*, 8 (1966), 707–710.

[30] Li, J., Huang, W., Sia, C. L., Chen, Z., Wu, T., and Wang, Q. Enhancing covid-19 epidemic forecasting accuracy by combining real-time and historical data from multiple internet-based sources: Analysis of social media data, online news articles, and search queries. *JMIR Public Health Surveill 8*, 6 (Jun 2022), e35266.

[31] Magdy, N., Sakr, M., and Elbahnasy, K. A generic trajectory similarity operator in moving object databases. *Egyptian Informatics Journal 18* (2017), 29–37.

[32] Marteau, P.-F. Time warp edit distance. *arXiv preprint arXiv:0802.3522* (2007).

[33] Marteau, P.-F. Time warp edit distance with stiffness adjustment for time series matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence 31*, 2 (2009), 306–318. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[34] Martin, A., Nateqi, J., Gruarin, S., Munsch, N., Abdarahmane, I., Zobel, M., and Knapp, B. An artificial intelligence-based first-line defence against COVID-19: digitally screening citizens for risks via a chatbot. *Scientific Reports 10*, 1 (2020), 19012. Number: 1 Publisher: Nature Publishing Group.

[35] Maus, V., Câmara, G., Cartaxo, R., Sanchez, A., Ramos, F. M., and de Queiroz, G. R. A time-weighted dynamic time warping method for land-use and land-cover mapping. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 9*, 8 (2016), 3729–3739.

Conference Name: IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing.

[36] Mavragani, A. Tracking covid-19 in europe: Infodemiology approach. *JMIR Public Health Surveill 6*, 2 (Apr 2020), e18941.

[37] Metcalf, T. G., Melnick, J. L., and Estes, M. K. Environmental virology: from detection of virus in sewage and water by isolation to identification by molecular biology - a trip of over 50 years. *Annual Review of Microbiology 49* (1995), 461–488. Publisher: Annual Reviews, Inc.

[38] M.k, V., and K, K. A survey on similarity measures in text mining. *Machine Learning and Applications: An International Journal 3*, 2 (2016), 19–28.

[39] Munsch, N., Martin, A., Gruarin, S., Nateqi, J., Abdarahmane, I., Weingartner-Ortner, R., and Knapp, B. Diagnostic accuracy of web-based COVID-19 symptom checkers: Comparison study. *Journal of Medical Internet Research 22*, 10 (Oct 2020), e21299. Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.

[40] Needleman, S. B., and Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology 48*, 3 (1970), 443–453.

[41] Nutton, V. Books, printing and medicine in the renaissance. *Medicina nei Secoli: Journal of History of Medicine and Medical Humanities 17*, 2 (2005), 421–442. Number: 2.

[42] Oldstone, M. B. A. *Viruses, plagues, and history: past, present, and future*, rev. and updated ed ed. Oxford University Press, 2020. OCLC: ocn301798298.

[43] Paassen, B., Mokbel, B., and Hammer, B. A toolbox for adaptive sequence dissimilarity measures for intelligent tutoring systems. In *Proceedings of the 8th International Conference on Educational Data Mining (EDM 2015)* (2015), O. C. Santos, J. G. Boticario, C. Romero, M. Pechenizkiy, A. Merceron, P. Mitros, J. M. Luna, C. Mihaescu, P. Moreno, A. Hershkovitz, S. Ventura, and M. Desmarais, Eds., International Educational Datamining Society, p. 632.

[44] Paparrizos, J., and Gravano, L. k-shape: Efficient and accurate clustering of time series. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data* (2015), SIGMOD '15, Association for Computing Machinery, pp. 1855–1870.

[45] Paparrizos, J., Liu, C., Elmore, A. J., and Franklin, M. J. Debunking four long-standing misconceptions of time-series distance measures. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data* (2020), SIGMOD '20, Association for Computing Machinery, pp. 1887–1905.

[46] Pitlik, S. D. COVID-19 compared to other pandemic diseases. *Rambam Maimonides Medical Journal 11*, 3 (2020), e0027.

[47] Polgreen, P. M., Chen, Y., Pennock, D. M., Nelson, F. D., and Weinstein, R. A. Using internet searches for influenza surveillance. *Clinical Infectious Diseases 47*, 11 (12 2008), 1443–1448.

[48] Saito, R., and Haruyama, S. Estimating time-series changes in social sentiment@ twitter in us metropolises during the covid-19 pandemic. *Journal of computational social science 6*, 1 (2023), 359–388.

[49] Sakoe, H., and Chiba, S. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing 26*, 1 (1978), 43–49. Conference Name: IEEE Transactions on Acoustics, Speech, and Signal Processing.

[50] Serrà, J., and Arcos, J. L. An empirical evaluation of similarity measures for time series classification. *Knowledge-Based Systems 67* (2014), 305–314.

[51] Shih, D.-H., Wu, Y.-H., Wu, T.-W., Chang, S.-C., and Shih, M.-H. Infodemiology of influenza-like illness: Utilizing google trends' big data for epidemic surveillance. *Journal of Clinical Medicine 13*, 7 (2024).

[52] Silver, N. *The signal and the noise: why so many predictions fail - but some don't*. Penguin Publishing Group, 2012. OCLC: 951456917.

[53] Smith, T. F., and Waterman, M. S. Identification of common molecular subsequences. *Journal of Molecular Biology 147*, 1 (1981), 195–197.

[54] Stolerman, L. M., Clemente, L., Poirier, C., Parag, K. V., Majumder, A., Masyn, S., Resch, B., and Santillana, M. Using digital traces to build prospective and real-time county-level early warning systems to anticipate covid-19 outbreaks in the united states. *Science Advances 9*, 3 (2023), eabq0199.

[55] team, T. p. d. pandas-dev/pandas: Pandas. Tech. rep., NumFOCUS, 2020.

[56] University, J. H. COVID-19 data repository by the center for systems science and engineering (CSSE) at johns hopkins university. `https://github.com/CSSEGISandData/COVID-19`, jun 2021. accessed: 2020-02-04 22:03:53.

[57] University, J. H. JH CSSE COVID-19 time series data. https://github.com/CSSEGISandData/COVID-19/blob/73115335b3e166942ea00e3a0bd0b3523fb858d8/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_global.csv, may 2021. accessed: 2021-05-26 05:06:16.

[58] Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, I., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors. SciPy 1.0: Fundamental algorithms for scientific computing in python. *Nature Methods 17*, 3 (2020), 261–272.

[59] Wang, X., Mueen, A., Ding, H., Trajcevski, G., Scheuermann, P., and Keogh, E. Experimental comparison of representation methods and distance measures for time series data. *Data Mining and Knowledge Discovery 26*, 2 (2013), 275–309.

[60] Wien, F. S. Symptom-checker für corona selbst-check. https://www.fsw.at/n/symptom-checker-fuer-corona-selbst-check, jun 2020. accessed: 2021-02-04 15:03:53.

[61] Yelin, I., Aharony, N., Tamar, E. S., Argoetti, A., Messer, E., Berenbaum, D., Shafran, E., Kuzli, A., Gandali, N., Shkedi, O., Hashimshony, T., Mandel-Gutfreund, Y., Halberthal, M., Geffen, Y., Szwarcwort-Cohen, M., and Kishony, R. Evaluation of COVID-19 RT-qPCR test in multi sample pools. *Clinical Infectious Diseases 71*, 16 (2020), 2073–2078.

[62] Zhang, X., Wang, M., Liu, K., Xie, J., and Xu, H. Using NDVI time series to diagnose vegetation recovery after major earthquake based on dynamic time warping and lower bound distance. *Ecological Indicators 94* (2018), 52–61.

[63] Zimmer, C. *A planet of viruses*, second edition ed. The University of Chicago Press, 2021.

[64] Zobel, M., Martin, A., Nateqi, J., and Knapp, B. Predicting global trends in COVID-19 cases via online symptom checkers self-assessments. *Available at SSRN 3729913*, ID 3729913 (2020).

12 Appendix

12.1 Graphics

12.2 Tables

iso	name	days	interval(-)	pcc	interval(+)	p-value
cz	Czechia	366	0.768188	0.807138	0.840135	2.516671e-85
at	Austria	366	0.738420	0.781755	0.818659	1.173709e-76
fr	France	366	0.730240	0.774754	0.812718	1.810292e-74
de	Germany	366	0.659740	0.713964	0.760793	2.648984e-58
it	Italy	366	0.619393	0.678803	0.730483	9.340581e-51
hu	Hungary	366	0.570151	0.635517	0.692888	8.447039e-43
eg	Egypt	366	0.462212	0.539175	0.608076	5.450789e-29
dz	Algeria	366	0.405421	0.487664	0.562075	2.911843e-23
ph	Philippines	366	0.371837	0.456931	0.534409	2.795131e-20
in	India	366	0.348233	0.435207	0.514753	2.397342e-18
sa	Saudi Arabia	366	0.332267	0.420456	0.501357	4.126549e-17
se	Sweden	366	0.203786	0.300029	0.390528	4.750021e-09
gr	Greece	366	0.188342	0.285343	0.376831	2.757231e-08
tr	Turkey	366	0.050364	0.152089	0.250691	3.537945e-03
iq	Iraq	366	0.023366	0.125575	0.225187	1.622999e-02
ca	Canada	366	-0.096589	0.005980	0.108424	9.092274e-01
gb	United Kingdom	366	-0.615416	-0.547442	-0.471385	5.278867e-30
us	United States	366	-0.775424	-0.731021	-0.679426	2.166600e-62
ma	Morocco	365	0.797575	0.832056	0.861118	3.428525e-95
jo	Jordan	365	0.523637	0.594249	0.656752	2.590554e-36
ch	Switzerland	365	0.514636	0.586220	0.649688	3.706332e-35
es	Spain	365	0.343258	0.430615	0.510587	5.901205e-18
nl	Netherlands	365	0.289832	0.381021	0.465355	4.313692e-14
ae	United Arab Emirates	365	0.036042	0.138042	0.237195	8.180446e-03
au	Australia	365	-0.171657	-0.070386	0.032358	1.790715e-01
ro	Romania	364	0.328587	0.417049	0.498258	7.807009e-17
sy	Syrian Arab Republic	364	0.002991	0.105469	0.205754	4.375154e-02
cy	Cyprus	364	-0.128416	-0.026252	0.076464	6.166602e-01
za	South Africa	364	-0.289294	-0.192493	-0.091794	2.116315e-04
pk	Pakistan	363	0.234002	0.328629	0.417088	1.152818e-10
rs	Serbia	362	0.221090	0.316429	0.405777	5.911568e-10
be	Belgium	361	0.814821	0.846615	0.873332	9.679073e-102
bg	Bulgaria	361	0.573288	0.638287	0.695303	2.854857e-43
ie	Ireland	360	-0.269506	-0.171741	-0.070471	9.706457e-04

sd	Sudan	358	0.355027	0.441470	0.520428	6.862318e-19
ng	Nigeria	358	-0.364885	-0.272570	-0.174948	1.174811e-07
ye	Yemen	357	-0.004941	0.097618	0.198145	6.209512e-02
il	Israel	354	-0.092325	0.010282	0.112673	8.445789e-01
mx	Mexico	353	0.097294	0.197831	0.294371	1.392498e-04
lb	Lebanon	351	-0.169362	-0.068033	0.034719	1.940789e-01
tn	Tunisia	350	0.529359	0.599345	0.661230	4.602306e-37
br	Brazil	345	-0.184121	-0.083180	0.019496	1.121405e-01
hr	Croatia	336	0.615940	0.675782	0.727869	3.714669e-50
sg	Singapore	326	0.340471	0.428041	0.508250	9.719579e-18
fi	Finland	316	0.350608	0.437398	0.516738	1.552419e-18
my	Malaysia	316	-0.215002	-0.115027	-0.012666	2.778068e-02
ru	Russian Federation	312	-0.028594	0.074133	0.175311	1.569667e-01
kr	Korea, Republic of	312	-0.149745	-0.047971	0.054808	3.601215e-01
nz	New Zealand	308	-0.221779	-0.122044	-0.019781	1.951278e-02
ke	Kenya	306	0.308689	0.398586	0.481426	2.185275e-15
sk	Slovakia	306	-0.085368	0.017294	0.119592	7.415977e-01
dk	Denmark	302	0.079385	0.180427	0.277799	5.234165e-04
pl	Poland	293	0.120265	0.220063	0.315456	2.158528e-05
pt	Portugal	286	0.072165	0.173392	0.271084	8.650277e-04
qa	Qatar	285	0.402448	0.484951	0.559640	5.490509e-23
ar	Argentina	279	0.053012	0.154682	0.253177	3.007348e-03
no	Norway	276	0.109093	0.209263	0.305226	5.471867e-05
gh	Ghana	273	-0.065843	0.036916	0.138901	4.813881e-01
bd	Bangladesh	264	0.030414	0.132511	0.231872	1.116088e-02
co	Colombia	262	0.049577	0.151318	0.249951	3.711371e-03
bh	Bahrain	257	-0.014738	0.087905	0.188714	9.311060e-02
id	Indonesia	250	-0.123667	-0.021428	0.081260	6.828400e-01
jp	Japan	249	-0.064817	0.037945	0.139911	4.692474e-01
si	Slovenia	246	0.253355	0.346856	0.433937	8.723828e-12
al	Albania	234	-0.259735	-0.161526	-0.060009	1.935348e-03
np	Nepal	230	0.076501	0.177618	0.275119	6.411195e-04
ua	Ukraine	228	0.078328	0.179398	0.276817	5.640093e-04
lk	Sri Lanka	220	-0.272987	-0.175385	-0.074209	7.517038e-04
th	Thailand	214	0.235085	0.329651	0.418035	1.001913e-10
cl	Chile	208	-0.293710	-0.197135	-0.096577	1.471499e-04
lu	Luxembourg	196	0.229459	0.324340	0.413115	2.065116e-10
lt	Lithuania	172	-0.050706	0.052075	0.153764	3.204558e-01
cm	Cameroon	168	-0.212116	-0.112042	-0.009643	3.212053e-02
mk	North Macedonia	160	0.026244	0.128409	0.227919	1.395614e-02
tz	Tanzania, United Republic of	137	-0.112095	-0.009697	0.092906	8.533265e-01

cn	China	132	-0.292273	-0.195624	-0.095019	1.657863e-04
ug	Uganda	110	-0.070275	0.032469	0.134531	5.357861e-01
mu	Mauritius	103	-0.047760	0.055020	0.156646	2.938187e-01
so	Somalia	81	0.005605	0.108053	0.208256	3.881375e-02
gt	Guatemala	62	-0.270711	-0.173002	-0.071765	8.889713e-04
bw	Botswana	44	0.308023	0.397967	0.480860	2.434939e-15
bz	Belize	26	-0.006229	0.096343	0.196908	6.560284e-02
gm	Gambia	20	-0.042095	0.060677	0.162179	2.468957e-01
gn	Guinea	16	-0.092304	0.010304	0.112695	8.442583e-01

Table 7: 84 Countries used for this thesis. *iso*: the iso-2 code of the country. *name*: Country name. *days*: Number of days where at least one request reached Symptoma originating from the respective country. *interval(-)*: Lower end of the confidence interval for the pearson correlation coefficient. *pcc*: The pearson correlation coefficient comparing Symptoma high risk cases with the number newly infected COVID-19 cases provided by Johns Hopkins University for the country in question. *interval(+)*: The upper end of the confidence interval for the pcc. *p-value*: p-value for the pcc, $\alpha = 0.05$.The table is primary ordered by *days* and secondary by *pcc*

rank	pcc	L_2	dtw	twed
1	cz	cz	de	cz
2	at	se	cz	de
3	fr	fr	fr	fr
4	de	tr	hu	dz
5	dz	de	dz	tr
6	hu	at	at	at
7	it	hu	se	se
8	eg	dz	tr	hu
9	sa	it	sa	gr
10	se	eg	it	ph
11	ph	ca	ph	sa
12	in	ph	eg	it
13	tr	sa	gr	in
14	gr	in	ca	iq
15	ca	iq	in	eg
16	iq	gr	iq	gb
17	gb	gb	gb	ca
18	us	us	us	us

Table 8: Ranks of the countries by distance measure. Sorted by rank.

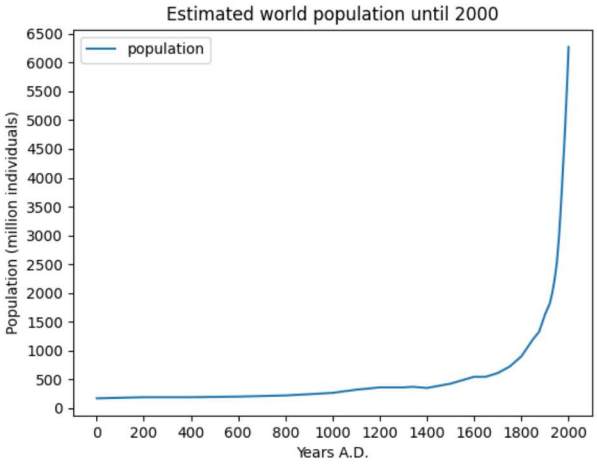


Figure 1: Estimated world population between 0 and 2.000 A.D.[7]

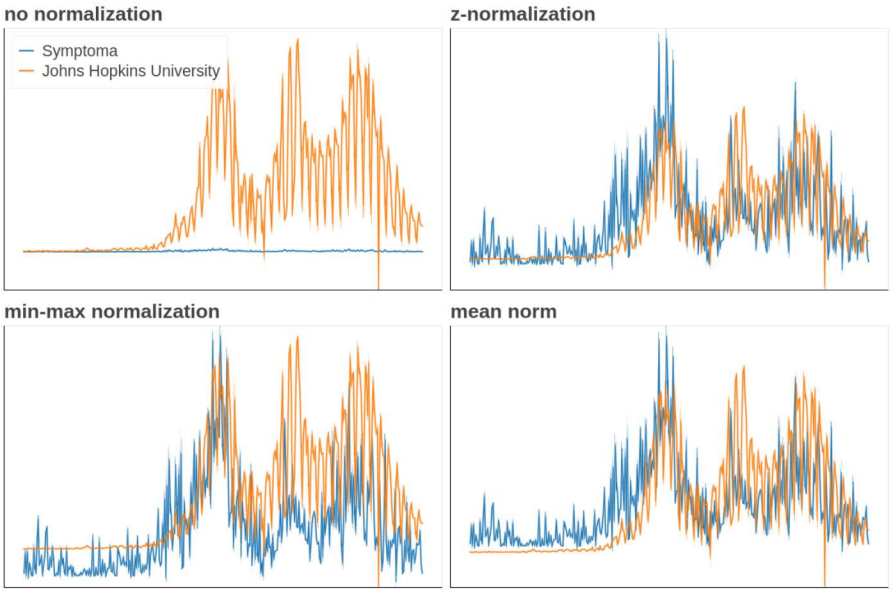


Figure 2: normalization examples for Symptoma high risk results compared with new infected cases of COVID-19 aggregated by Johns Hopkins University in Czechia.

	1	1	2	1	2	3	4	3	2	1
1	0	0	1	1	2	6	15	19	20	20
1	0	0	1	1	2	6	15	19	20	20
1	0	0	1	1	2	6	15	19	20	20
1	0	0	1	1	2	6	15	19	20	20
2	1	1	0	1	1	2	6	7	7	8
2	2	2	0	1	1	2	6	7	7	8
3	6	6	1	4	2	1	2	2	3	7

Figure 3: Example alignment of two series $x = \{1, 1, 2, 1, 2, 3, 4, 3, 2, 1\}$, $y = \{1, 1, 1, 1, 2, 2, 3\}$ using DTW. The numbers in the cell represent the cost of the respective mapping, calculated via equation (21) and (20). The blue path represents the best alignment determined by backtracking.

```
Province/State,Country/Region,Lat,Long,1/22/20,1/23/20,1/24/20,1/25/20,1/26/20,1/27/20
,Afghanistan,33.93911,67.709953,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
,Albania,41.1533,20.1683,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
,Algeria,28.0339,1.6596,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
,Andorra,42.5063,1.5218,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
,Angola,-11.2027,17.8739,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
,Antigua and Barbuda,17.0608,-61.7964,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
,Argentina,-38.4161,-63.6167,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
,Armenia,40.0691,45.0382,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
Australian Capital Territory,Australia,-35.4735,149.0124,0,0,0,0,0,0,0,0,0,0,0,0,0,0
New South Wales,Australia,-33.8688,151.2093,0,0,0,0,3,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,
Northern Territory,Australia,-12.4634,130.8456,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
Queensland,Australia,-27.4698,153.0251,0,0,0,0,0,0,0,1,3,2,3,2,2,3,3,4,5,5,5,5,5,5,5,5
South Australia,Australia,-34.9285,138.6007,0,0,0,0,0,0,0,0,0,0,1,2,2,2,2,2,2,2,2,2,
Tasmania,Australia,-42.8821,147.3272,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
Victoria,Australia,-37.8136,144.9631,0,0,0,0,1,1,1,1,2,3,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4
Western Australia,Australia,-31.9505,115.8605,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
```

Figure 4: Created June 15th 2021. This Screenshot shows part of the source data from the file *"time_series_covid19_confirmed_global.csv"* [57]

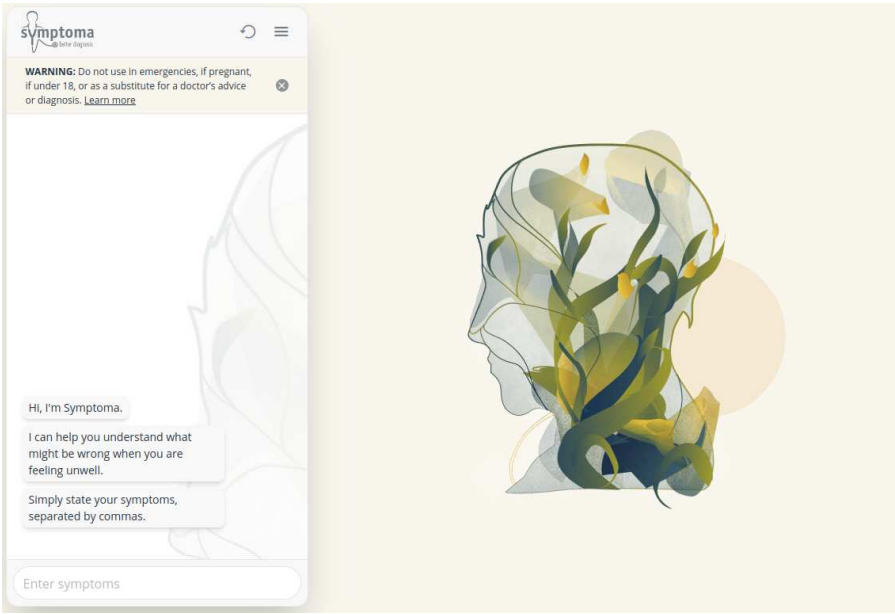


Figure 5: Startpage of symptoma.com

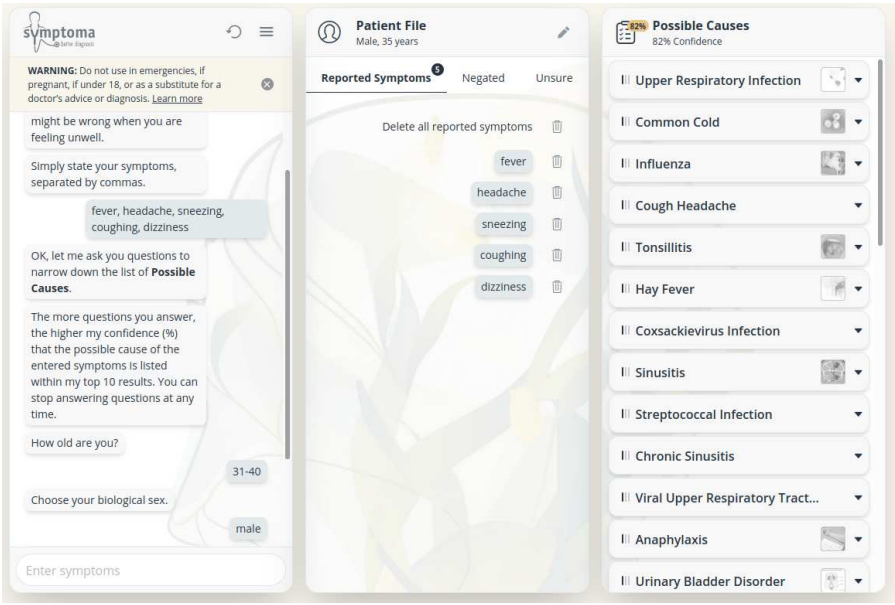


Figure 6: Assessment provided by symptoma.com after enough information has been gathered.

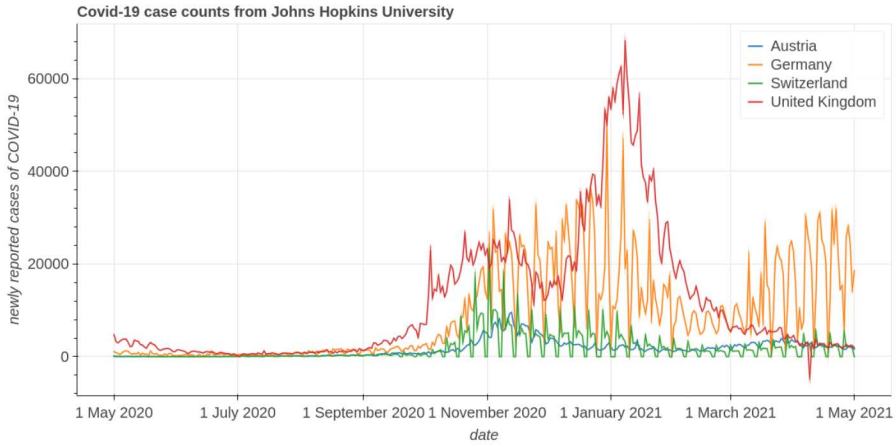


Figure 7: Non normalised *JH* for Germany, Switzerland, United Kingdom and Austria

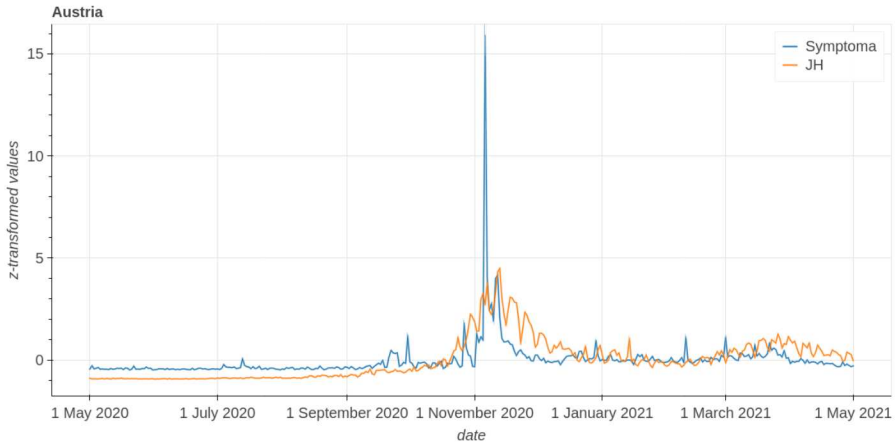


Figure 8: Comparison of z-nomalized *Symptoma data* vs *JH* for Austria

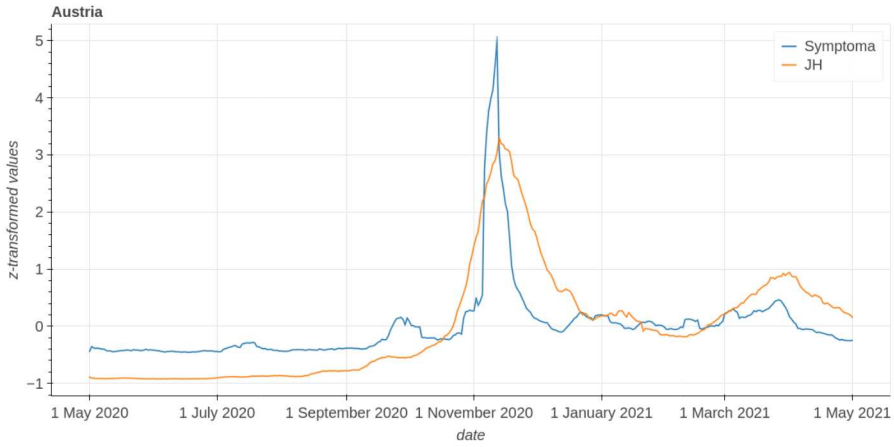


Figure 9: Comparison of fully preprocessed *Symptoma* vs *JH* for Austria

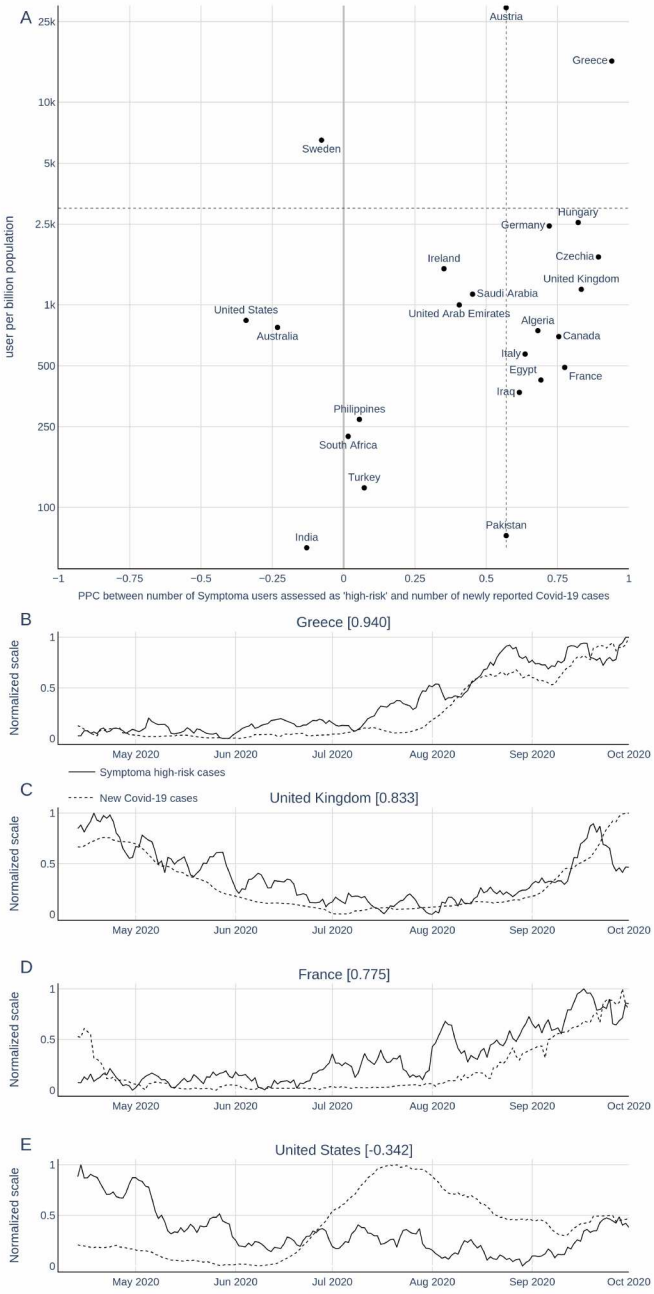


Figure 24: Figure 1 from [64]

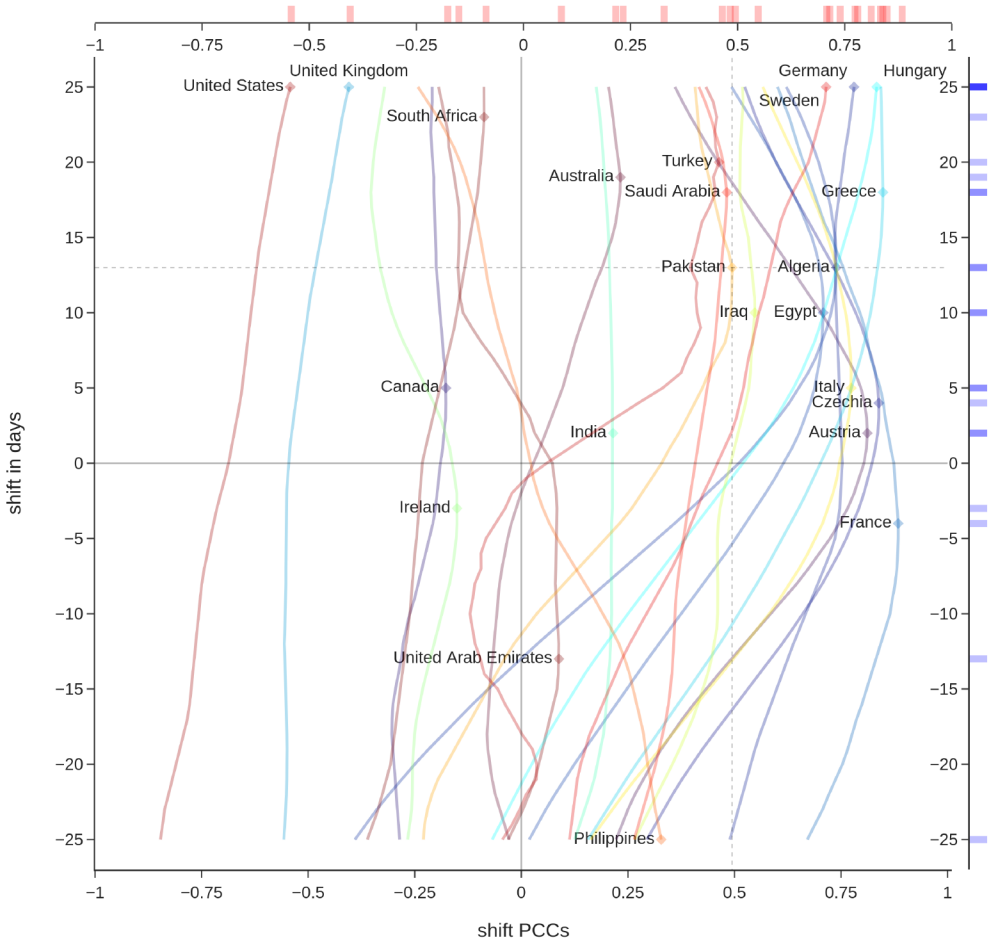


Figure 25: Figure 2 from [64]

Diese Masterarbeit untersucht die Ähnlichkeit zwischen online Testdaten bestätigten globalen COVID-19-Fällen. Erstere wurden durch den Online-Symptom-Checker Symptoma generiert und letztere von der Johns Hopkins University (JHU) aggregiert. Diese Arbeit trägt zur wissenschaftlichen Literatur im Bereich der Infodemiologie[11] bei, einer Disziplin, die sich mit der Nutzung internetbasierter Daten zur Überwachung und Analyse epidemiologischer Entwicklungen befasst.

Die zentralen Ziele dieser Arbeit sind es, zu analysieren, inwiefern der Online-Symptom-Checker Symptoma spezifische Zeitreihen für unterschiedliche Krankheitsbilder erzeugt und ob diese mit den offiziell berichteten COVID-19-Fällen in einer systematischen Beziehung stehen. Es wurden die 40 häufigsten Krankheitsvorschläge durch Symptoma in Österreich ausgewählt. Für jede dieser Erkrankungen wurden Zeitreihen generiert, indem die Anzahl der NutzerInnen aggregiert wurde, bei denen die jeweilige Krankheit unter den 30 wahrscheinlichsten Ursachen von Symptoma aufgeführt wurde. Anschließend wurden diese Zeitreihen mit den bestätigten COVID-19-Fällen verglichen, wobei der Pearson-Korrelationskoeffizient (PCC) zur Bestimmung der Ähnlichkeit herangezogen wurde. Zudem wurde der Überlappungskoeffizient der Symptome berechnet, um strukturelle Gemeinsamkeiten zwischen den Krankheitsbildern zu quantifizieren. Krankheiten mit einem niedrigen Symptom-Überlappung mit COVID-19 zeigten ebenfalls eine niedrige Korrelation mit der Zeitreihe aus durch Symptoma identifizierten COVID-19 Fällen. Selbst Atemwegserkrankungen, die wie erwartet eine hohe Symptom-Überlappung aufweisen zeigen einen PCC von maximal 0,44. Demnach sind die Zeitreihen generiert mit den Nutzerdaten des Online-Symptom-Checker Symptoma unter sich einzigartig. Diese Ergebnisse zeigen, dass Symptoma selbst ähnliche Krankheiten verlässlich voneinander unterscheiden kann.
Die Datenbasis der zweiten Analyse umfasst aggregierte Nutzungsdaten von Symptoma sowie Fallzahlen der JHU, die den Zeitraum vom 1. Mai 2020 bis zum 1. Mai 2021 abdecken. Um zeitliche Verschiebungen in den Fallzahlen zu berücksichtigen, wurde der PCC für eine optimale Verschiebung (lag) zwischen -30 und +30 Tagen berechnet. Diese Berechnungen erfolgten für 18 von 84 untersuchten Ländern, wobei die Auswahl der Länder auf der Verfügbarkeit durchgängig erfasster Daten über 366 Tage basierte. Die Analyse ergab eine durchschnittliche zeitliche Verschiebung von +10 Tagen sowie eine mittlere Korrelation von 0,55 zwischen den Symptoma-Daten und den COVID-19-Fallzahlen der JHU in den ausgewählten Ländern. Diese Ergebnisse wurden durch zusätzliche Ähnlichkeitsmaße bestätigt, darunter die euklidische Distanz (L2-Norm), das Dynamic Time Warping (DTW) sowie die Time Warped Edit Distance (TWED).
Die festgestellte positive mediane Verschiebung weist darauf hin, dass Fälle infektiöser Erkrankungen tendenziell früher in Symptoma-Daten sichtbar werden als in den offiziell gemeldeten Fallzahlen. Daraus lässt sich ableiten, dass Online-Symptom-Checker wie Symptoma potenziell als Frühwarnsysteme für aufkommende Krankheitsausbrüche genutzt werden können. Gleichzeitig werden in dieser Arbeit die Grenzen einer solchen Anwendung diskutiert. Es konnte beobachtet werden, dass die Korrelation zwischen den offiziellen Fallzahlen und den Symptoma-Daten nach der zweiten Infektionswelle in den meisten Ländern signifikant abnahm. Dies wird als Hinweis auf eine zunehmende Ermüdung in der Bevölkerung hinsichtlich der COVID-19-Berichterstattung interpretiert, ein Phänomen, das auch in anderen wissenschaftlichen Arbeiten dokumentiert wurde[36].
Die Ergebnisse dieser Arbeit verdeutlichen das Potenzial digitaler Gesundheitsdaten für die epidemiologische Überwachung, zeigen aber gleichzeitig die Notwendigkeit einer differenzierten Betrachtung der Limitationen und methodischen Herausforderungen auf. Zukünftige Forschungen sollten sich darauf konzentrieren, die Übertragbarkeit der Ergebnisse für weitere geografische Regionen und Krankheitsbilder zu untersuchen sowie alternative algorithmische Ansätze zur Optimierung der Datenanalyse einzusetzen.
Durch die Verbindung von internetbasierten Gesundheitsdaten mit epidemiologischen Modellierungen leistet diese Arbeit einen Beitrag zur Weiterentwicklung digitaler Überwachungsmethoden und bietet eine Grundlage für die mögliche Implementierung von Online-Symptom-Checkern als unterstützende Werkzeuge im Bereich des öffentlichen Gesundheitswesens.