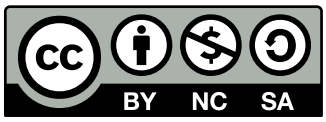
Aalto University
School of Science

Master's Programme in Mathematics and Operations Research

On clustering of countries based on causes of mortality

Antti Ahvenjärvi

Master's Thesis
2026



Aalto University
School of Science

Author Antti Ahvenjärvi

Title On clustering of countries based on causes of mortality

Degree programme Mathematics and Operations Research

Major Systems and Operations Research

Supervisor Prof. Pauliina Ilmonen

Advisor Prof. Pauliina Ilmonen

Date 6 February 2026 **Number of pages** 74 **Language** English

Abstract

Comparative analysis of health data across countries helps understand inequalities in health outcomes and evaluate the effects of health policies. Mortality profiles are defined by cause-specific mortality rates, and they reflect the living conditions within a country. Thus, using mortality profiles to group countries with similar mortality patterns can provide meaningful information about the similarities and differences in health outcomes and living conditions across countries. When multiple causes of death are considered in the analysis, the setting is inherently multivariate, suggesting the use of multivariate data analysis methods.

This thesis examines clusters of countries identified from cause-specific mortality data across 26 countries, mainly from Europe, Latin America, and North America. Clustering is performed separately for three years: 2001, 2018, and 2021. The data used for clustering consist of gender-specific age-standardized death rates for five high-level causes of death, including neoplasms, accidents and assaults, diseases of the circulatory system, diseases of the respiratory system, and intentional self-harm. The clusters are derived from the data by applying agglomerative hierarchical clustering using Ward's method. For each considered year, two-, three-, and four-cluster solutions are examined and compared between the years. To assess the socioeconomic similarity within the identified clusters, the Human Development Index (HDI) is used as an external validation measure.

The clustering results suggest that interpretable groups of countries in terms of geographical and socioeconomic similarity can be identified from the mortality clusters. Across all examined years, Eastern European and Latin American countries are often clustered into distinct clusters. Countries within clusters also tend to have similar HDI levels, with high-HDI countries clustering together almost always. The changes in the cluster composition from 2018 to 2021 are driven by increased mortality from respiratory diseases, potentially linked to the COVID-19 pandemic. However, this connection is not extensively studied in this thesis. Overall, this thesis demonstrates the usefulness of clustering in comparative health analysis.

Keywords Age-standardized death rates, agglomerative hierarchical clustering, cause-specific mortality, Human Development Index, ICD-10, Ward's method

Aalto University
School of Science

Tekijä Antti Ahvenjärvi

Työn nimi Maiden klusteroinnista kuolinsyydatan perusteella

Koulutusohjelma Matematiikka ja operaatiotutkimus

Pääaine Systeemi- ja operaatiotutkimus

Työn valvoja Prof. Pauliina Ilmonen

Työn ohjaaja Prof. Pauliina Ilmonen

Päivämäärä 6. helmikuuta 2026 **Sivumäärä** 74 **Kieli** englanti

Tiivistelmä

Terveyteen liittyvän datan analysoinnilla on keskeinen rooli, kun halutaan ymmärtää maiden välisiä eroja ja arvioida terveyspoliittisten päätösten vaikutuksia. Kuolleisuusluvut ovat tärkeitä maiden elinolosuhteita arvioitaessa, ja näin ollen maiden ryhmittely kuolleisuusdatan perusteella voi antaa merkityksellistä tietoa maiden välisistä eroista ja samankaltaisuuksista. Kun ryhmittelyssä käytetty kuolleisuusdata sisältää useamman kuin yhden kuolinsyyn, voidaan puhua moniulotteisesta datasta, jonka analysointiin käytetään siihen soveltuvia menetelmiä.

Tämä diplomityö tutkii kuolinsyykohtaisesta kuolleisuusdatasta löydettyjä maaklustereita vuosille 2001, 2018 ja 2021. Klusterianalyysin kohteena on 26 maata, joiden joukossa on maita niin Euroopasta, Latinalaisesta Amerikasta kuin Pohjois-Amerikastakin. Työssä käytetty kuolleisuusdata muodostuu ikävakioiduista kuolleisuusluvuista molemmille sukupuolille ja viidelle ylätason kuolinsyylle: syövät, onnettomuudet ja pahoinpitelyt, hengityselimistön sairaudet, sydän- ja verenkiertoelimistön sairaudet sekä itsemurhat. Klusterointi suoritetaan kasaavalla hierarkkisella klusteroinnilla käyttämällä Wardin menetelmää, ja saaduista klusteriratkaisuista analysoidaan tarkemmin kahden, kolmen ja neljän klusterin ratkaisut jokaiselle analysoitavalle vuodelle. Lisäksi työssä arvioidaan kuolleisuusdatan perusteella muodostettujen klustereiden sosioekonomisia samankaltaisuuksia vertaamalla klusteroitujen maiden inhimillisen kehityksen indeksin arvoja.

Analyysin tuloksena huomattiin, että kuolleisuusdatasta muodostetut klusterit sisältävät usein maantieteellisesti ja sosioekonomisesti samankaltaisia maita. Esimerkiksi Itä-Euroopan ja Latinalaisen Amerikan maat klusteroituvat usein omiin klustereihinsa, ja samassa klusterissa olevilla mailla on myös monesti keskenään samankaltaiset arvot inhimillisen kehityksen indeksillä mitattuna. Vuoden 2021 klusterirakenteessa havaitaan muutos verrattuna aiempiin tarkasteltuihin vuosiin. Tämän muutoksen taustalla on mahdollisesti COVID-19 pandemian aiheuttama keuhkotautikuolleisuuden kasvu, mutta pandemian tarkempi tarkastelu tässä yhteydessä on kuitenkin jätetty työn rajauksen ulkopuolelle. Kaikkiaan tässä työssä saadut tulokset havainnollistavat klusteroinnin käytettävyyttä vertailevassa terveysdatan analysoinnissa.

Avainsanat ICD-10, ikävakioitu kuolleisuus, inhimillisen kehityksen indeksi, kasaava hierarkkinen klusterointi, kuolinsyykohtainen kuolleisuus, Wardin menetelmä

Contents

1 Introduction

Examining health data across countries is essential for identifying inequalities and analyzing the outcomes of health policies. Although global life expectancy has increased over time, substantial differences in health outcomes and mortality profiles between countries are still observed (Aksan and Chakraborty, 2023). As the mortality profiles can be seen as the result of the living conditions within a country, comparative analysis of these profiles can provide valuable insights into assessing different health care systems and laying the foundation for future development.

Traditional mortality analysis is based on analyzing and comparing individual indicators, such as life expectancy and cause-specific death rates (Léger and Mazzuco, 2021). When more than one indicator is considered, the setting turns multivariate. In multivariate data analysis, comparisons between single indicators often fail to fully describe the similarities and differences within the data set. To perform comparative analysis in this multivariate setting, cluster analysis can be applied. Cluster analysis is a tool for finding groups that lie within a data set without the need for external parameters or data (Härdle and Simar, 2015). Thus, it can be, for example, used to find clusters of countries by simultaneously taking into account multiple health indicators, such as causes of death.

This thesis applies hierarchical clustering to explore similarities in cause-specific mortality across countries and over time. The analysis was performed using gender-specific age-standardized death rates (ASDRs) for five high-level causes of death obtained from the World Health Organization Mortality Database. The examined causes of death are neoplasms, accidents and assaults, diseases of the circulatory system, diseases of the respiratory system, and intentional self-harm. These causes are selected for the analysis as they reflect various aspects of living conditions, such as safety and psychological well-being, as well as cover the leading causes of mortality. Age standardization enables meaningful comparisons across countries with differing age structures.

The primary objective of this thesis is to investigate how countries are clustered based on cause-specific mortality and to examine how these clusters evolve over time. The analysis focuses on three years: 2001, 2018, and 2021. The data set includes data from 26 countries, including those from Europe, Latin America, and North America. Analyzing the geographical and socioeconomic similarities within the clusters is of particular interest. To evaluate the socioeconomic similarity, we use the Human Development Index (HDI) to externally validate the clusters formed from the mortality data.

In this thesis, AI (ChatGPT) is used for grammatical checks and improving the flow of the text.

The remainder of this thesis is organized as follows. Section 2 introduces the ICD-10 classification system used to code causes of death and covers relevant literature on mortality patterns and on the selected causes of death for this study. Section 3 outlines the clustering methodology and cluster validation methods. Section 4 presents the data and related preprocessing. Section 5 presents the clustering results for each considered year and examines how the clusters align with HDI. Section 6 discusses

the findings and limitations. Finally, Section 7 concludes the thesis by summarizing the main results.

2 Causes of mortality

2.1 Recording of deaths and ICD classification

Gathering reliable cause of death (COD) data is crucial for making informed decisions and plans within health-related systems and for driving appropriate policies (Cobos Muñoz et al., 2018). The medical registration of deaths is handled by the civil registration and vital statistics (CRVS) systems in each country. Although the methods and data quality vary considerably across national CRVS systems, particularly in developing countries, ongoing efforts aim to strengthen and standardize these processes (Cobos Muñoz et al., 2020). Having data in a standard format is vital not only for following the national progress but also for making inter-country comparisons in the field of health statistics.

The first regarded steps toward standard COD classification were taken in the 1700s by Francois Boissier de Sauvages, who presented a taxonomy with 10 major classes of diseases (Boissier de Sauvages, 1763). With the requirement for statistically comprehensive COD data, approaches in classification evolved, and a first *International Classification Of Causes Of Deaths* (ILCD) was built upon the work by a committee led by Jacques Bertillon (Moriyama et al., 2011). In 1948, the World Health Organization (WHO) took charge of ILCD and, by incorporating causes of morbidity into the system, gave life to the current standard known as *International Classification Of Diseases* (ICD) (Hirsch et al., 2016).

Since then, ICD has become the standard for classifying health data, and it was developed with the purpose of systematically recording mortality and morbidity data to allow practical data analysis, including inter-country comparisons (World Health Organization, 2019). The ICD system is regularly updated and revised by WHO, with the most widely used standard today being ICD-10, the 10th version of ICD (Otero Varela et al., 2021), while the adaptation of the latest version, ICD-11, is still in progress. Thus, the data used in this thesis follows the ICD-10 system.

The WHO standards require its members to organize national medical certification of all deaths using the ICD (De Savigny et al., 2017). However, there are still countries in which a considerable number of deaths are not recorded by CRVS systems. According to World Health Organization (2021), approximately 40% of deaths are not registered annually, highlighting significant deficiencies in civil registration and vital statistics systems. This is common in low-income countries, where standard COD death data mainly cover in-hospital deaths in urban areas, making the data not representative of the whole nation (Murray et al., 2007). On the other side of the spectrum, there are countries such as Finland where the medically certified death coverage is practically 100 percent (Official Statistics of Finland (OSF), 2020).

2.2 ICD-10

ICD-10 contains 22 chapters that define the 'family' of diagnoses. The chapters are further divided into blocks of homogeneous categories. Inside a block, each category defines either a single condition or a block of conditions with common characteristics.

The top-level chapters in ICD-10 are presented in Table 1, and Figure 1 depicts the taxonomy in the system by demonstrating how lung cancer is coded.

Registering a death according to ICD-10 requires filling out a medical certification that contains, for example, the date of birth, date of death, COD, manner of death (accident, suicide, etc.), and the signature of the physician or medical examiner. The WHO instruction manual provides details and procedures for implementing and applying the ICD-10 classification (World Health Organization, 2019). Only medically certified deaths are accepted and published in the WHO mortality database, which is the primary source for data used in this thesis.

Table 1: Chapters of ICD-10 system.

Number	Title
I	Certain infectious and parasitic diseases
II	Neoplasms
III	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
IV	Endocrine, nutritional and metabolic diseases
V	Mental and behavioural disorders
VI	Diseases of the nervous system
VII	Diseases of the eye and adnexa
VIII	Diseases of the ear and mastoid process
IX	Diseases of the circulatory system
X	Diseases of the respiratory system
XI	Diseases of the digestive system
XII	Diseases of the skin and subcutaneous tissue
XIII	Diseases of the musculoskeletal system and connective tissue
XIV	Diseases of the genitourinary system
XV	Pregnancy, childbirth and the puerperium
XVI	Certain conditions originating in the perinatal period
XVII	Congenital malformations, deformations and chromosomal abnormalities
XVIII	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
XIX	Injury, poisoning, and certain other consequences of external causes
XX	External causes of morbidity and mortality
XXI	Factors influencing health status and contact with health services
XXII	Codes for special purposes

2.3 Selected CODs

In order to cluster countries based on COD data, the causes to be considered must be selected. To get a well-rounded and top-level view of different causes of mortality, the following five causes have been selected for our cluster analysis: neoplasms, diseases of the circulatory system, accidents and assaults, diseases of the respiratory system,

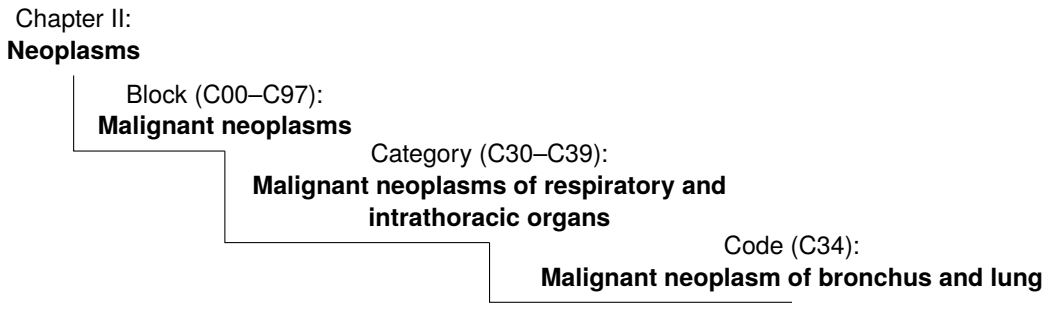


Figure 1: ICD-10 hierarchy for malignant neoplasms down to code C34 (lung cancer).

and intentional self-harm. Neoplasms and diseases of the respiratory and circulatory systems have been among the leading causes of death in the 21st century (Mathers et al., 2017). Accidents are also a major cause of death, with Heron (2018) reporting that they ranked among the top four leading causes of death in the United States regardless of the racial group in 2016. Assaults were included alongside accidents in the cluster analysis, as both serve as indicators of the overall safety within a country.

To further increase the variety of causes considered, intentional self-harm was selected to highlight psychological well-being within a society. The next five sections give more detailed descriptions of all the selected CODs.

2.3.1 Neoplasms

Neoplasms, commonly referred to as tumors, cover conditions characterized by abnormal tissue growth that often realizes as masses within the body (Kumar et al., 2005). This uncontrolled growth results from genetic alterations at the cellular level that disrupt normal regulatory mechanisms, leading to excessive cell proliferation. In the ICD-10 classification system, neoplasms are included in Chapter II and are categorized into four blocks: malignant, in situ, benign, and neoplasms of uncertain or unknown behavior. The most recognized form of neoplasms is malignant neoplasms, commonly known as cancers (National Cancer Institute).

The study of causes of cancer has been an interest in the medical field throughout history, with an aim to identify carcinogens (Blackadar, 2016). The list of known carcinogens includes, for example, smoking, alcohol use, diets with low fruit and vegetable intake, and certain viruses (Danaei et al., 2005). At the individual level, cancer prevention involves avoiding exposure to carcinogens, while at the population level, screening programs have proven to be successful national preventative measures (Wild et al., 2019).

The number of cancer-related deaths has increased over time and is expected to continue rising. Force et al. (2025) estimate a growth of 74,5% in global cancer mortality from 2024 to 2050. This trend is driven by population growth and population aging, given that cancer is more common among the elderly population (Are et al., 2013). However, when age-standardization is used, we observe that in more developed countries, the cancer mortality has actually stalled or decreased in recent years (Torre et al., 2016). This phenomenon is driven by increased cancer screening and advances

in cancer treatment. A similar level of preventative measures is not available in the low-income countries, thus resulting in an increasing number of cancer-related deaths.

In 2022, the most common causes of cancer mortality were lung, liver, and stomach cancer among males, and breast, lung, and cervix uteri cancer among females (Bray et al., 2024).

2.3.2 Diseases of the circulatory system

Diseases of the circulatory system cover a wide range of conditions affecting the heart and blood vessels, collectively known as cardiovascular diseases (CVD). CVDs remain the leading causes of mortality globally, with ischemic heart disease ranked as the top cause of death in 2021 (Naghavi et al., 2024). Strokes, caused by blockage of blood flow to the brain or by intracranial bleeding, are also among the leading causes of death (Qian et al., 2025). In 2019, ischemic heart disease and stroke together accounted for 85% of deaths due to CVDs Di Cesare et al. (2024). In the ICD-10 classification system, CVDs are included in Chapter IX.

Although the absolute number of CVD deaths has risen over time, mainly due to population growth and aging, the global age-standardized death rates have been declining in recent decades (Ezzati et al., 2015). This decline has been caused by advances in prevention, diagnosis, and treatment, as well as improved understanding of risk factors including smoking, obesity, high LDL cholesterol, and high blood pressure (Mendis et al., 2011). In particular, blood pressure control has been identified as one of the most important practices in reducing cardiovascular mortality (Lackland et al., 2014). However, advances in preventive measures are not equally experienced across countries. Pu et al. (2023) estimates that low-income countries will experience an increase in age-standardized CVD-related mortality rates from 2020 to 2030.

2.3.3 Diseases of the respiratory system

The diseases of the respiratory system are among the leading causes of death globally (Naghavi et al., 2024). The respiratory system, which includes the lungs, bronchi, trachea, and associated structures, can be affected by a range of diseases that are categorized as either noncommunicable or communicable. The diseases of the respiratory system are contained in Chapter X of the ICD-10 classification system, with additional codes for COVID-19 and SARS included in Chapter XXII.

Noncommunicable respiratory diseases include many chronic conditions, such as chronic obstructive pulmonary disease, which was the third leading cause of death in 2019 (Naghavi et al., 2024). The main environmental risk factors for these conditions are tobacco smoking and air pollution (Ahmed et al., 2017). Communicable diseases, on the other hand, include infectious diseases often caused by viruses or bacteria. Major communicable respiratory diseases cover infectious conditions of the lower respiratory system, including pneumonia, bronchitis, and influenza (Safiri et al., 2023). With the emergence of the COVID-19 pandemic, respiratory infections gained global attention, and COVID-19 became the second leading cause of death in 2021 (Naghavi et al., 2024).

When the excess mortality due to COVID-19 estimated by Wang et al. (2022) to be 18,2 million between 2020 and 2021, is not taken into account, the mortality from communicable respiratory diseases has seen a major decline since 1990 (Bender et al., 2024). However, the pace of this decline has varied significantly across countries. Mortality rates from infectious respiratory diseases remain substantially higher in low-income countries, where access to healthcare, vaccination coverage, and hygiene infrastructure is limited (Safiri et al., 2023).

2.3.4 Accidents and assaults

The rate of accident and assault related mortality can be seen as a straightforward indicator of the level of safety within a society. Globally, the mortality due to assaults and accidents has been greater among males than females (Haagsma et al., 2016). In the ICD-10 classification system, these causes of death fall under Chapter XX: External causes of morbidity and mortality. Accidental causes correspond to codes V01-X59, Y40-Y86, Y88, and Y89, but, similarly to the WHO interactive mortality database World Health Organization (2025c), we exclude codes X41, X42, X44, and X45, which relate to drug- and alcohol-related poisonings. Accidents include causes such as road traffic injuries, falls, fires, and drownings. Assault related deaths are covered by codes X85-Y09 and Y87.1, with different codes reflecting the manner of the assault.

Injuries from accidents and assaults are a major cause of premature death, covering approximately 8% of all deaths during 2000–2019 (World Health Organization, 2023). In the United States, accidental injuries have been the leading cause of death among adolescents, with road traffic accidents being the predominant cause (West et al., 2021). Globally, adolescent mortality from road traffic accidents has declined from 1990 to 2019 (Khan et al., 2021). This trend has been observed across all income levels, but improvements have been more pronounced in high-income countries. Among elderly people, accidental falls are the leading cause of injury-related mortality (James et al., 2020).

In contrast, assault related mortality is divided heterogeneously, with over 50% of all firearm deaths occurring in Brazil, the United States, Mexico, Colombia, Guatemala, and Venezuela (Naghavi et al., 2018). High levels of violent mortality in several Latin American countries are strongly associated with organized crime and drug trafficking (Bisca et al., 2024). In the United States, firearm mortality has been linked to regulatory environments with stricter firearm legislation, particularly background checks, found to be associated with lower firearm-related deaths (Kalesan et al., 2016).

2.3.5 Intentional self-harm

Suicide, defined as death caused by intentional self-harm, represents a significant global public health challenge. Globally, suicides account for approximately 1% of all deaths and constitute the third leading cause of death among individuals aged 15–29 (World Health Organization, 2025b). In the ICD-10 classification system, intentional

self-harm is listed in Chapter XX under codes X60–X84 and Y87.0, which specify the methods involved in the suicide.

Suicide mortality has consistently been higher among males than females, despite females reporting higher rates of suicidal ideation and non-fatal attempts. This pattern is commonly referred to as the gender paradox (Bennett et al., 2023). Globally, suicide rates declined between 1990 and 2019 (Yip et al., 2022).

Multiple factors have been observed to contribute to suicide risk. Mental health disorders, particularly depressive symptoms, internal entrapment, and perceived burdensomeness, are strongly associated with suicidal behavior (De Beurs et al., 2019). In addition, social and environmental factors play a major role. Na et al. (2025) identified childhood maltreatment, divorce, and exposure to parental suicide as significant risk factors, while marriage and religious affiliation were identified as protective factors. Despite its complexity, suicide is considered preventable, with effective strategies including improved mental health care and restriction of access to means (Mann et al., 2021).

2.4 Clusters in mortality data

Clustering countries based on mortality data can be used to identify groups with similar epidemiological profiles and to guide the direction of health policies and public health interventions. Such clustering provides insight into how differences in socioeconomic conditions and health care systems contribute to differences in mortality outcomes across countries.

A central component of comparative analysis is examining whether clusters align with economic, developmental, or social characteristics. Economic status is often measured by gross domestic product (GDP) per capita, while composite indices offer a broader perspective. For example, the Global Burden of Disease (GBD) study introduced the Socio-demographic Index (SDI), which summarizes income per capita, education level, and fertility into a single measure that has a value between zero and one (Kassebaum et al., 2016). Similarly, the Human Development Index (HDI), first introduced by UNDP (1990), combines life expectancy, education, and income for assessing the level of development of a country. Geographic proximity is also an interesting factor when analyzing the clustering results, as neighboring countries can share similar cultural, environmental, and health system characteristics.

The literature on clustering of mortality data covers diverse methodological approaches with different perspectives. Léger and Mazzuco (2021) performed functional clustering of age-specific mortality curves for 32 countries from 1960 to 2018, finding consistent declines in infant mortality and a shift and compression of deaths toward older ages. Among the studied countries, the Nordic countries were observed to lead this transition, while the Eastern European countries lagged. Lotrič Dolinar et al. (2019) clustered EU countries using age-standardized death rates by sex and age group for neoplasms, cardiovascular diseases, and respiratory diseases. They identified four clusters that aligned geographically and largely followed an East–West division. Işikhan and Güleç (2018) applied hierarchical clustering for 146 countries using cause of death and health-risk data, resulting in six regionally organized clusters

that highlighted the roles of climate and ethnicity rather than socioeconomic factors in the cluster separation. In the context of the COVID-19 pandemic, Rizvi et al. (2021) demonstrated that the prevalence of diseases such as tuberculosis, respiratory infections, and cardiovascular diseases was strongly linked with pandemic-related mortality clusters.

Studies of COD often discuss the phenomenon known as the epidemiological transition, characterized by a global shift from communicable to noncommunicable diseases (Santosa et al., 2014). Regions with higher levels of development tend to exhibit a greater proportion of mortality from noncommunicable diseases, whereas communicable, maternal, and perinatal mortality remain disproportionately high in areas with lower-development (Wang and Wang, 2020). These differences reflect inequalities in access to healthcare, vaccination coverage, public health infrastructure, and general living conditions.

3 Clustering

When having multivariate data of individuals, it might be of interest to examine whether homogeneous groups, i.e., clusters, can be formed from those individuals in such a way that the similarity within a cluster and the differences between clusters are maximized according to some measure (Härdle and Simar, 2015). Finding such clusters is interesting across a variety of applications, from customer segmentation to image analysis to, in our case, identifying groups of similar mortality patterns among countries. Clustering is used to derive information from the data itself, rather than relying on external factors.

The proximity of two individuals in a multivariate setting is a matter of the used metric. Thus, the clusters found from a dataset are not unique and depend on the applied similarity and distance measures and the algorithm used (Shirkhorshidi et al., 2015). Algorithms can be divided into categories based on how they seek to form the clusters. At a high level, clustering methods can be divided into partitioning and hierarchical methods. Hierarchical clustering provides a hierarchy of clusters without a predetermined number of clusters, whereas in partitioning clustering, the number of clusters is specified as an input. (Madhulatha, 2012) In this thesis, we use hierarchical clustering, Ward's method in particular.

The principles of hierarchical clustering are presented in Section 3.2. Section 3.1 covers different similarity and distance measures.

3.1 Proximity measures

In order to form groups of similar data objects from a multivariate dataset, one has to define what is meant by the notion 'similar'. The similarity is defined by a measure $d(x_i, x_j)$ that returns a value indicating the proximity of the data points x_i and x_j, which are feature vectors for objects i and j. The measure is evaluated for each pair of objects resulting in a proximity matrix $\mathcal{D}(n \times n)$, where n is the number of objects in the dataset and each entry, d_{ij}, in the matrix stands for the value of the measure, i.e., $d(x_i, x_j)$. (Härdle and Simar, 2015)

The proximity measure can be classified as a similarity or a distance measure. In the case of a similarity measure, a larger value of $d(x_i, x_j)$ indicates greater similarity, while with a distance measure, a larger value indicates greater dissimilarity. (Mehta et al., 2020) There exists a variety of measures for different types of data, including binary, nominal, and ordinal data. However, given the nature of the data used in this thesis, we focus only on measures for numeric data.

3.1.1 Similarity measures

For a pair of data objects in a dataset, a similarity measure outputs a value indicating the degree of similarity between the objects. Formally, a similarity measure d for a dataset $\mathcal{X}$ can be defined through the following properties (Lesot et al., 2009):

$$\text{Positivity:} \quad d(x_i, x_j) \geq 0, \quad \forall i, j \in \mathcal{X}. \tag{1}$$

$$\text{Symmetry:} \quad d(x_i, x_j) = d(x_j, x_i), \quad \forall i, j \in \mathcal{X}. \tag{2}$$

$$\text{Maximality:} \quad d(x_i, x_i) \geq d(x_i, x_j), \quad \forall i, j \in \mathcal{X}.$$

In addition to these definitions, similarity measures are sometimes normalized to have values in the interval $[0, 1]$. An example of a similarity measure is normalized cosine similarity, which measures the similarity of objects i and j as the cosine of the angle between the corresponding feature vectors x_i and x_j:

$$d(x_i, x_j) = \frac{1}{2}\left(\frac{x_i^T \cdot x_j}{\|x_i\| \, \|x_j\|} + 1\right).$$

Similarity measures can also be derived from distance measures. Because distance is zero when two objects are identical, the similarity measure can be defined as: $d = 1 - d'$, where d' is a normalized distance measure.

3.1.2 Distance measures

Distance measures are functions that map the dissimilarity of a pair of data objects to a numeric value. A distance function d is defined on a dataset $\mathcal{X}$ to satisfy conditions (1), (2), and it is defined as a proper metric if it also satisfies the following conditions (Xu and Wunsch, 2005):

$$\text{Triangle inequality:} \quad d(x_i, x_j) \leq d(x_i, x_k) + d(x_k, x_j), \quad \forall i, j, k.$$

$$\text{Reflexivity:} \quad d(x_i, x_j) = 0, \quad \text{iff} \quad x_i = x_j.$$

Often used distance functions include Euclidean distance, which is defined as:

$$d(x_i, x_j) = \left(\sum_{p=1}^{m} |x_{ip} - x_{jp}|^2\right)^{\frac{1}{2}},$$

where m is the number of features in the data vector x_i for all i. Manhattan distance, also known as the L_1 norm, is defined as the sum of absolute differences of the features between data points x_i and x_j:

$$d(x_i, x_j) = \sum_{p=1}^{m} |x_{ip} - x_{jp}|.$$

Another widely used distance measure is the Mahalanobis distance. It is defined as:

$$d(x_i, x_j) = \sqrt{(x_i - x_j)^T \mathbf{\Sigma}^{-1} (x_i - x_j)},$$

where $\mathbf{\Sigma}^{-1}$ is the inverse of the data covariance matrix. By using the covariance matrix, the Mahalanobis distance takes into account the feature correlations. (Lesot et al., 2009)

In this thesis, we use Euclidean distance in the clustering process. The Euclidean distance is sensitive to variables with different scales (Xu and Wunsch, 2005). Thus, we standardize the data in this thesis to have a mean of zero and a standard deviation of one using the following equation:

$$z_{ip} = \frac{x_{ip} - m_p}{s_p},$$

where m_p is the mean of the feature p and s_p is the standard deviation of the same feature.

3.2 Hierarchical clustering

Hierarchical clustering algorithms can be divided into agglomerative and divisive methods. Agglomerative methods begin with each data object as a separate cluster and iteratively reduce the number of clusters by merging them. In contrast, divisive methods begin with a single cluster containing all data objects and successively split it into smaller clusters.(Härdle and Simar, 2015)

The heart of hierarchical clustering is building a hierarchy of clusters from which a preferred clustering of the data can be obtained. Thus, hierarchical clustering does not require setting an initial number of clusters.

3.2.1 Agglomerative hierarchical clustering

Agglomerative hierarchical clustering algorithms follow, in principle, the following (Mehta et al., 2020):

Algorithm Agglomerative clustering algorithms

1: Begin with each data point as a cluster.
2: Compute the distance matrix of individual points.
3: **while** More than one cluster left **do**
4: Merge two closest clusters together.
5: Update the distance matrix.
6: **end while**

Because the two closest clusters are always merged together, the algorithms can be classified as greedy (Murtagh and Contreras, 2012). The distance between data points is defined by the chosen distance function, but measuring the distance between clusters is the key difference between agglomerative techniques. To be able to define this

inter-cluster closeness, one has to select a linkage criterion. Common linkage criteria include single linkage, complete linkage, and Ward's method. Single linkage gives the distance between clusters as the shortest distance between a pair of data points from those clusters, while complete linkage defines the distance between clusters as the furthest distances between their data points (Mehta et al., 2020). Ward's method, also known as the minimum variance method, selects the merge that results in the smallest increase in within-cluster variance, which is used as a measure of heterogeneity. The within-cluster variance to be minimized when merging clusters in Ward's method is given by (Kaufman and Rousseeuw, 1990):

$$D(A, B) = \frac{|A||B|}{|A| + |B|} \|c_A - c_B\|^2,$$

where A and B are different clusters, $|\cdot|$ gives the number of data points in a cluster, and c_A and c_B are their centroids, respectively. The centroid of a cluster is its center, typically defined as the mean of the data points within the cluster.

The choice of the linkage criterion has a significant impact on clustering results, as it determines the flow of the cluster merging process. Single linkage is prone to merging vastly different clusters only because the clusters share a pair of data points with a small distance. This is called the chaining effect (Jarman, 2020). However, this might be convenient when the clusters are long and chain-like. Complete linkage, on the other hand, seeks to form well-separated clusters, as the data points in a cluster tend to be closer when the linkage is done through the furthest distance between a pair of data points (Härdle and Simar, 2015). Ward's method also produces well-separated clusters because it seeks to minimize within-cluster variance. However, Ward's method has been shown to outperform complete linkage in some practical applications (Sharma et al., 2019). Also, by taking into account the cluster size, Ward's method uses more information in clustering than single and complete linkage, which use only information about the distance between pairs of data points.

3.2.2 Number of clusters

Because the number of clusters is not predefined for hierarchical clustering, the final clustering results must be selected from the hierarchy according to some rule. Agglomerative hierarchical clustering results can be presented as a dendrogram, a tree representation of the merging process. The leaves of the dendrogram represent the starting point of the clustering, where each data point is in its own cluster. The root of the tree is the single cluster containing all data points. As the height of the dendrogram often represents the closeness of merged clusters defined by the selected linkage criterion, the dendrogram can be used as a tool for selecting the clustering result by cutting the tree at a certain height. The resulting clusters are then presented by the branches below the cut. The cut can be performed at a predefined height or after visual analysis of the dendrogram. (Xu and Wunsch, 2005)

For cluster validity, there is also a variety of indices that can be used to validate clustering results and decide on the number of clusters. The R package `NbClust`, introduced by Charrad et al. (2014), provides thirty cluster validation indices, listed in

Table 2. `NbClust` can be evaluated on a data set with a desired linkage and distance function and for a desired range of the number of clusters. Thus, the final number of clusters can be decided based on the majority rule, i.e., selecting the number of clusters that is favored by the majority of indices. The indices in the package examine and validate the clustering results by comparing different properties of within- and between-cluster measures.

Table 2: Indices in `NbClust` package.

	Name of the index	Name in `NbClust`
1	Calinski and Harabasz (CH) index	`"ch"`
2	Duda index	`"duda"`
3	Pseudo t^2 index	`"pseudot2"`
4	Cindex	`"cindex"`
5	Gamma index	`"gamma"`
6	Beale index	`"beale"`
7	Cubic Clustering Criterion	`"ccc"`
8	Point biserial correlation index	`"ptbiserial"`
9	Gplus index	`"gplus"`
10	Davies and Bouldin index	`"db"`
11	Frey index	`"frey"`
12	Hartigan index	`"hartigan"`
13	Tau index	`"tau"`
14	Ratkowsky index	`"ratkowsky"`
15	Scott index	`"scott"`
16	Marriot index	`"marriot"`
17	Ball index	`"ball"`
18	Trcovw index	`"trcovw"`
19	Tracew index	`"tracew"`
20	Friedman index	`"friedman"`
21	McClain and Rao index	`"mcclain"`
22	Rubin index	`"rubin"`
23	Krzanowski and Lai index	`"kl"`
24	Silhouette index	`"silhouette"`
25	Gap statistic	`"gap"`
26	Dindex	`"dindex"`
27	Dunn index	`"dunn"`
28	Hubert statistic	`"hubert"`
29	SD validity index	`"sdindex"`
30	SDbw validity index	`"sdbw"`

4 Mortality data

4.1 Data sources and processing

The mortality data used in this thesis were obtained from the WHO Mortality Database (World Health Organization, 2025a). The raw CSV files in the database report the number of deaths by country, sex, year, age group, and ICD code. These files were merged, and deaths corresponding to individual ICD codes were aggregated under their respective top-level CODs. This process produced total death counts for each selected top-level COD analyzed in this thesis. For some country–sex–year–COD combinations, a small number of deaths were reported with an unknown age group. For each such combination, we calculated the proportion of deaths across the known age groups within that same country–sex–year–COD combination, and redistributed the unknown-age deaths to the age groups according to these proportions.

To perform inter-country analysis, we need death rates that are comparable between countries. Age-standardization is a method used to provide comparable rates across populations with different age structures. In this thesis, we use age-standardized death rates (ASDR) per 100 000 people. The standard population used to calculate ASDR is the WHO world standard population described in Ahmad et al. (2001) and shown in Table (4). The population estimates for each country and age group are from the United Nations and obtained from United Nations, Department of Economic and Social Affairs, Population Division (2024). The ASDR for a cause i is calculated using the direct method of standardization by:

$$\text{ASDR}_i = \sum_a \left(\frac{f_a d_{i_a}}{p_a} \times 10^5\right),$$

where f_a is the WHO standard population percent for age group a from Table 4, d_a is the deaths inside age group a from WHO mortality database and p_a is the UN population estimate in age group a. The fraction is multiplied by 10^5 to obtain the age-standardized rate per 100 000 persons.

The countries selected for this thesis are the ones for which we have ICD-10-coded mortality data in the database for years between 2001 and 2021, and which are classified as having high-quality data in World Health Organization (2024). Data for a country is classified as high quality when it has an average usability of 80% or higher from 2010 onward. WHO defines usability as "the percentage of all deaths which are registered with a meaningful cause of death information". The usability is thus calculated as a percentage of deaths registered with medical certification multiplied by the percentage of meaningful cause-of-death codes among registered deaths. Using only high-quality data makes comparisons and clustering among countries more meaningful and reliable. Selected countries were also required to have a population of at least one million. The selected 26 countries, along with their three-character codes, are shown in Table 3.

The following chapters present the data for selected countries and selected causes of death discussed in Section 2.3. The years 2001, 2018, and 2021 are shown in more detail, as the clustering presented in Section 5 is performed on the ASDRs for those years.

Table 3: Countries selected for the analysis with their ISO 3166-1 alpha-3 codes.

Code	Country
BEL	Belgium
BRA	Brazil
CAN	Canada
CHL	Chile
COL	Colombia
CRI	Costa Rica
HRV	Croatia
CUB	Cuba
CZE	Czech Republic
DNK	Denmark
EST	Estonia
FIN	Finland
HUN	Hungary

Code	Country
ISR	Israel
JPN	Japan
LVA	Latvia
LTU	Lithuania
MEX	Mexico
NLD	Netherlands
NIC	Nicaragua
KOR	Republic of Korea
ESP	Spain
SWE	Sweden
CHE	Switzerland
GBR	United Kingdom
USA	United States of America

Table 4: WHO World Standard Population Distribution (percent)

Age group	WHO World Standard (%)
0–4	8.86
5–9	8.69
10–14	8.60
15–19	8.47
20–24	8.22
25–29	7.93
30–34	7.61
35–39	7.15
40–44	6.59
45–49	6.04
50–54	5.37
55–59	4.55
60–64	3.72
65–69	2.96
70–74	2.21
75–79	1.52
80–84	0.91
85+	0.63

4.2 Neoplasms

Deaths due to neoplasms include all deaths classified under ICD-10 codes C00-D48.9. Figure 2 illustrates the evolution of ASDRs by country and sex for neoplasms, with three outlying countries highlighted after visual observation of the time series. Figure 3 presents snapshots for the years 2001, 2018, and 2021, while Table 5 summarizes the corresponding descriptive statistics.

Across all three years, the mean ASDR is higher for males than for females. A consistent decline in rates is evident both visually and numerically from the mean values. Among males, Nicaragua and Mexico exhibit substantially lower ASDRs throughout the observation period, whereas Hungary maintains notably high rates for both sexes. In 2021, however, Latvia surpassed Hungary to record the highest male ASDR. Among females, Nicaragua and Japan rank consistently among the lowest, while the Republic of Korea shows a distinct decline over time, reaching the lowest female rate by 2013.

In addition to the overall decline in ASDRs, the standard deviation, which reflects variability across countries, also decreased from 2001 to 2021. This indicates a reduction in cross-country disparities, as reflected by the narrowing gap between the maximum and minimum ASDR values. The mean ASDR decreased by 25% for males and 19% for females from 2001 to 2021.

Table 5: Neoplasms – summary statistics

(a) Male

Statistic	2001	2018	2021
Mean	178.71	141.90	133.60
Median	172.97	126.18	117.90
SD	48.42	41.27	36.77
Min	79.75	74.31	82.96
Max	294.85	227.43	205.19

(b) Female

Statistic	2001	2018	2021
Mean	109.11	92.35	88.23
Median	111.51	91.10	86.56
SD	20.71	16.41	15.95
Min	67.23	59.18	55.66
Max	155.67	131.78	125.54

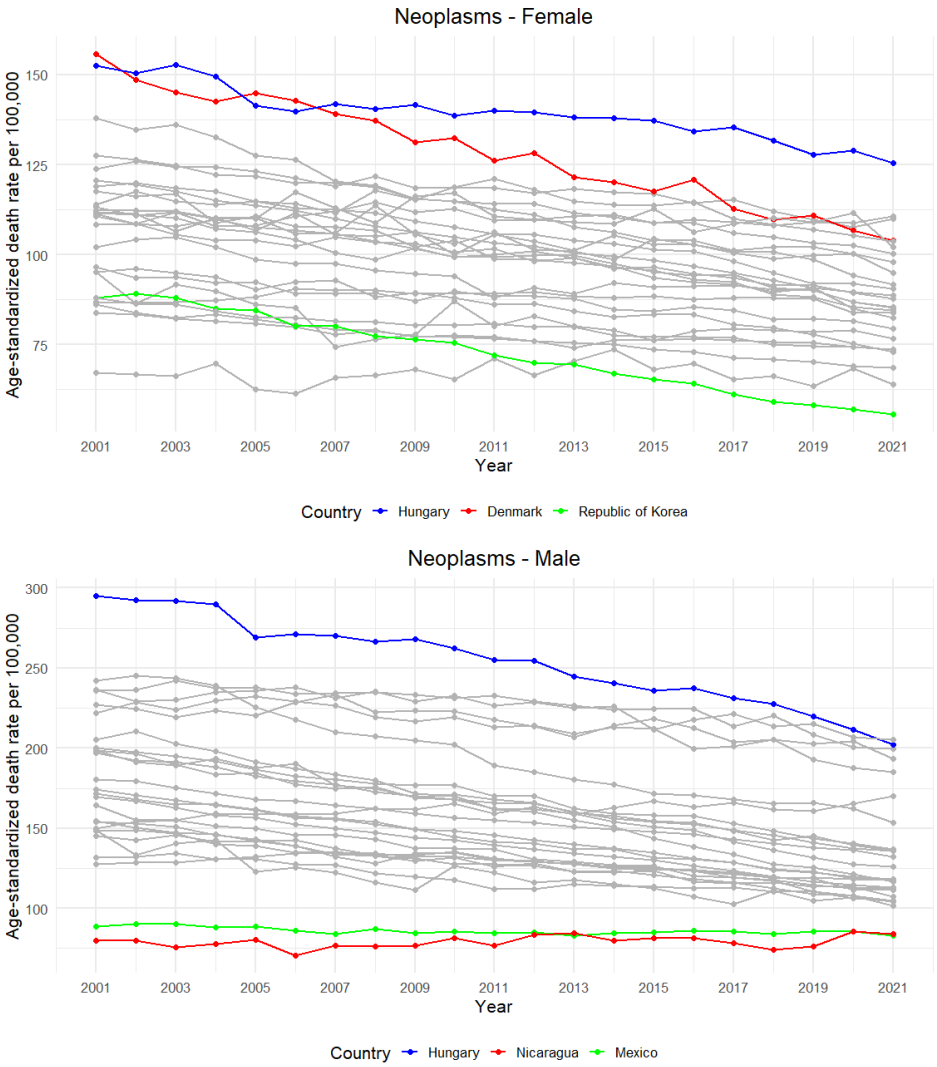


Figure 2: Time series of death rates due to neoplasms.

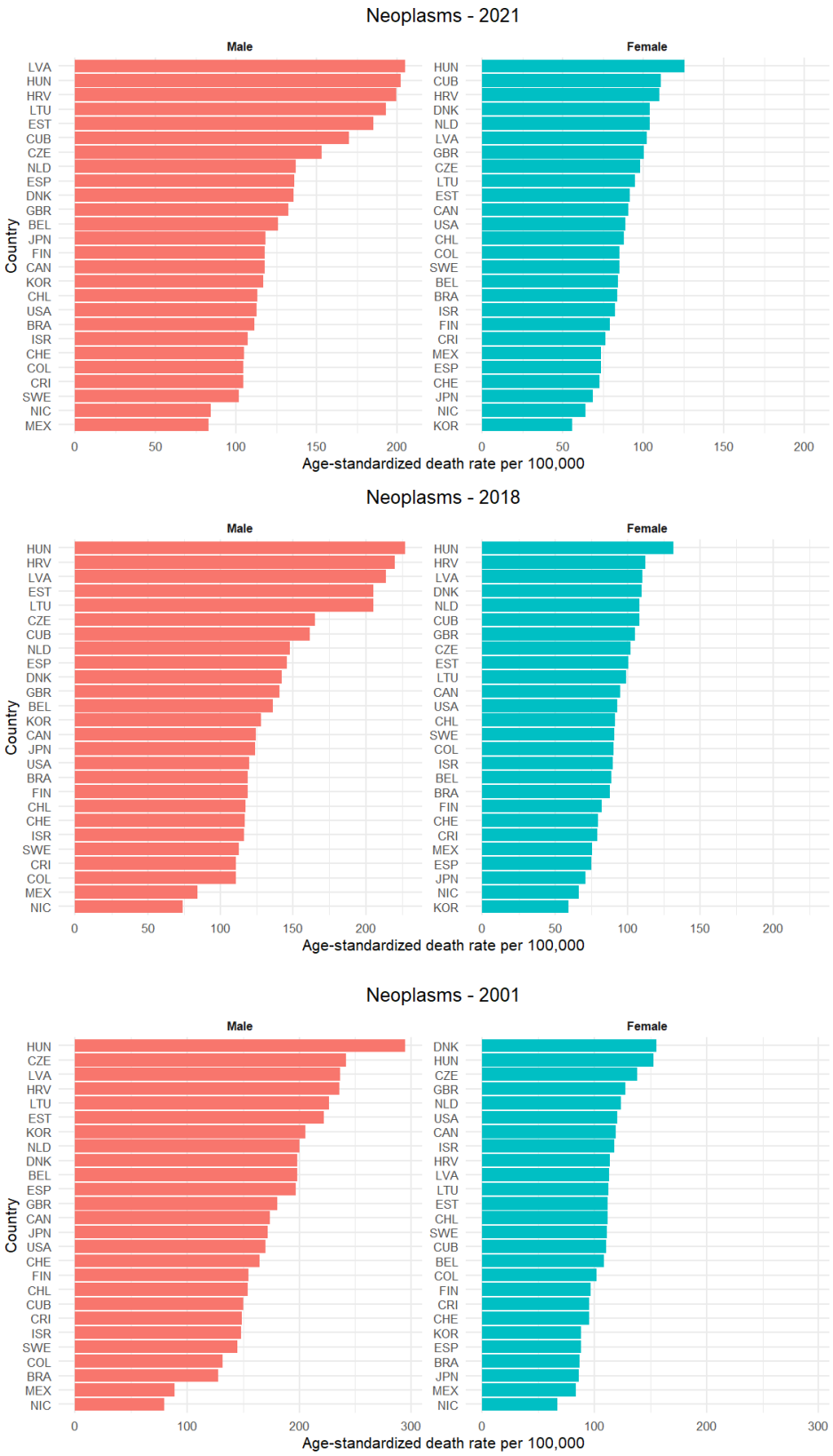


Figure 3: Yearly bar charts of death rates due to neoplasms.

4.3 Diseases of the respiratory system

Deaths due to diseases of the respiratory system include those classified under ICD-10 codes J00-J98.9 as well as codes within the U04, U07, U09, and U10 blocks, including SARS and COVID-19–related deaths. Figure 4 illustrates the evolution of ASDRs by country and sex for respiratory diseases, with three of the most extreme-behavior countries highlighted based on visual inspection of the time series. Figure 5 presents snapshots for the years 2001, 2018, and 2021, while Table 6 summarizes the corresponding descriptive statistics.

Colombia, Brazil, and Mexico exhibit the highest ASDRs among both males and females across the study period. A global increase in ASDRs is evident after 2019, primarily reflecting the impact of the COVID-19 pandemic. For both sexes, the mean ASDR declined between 2001 and 2018 but rose substantially again by 2021. The decrease in the mean from 2001 to 2018 is 20%, and the increase from 2018 to 2021 is 179% for males. For females, the decrease in the mean from 2001 to 2018 is 10%, and the increase from 2018 to 2021 is 157%. Also, the increase in the standard deviation from 2018 to 2021 is considerable, suggesting significant differences across countries in the effectiveness of COVID-19 responses.

Overall, the mean ASDR is higher for males than for females. Among males, Sweden consistently ranks among the countries with the lowest rates across all three years, with Finland recording the lowest male ASDR in 2018 and 2021. Among females, Latvia, Lithuania, and Estonia had some of the lowest rates in 2001 and 2018, but not in 2021, suggesting relatively greater mortality impacts from COVID-19 compared to other countries with low rates. For 2021, Finland also exhibits the lowest ASDR among females.

Table 6: Diseases of the respiratory system – summary statistics

(a) Male

Statistic	2001	2018	2021
Mean	59.89	47.67	133.02
Median	59.75	48.30	118.15
SD	17.87	15.46	86.67
Min	33.85	20.94	24.68
Max	104.18	84.71	349.75

(b) Female

Statistic	2001	2018	2021
Mean	31.75	28.70	73.90
Median	28.16	27.05	65.64
SD	15.12	13.82	51.29
Min	11.03	9.82	11.44
Max	67.53	59.41	193.91

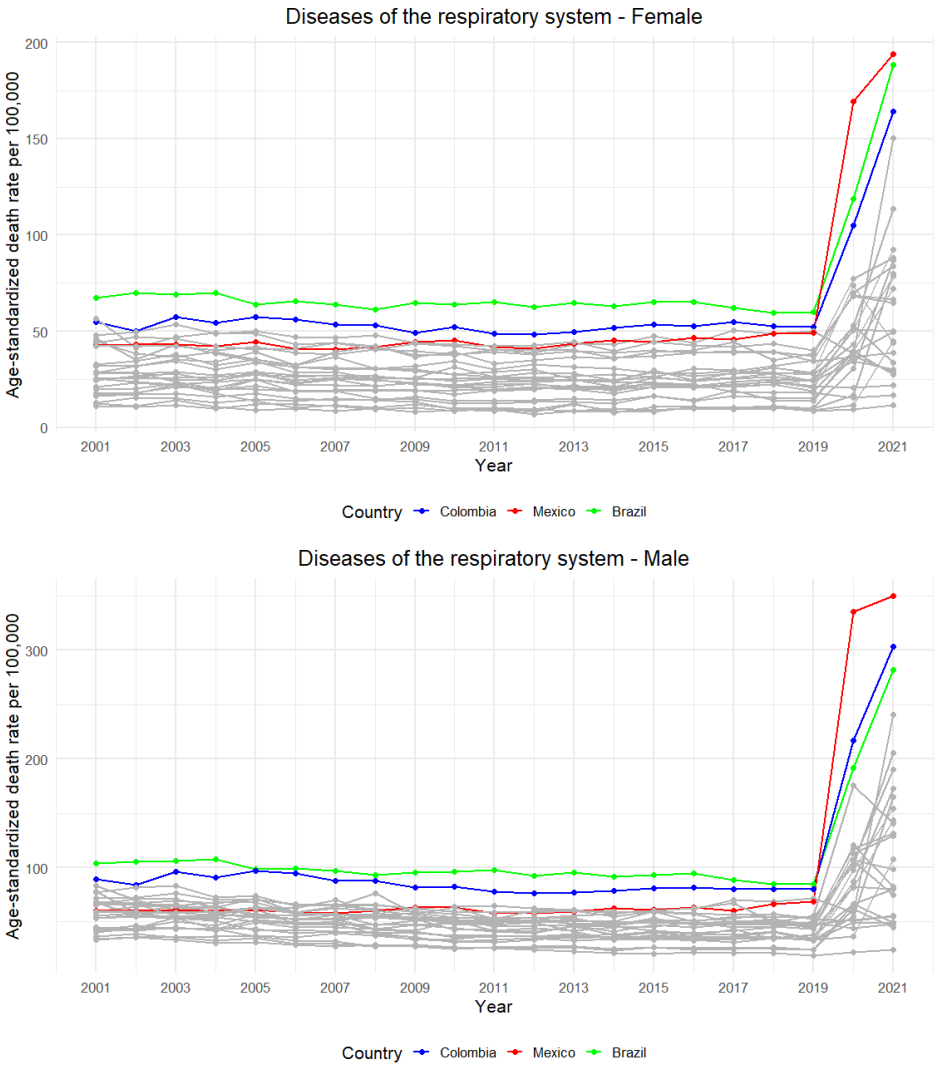


Figure 4: Time series of death rates due to diseases of the respiratory system.

Figure 5: Yearly bar charts of death rates due to diseases of the respiratory system.

4.4 Diseases of the circulatory system

Deaths due to diseases of the circulatory system include ICD-10 codes I00-I99. Figure 6 illustrates the evolution of ASDRs by country and sex for circulatory diseases, with three of the most uniquely behaving countries highlighted after visual observation of the time series. Figure 7 presents the ASDR snapshots for the years 2001, 2018, and 2021, and Table 7 summarizes the corresponding descriptive statistics.

Latvia and Lithuania consistently exhibit high ASDRs for both sexes throughout the observation period. The global trend shows a clear decline in ASDR from 2001 to 2018, with the decrease in the mean ASDR being 31% for males and 39% for females. However, an increase in the mean ASDR is observed from 2018 to 2021. That increase is 9% for males and 7% for females. Nicaragua, Cuba, and Latvia are examples of countries with a notable rise in ASDR between 2018 and 2021. Despite this increase in the mean, the median ASDR continues to decline, suggesting that the observed increase is driven by a limited number of countries rather than a global trend.

Japan, the Republic of Korea, and Israel are among the countries with the lowest ASDRs across all years for both sexes.

Table 7: Diseases of the circulatory – summary statistics

(a) Male

Statistic	2001	2018	2021
Mean	283.19	178.96	194.52
Median	237.81	145.59	137.39
SD	137.13	98.31	123.29
Min	136.75	77.91	68.53
Max	630.63	432.57	463.57

(b) Female

Statistic	2001	2018	2021
Mean	180.50	109.95	117.31
Median	146.43	81.66	78.15
SD	79.26	54.78	69.59
Min	80.31	48.90	41.77
Max	356.34	227.82	266.34

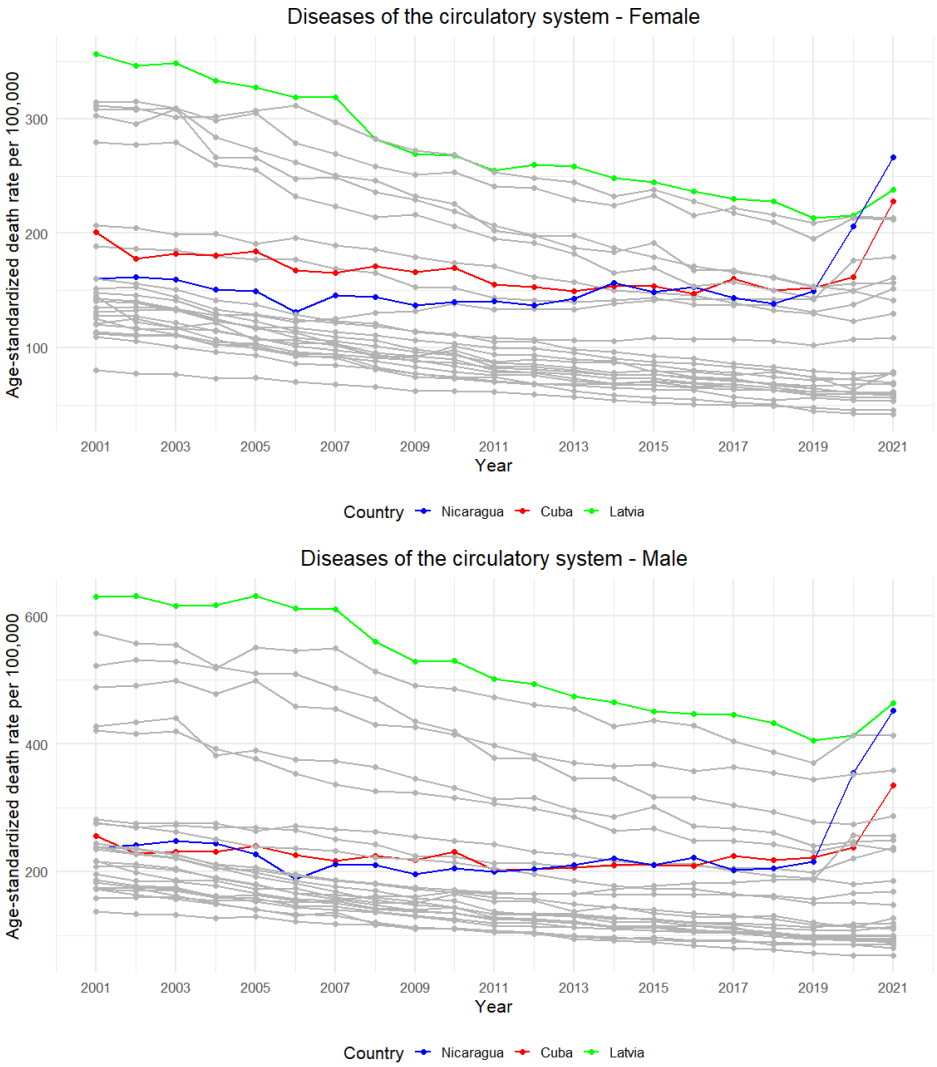


Figure 6: Time series of the death rates due to diseases of the circulatory system.

Figure 7: Yearly bar charts of the death rates due to diseases of the circulatory system.

4.5 Intentional self-harm

Deaths due to intentional self-harm include ICD-10 codes X60-X84.9. Figure 8 illustrates the evolution of ASDRs by country and sex for intentional self-harm, with three of the most outlying countries highlighted after visual observation of the time series. Figure 9 presents the ASDR snapshots for 2001, 2018, and 2021, and Table 8 provides the corresponding summary statistics.

Eastern European countries, including Estonia, Hungary, Latvia, and Lithuania, consistently exhibit some of the highest ASDRs for males over the observation period. Among females, the Republic of Korea shows considerably higher ASDRs than any other country in nearly all years. For males, Lithuania records the highest rates throughout most of the period, except in 2021, when it is surpassed by the Republic of Korea.

For males, both the median and mean ASDR decline steadily from 2001 to 2021, with the decline being 32%. Among females, the mean ASDR decreases by 19% between 2001 and 2018, but increases modestly by 1% from 2018 to 2021.

The countries with the greatest relative decrease in the ASDR over the observation period are Lithuania, for males and females, with decreases of 64% and 54%, respectively. The Republic of Korea has the greatest relative increase in the ASDR, being 27% for males and 59% for females.

Countries among the lowest ASDRs for both sexes include Spain, Mexico, Colombia, and Israel. Overall, the ASDR remains consistently higher for males than for females across all years.

Table 8: Intentional sef-harm – summary statistics

(a) Male

Statistic	2001	2018	2021
Mean	21.93	15.81	14.85
Median	17.50	15.26	13.62
SD	14.85	6.28	4.98
Min	7.18	8.19	6.83
Max	71.32	33.31	26.56

(b) Female

Statistic	2001	2018	2021
Mean	5.64	4.55	4.60
Median	5.28	4.15	3.99
SD	2.96	2.23	2.44
Min	1.34	1.93	2.10
Max	12.08	11.66	13.15

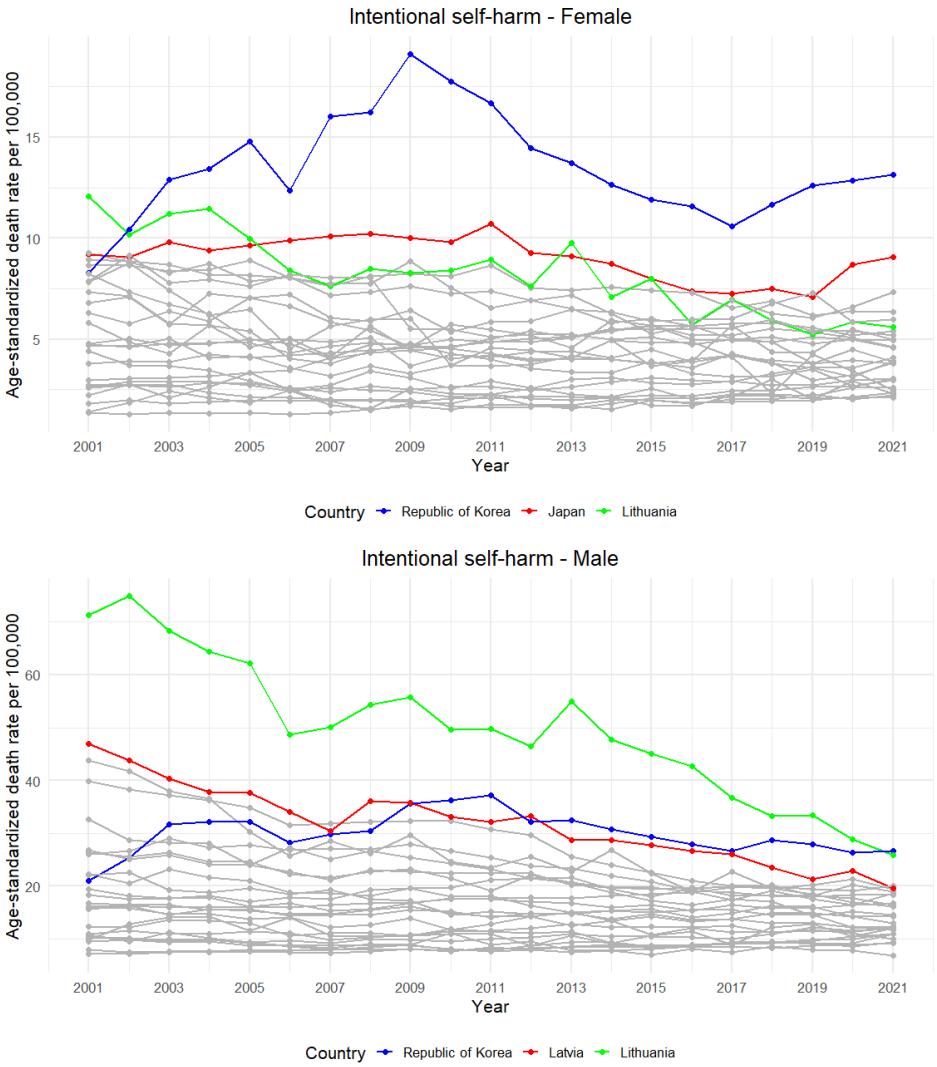


Figure 8: Time series of death rates due to intentional self-harm.

Figure 9: Yearly bar charts of death rates due to intentional self-harm.

4.6 Accidents and Assaults

Deaths due to accidents and assaults include those classified under ICD-10 codes V01-X59.9, Y40-Y86, X85-Y09.9, as well as blocks Y88, Y89, and code Y87.1, excluding blocks X41, X42, X44, and X45. Figure 10 illustrates the evolution of ASDRs by country and sex for these causes, with three uniquely behaving countries highlighted after visual observation of the time series. Figure 11 presents the ASDR snapshots for 2001, 2018, and 2021, while Table 9 reports the corresponding summary statistics.

Before 2009, Lithuania and Latvia exhibited ASDRs for females that were considerably higher than the overall mean. After 2009, Cuba showed the highest female ASDR. The country with the lowest female ASDR varies across the three examined years, although Spain consistently ranks among the three lowest in each year. Between 2001 and 2021, the mean ASDR among females decreased by approximately 38%.

For males, the mean ASDR is substantially higher than for females. Colombia, Mexico, Brazil, Latvia, and Lithuania remain among the countries with the highest ASDRs throughout the observation period, whereas the United Kingdom consistently ranks among the lowest ASDRs in 2001, 2018, and 2021. As in females, the male mean ASDR declines by approximately 43% between 2001 and 2021.

Table 9: Accidents and assaults – summary statistics

(a) Male

Statistic	2001	2018	2021
Mean	65.90	39.10	37.59
Median	50.52	33.16	28.57
SD	46.19	24.25	24.18
Min	19.89	14.48	13.92
Max	196.02	96.47	91.09

(b) Female

Statistic	2001	2018	2021
Mean	19.57	12.52	12.14
Median	18.16	12.00	11.54
SD	8.39	4.41	4.78
Min	8.81	6.97	6.01
Max	41.83	24.86	25.17

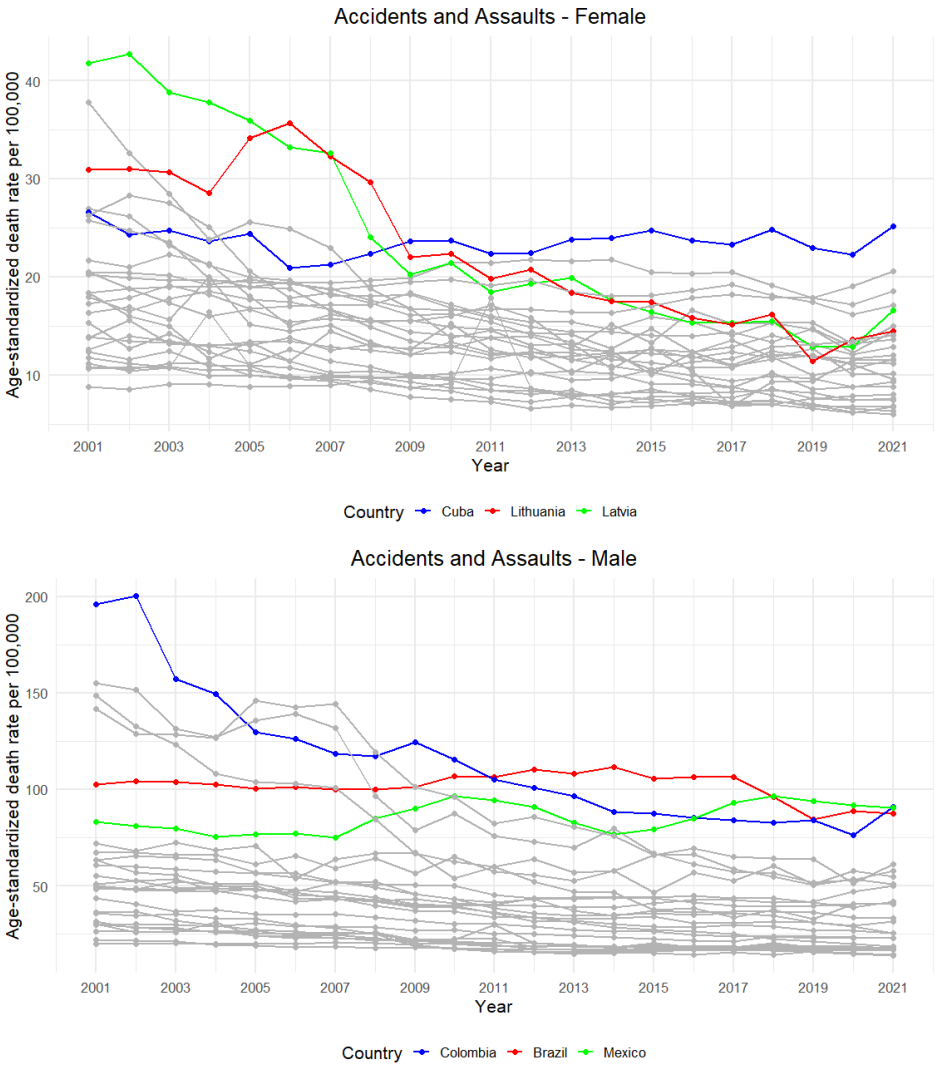


Figure 10: Time series of death rates due to accidents and assaults.

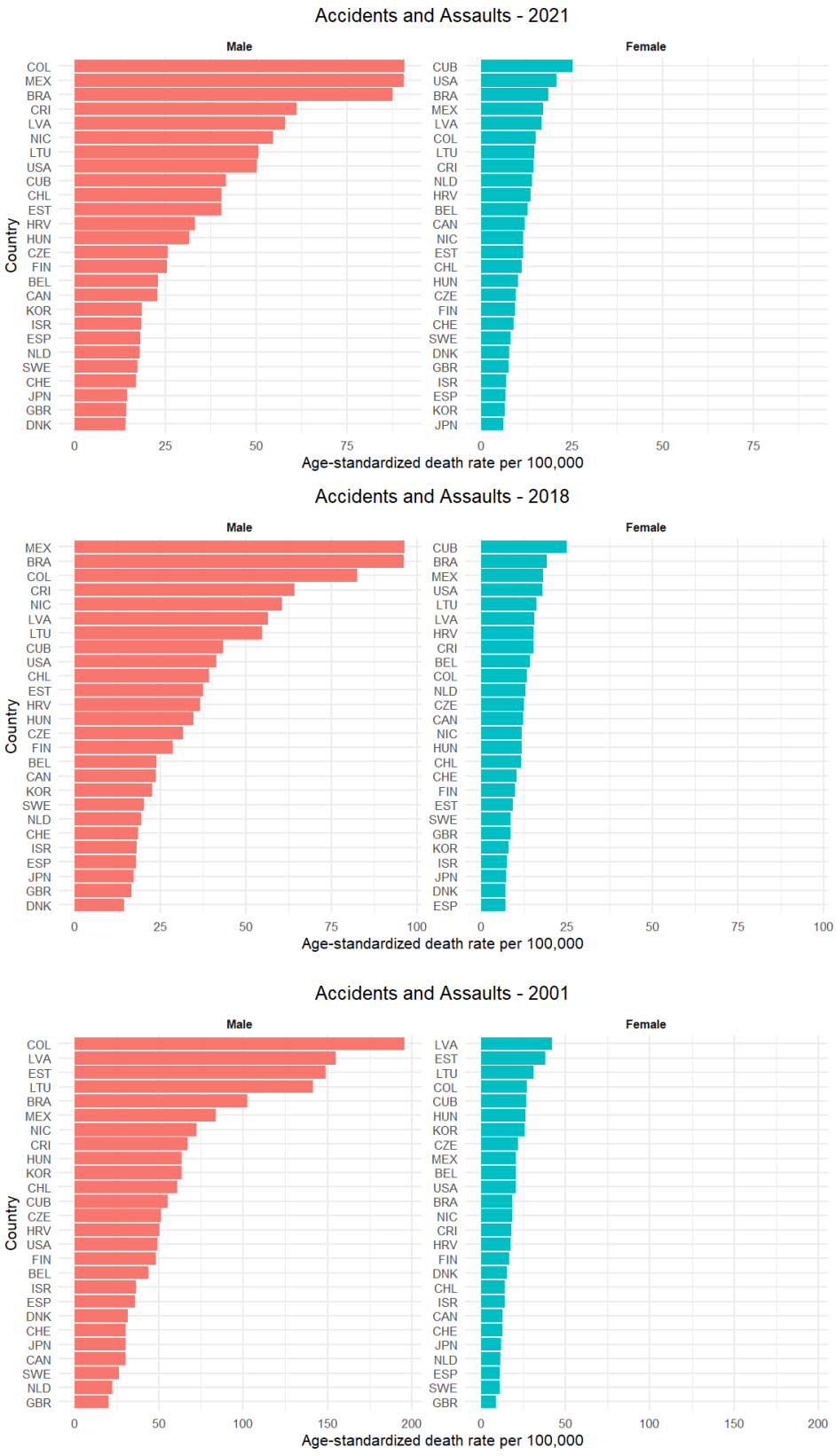


Figure 11: Yearly bar charts of death rates due to accidents and assaults.

5 Clustering of mortality data

We conduct a cluster analysis using the country-level mortality data described in Section 4. The analysis is performed separately for 2001, 2018, and 2021 to examine changes in the cluster structure. For each country and year, the data consist of ten variables corresponding to ASDRs for the five selected causes of death, each for both genders.

As outlined in Section 3, agglomerative hierarchical clustering is applied using Ward's method. The clustering is implemented using the `hclust` function in base R, with Euclidean distance measure. Prior to clustering, all variables are standardized to have a zero mean and a standard deviation of one to ensure that differences in scale do not influence distance calculations.

The selection of the number of clusters is guided by a combination of visual inspection of the dendrograms and formal cluster validation criteria. In particular, we use the `NbClust` package in R (Charrad et al., 2014), which includes 30 validation indices for determining an optimal number of clusters. These indices are listed in Table 2. Based on these diagnostics, we present and compare clustering solutions with two, three, and four clusters for each considered year.

To better interpret the clusters, the clustering results are visualized using heatmaps that display the ASDRs for countries grouped by cluster. These visualizations assist in identifying defining characteristics and differences between clusters. Furthermore, in Section 5.4, we incorporate the Human Development Index (HDI) as an external indicator to assess how levels of socioeconomic development align with the mortality-based clusters. HDI is not included in the clustering process itself but is used solely to interpret and validate the clustering results.

5.1 Clustering for the year 2001 data

The dendrogram for the hierarchical clustering of the 2001 data is presented in Figure 12. A visual inspection of the dendrogram suggests that a reasonable choice for the number of clusters is two, separating Eastern European countries from the rest. This choice is also supported by the results from `NbClust`, evaluated between two and eight clusters. Of the 30 validation indices, 13 favored a two-cluster solution, while the second-most-frequently suggested solution, seven clusters, was supported by only five indices.

The country compositions for two, three, and four clusters are given in the heatmaps shown in Figures 13, 14, and 15, respectively. In the two-cluster solution, countries are divided into an Eastern European cluster and a cluster containing all other countries. As shown in Figure 13, the primary distinguishing characteristic between these clusters is mortality from diseases of the circulatory system, which exhibits substantially higher ASDRs in the Eastern European cluster compared to the other cluster. In addition, the Eastern European countries also display higher average ASDRs for accidents and assaults, intentional self-harm, and neoplasms compared to the countries in the other cluster. In contrast, mortality from diseases of the respiratory system is, on average, lower in this cluster than in the cluster with the remaining countries.

Introducing a third cluster, as shown in Figure 14, leads to the formation of a distinct Latin American cluster consisting of Mexico, Nicaragua, Brazil, and Colombia. The defining characteristics of this cluster are comparatively low ASDRs for neoplasms and intentional self-harm. The remaining Latin American countries, Cuba, Chile, and Costa Rica, remain grouped with the larger cluster containing the rest of the countries.

When a fourth cluster is introduced, the original Eastern European cluster further divides into two subclusters: Lithuania, Estonia, and Latvia (East1), and Hungary, Croatia, and the Czech Republic (East2), as illustrated in Figure 15. This split is primarily driven by higher ASDRs in the East1 group due to accidents and assaults relative to East2. Furthermore, intentional self-harm ASDRs are also consistently higher in East1, emphasizing the distinction between these two Eastern European subclusters.

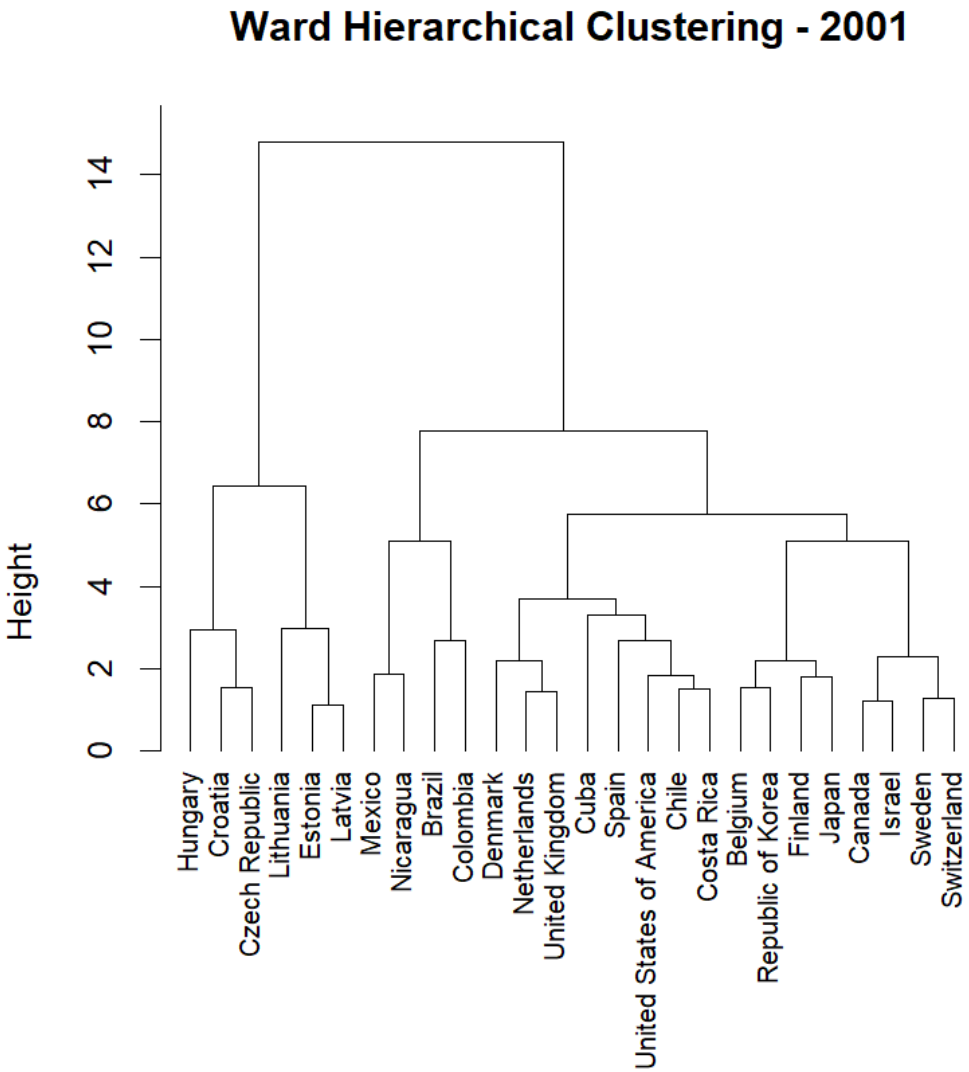


Figure 12: Dendrogram of the clustering for the year 2001 data.

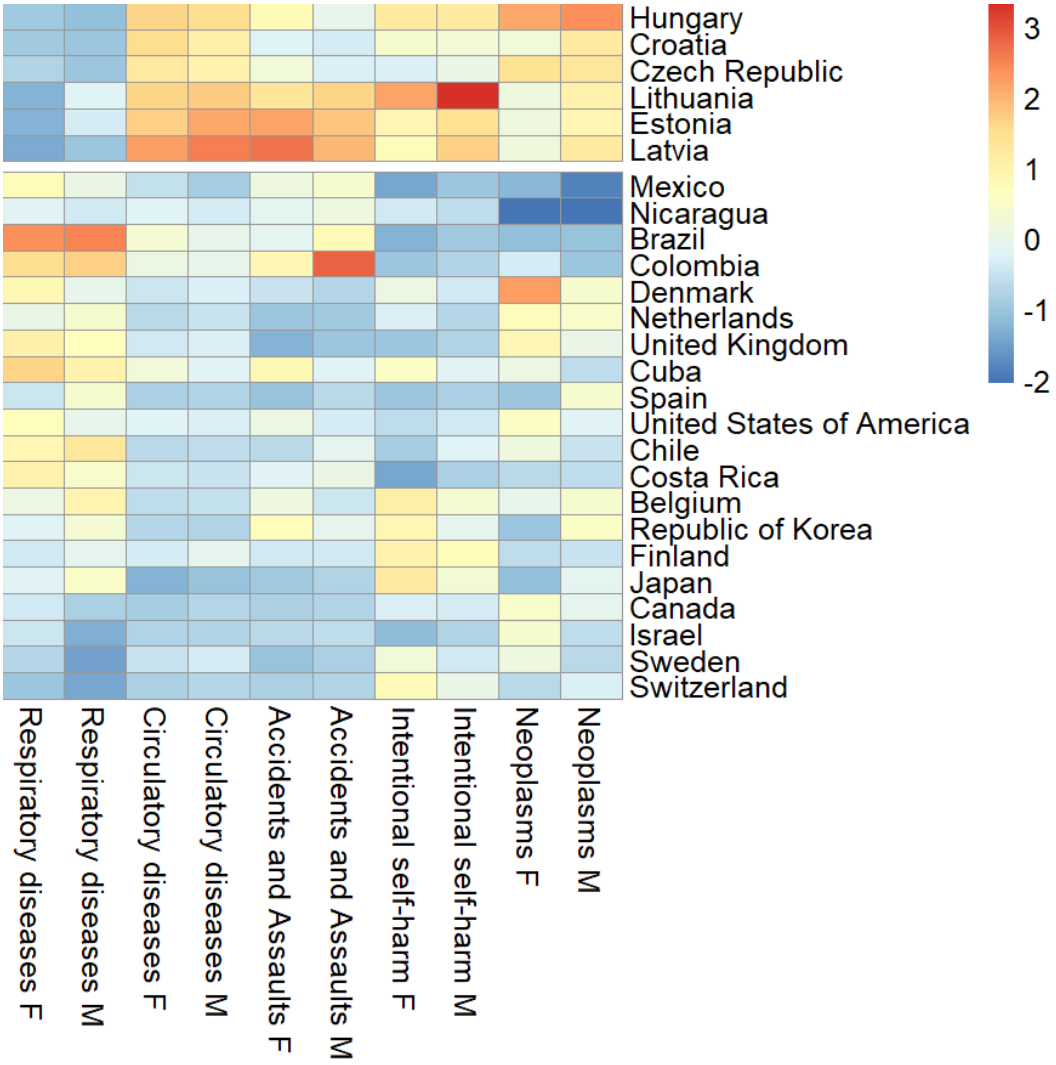


Figure 13: Heatmap of the standardized ASDRs of the 2001 data. Two clusters are given as separate blocks.

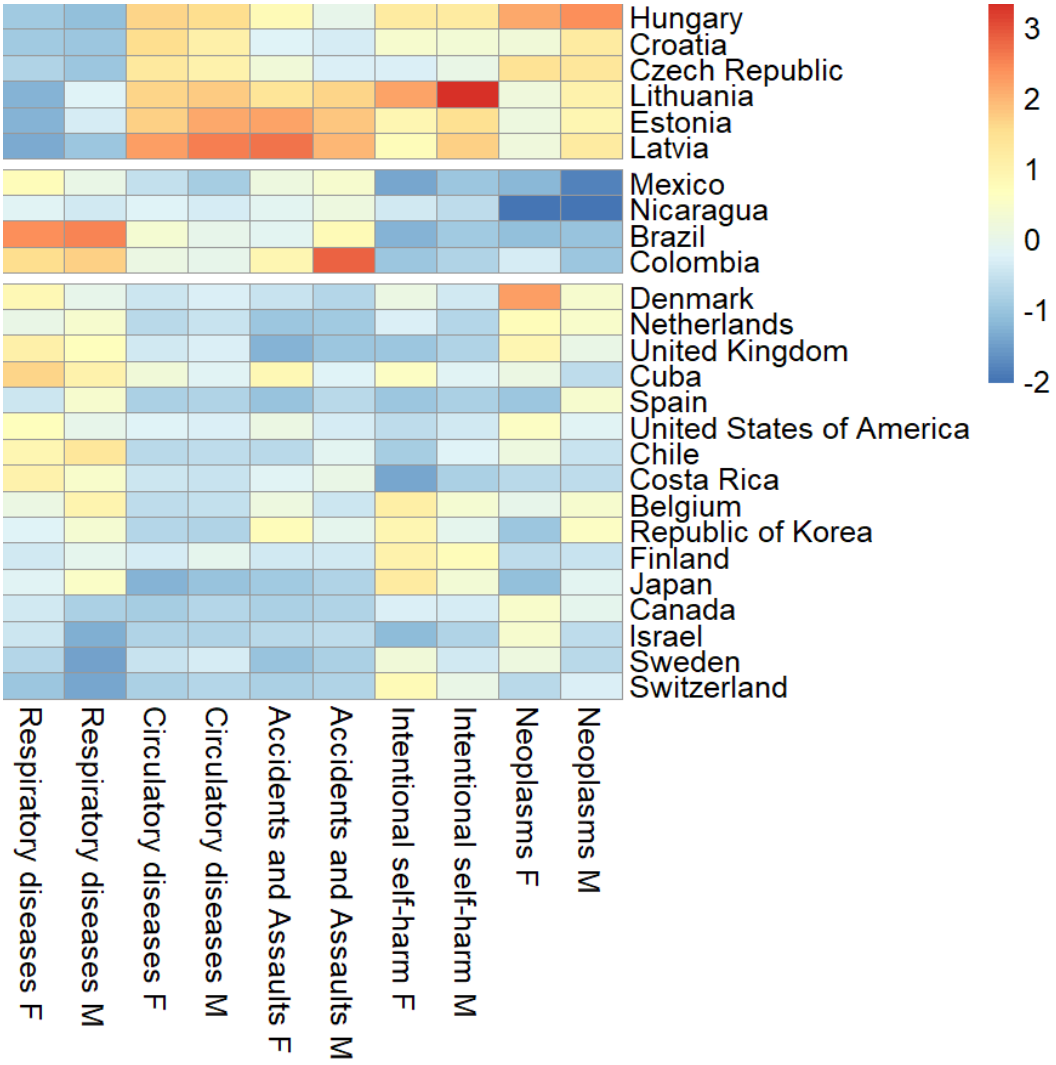


Figure 14: Heatmap of the standardized ASDRs of the 2001 data. Three clusters are given as separate blocks.

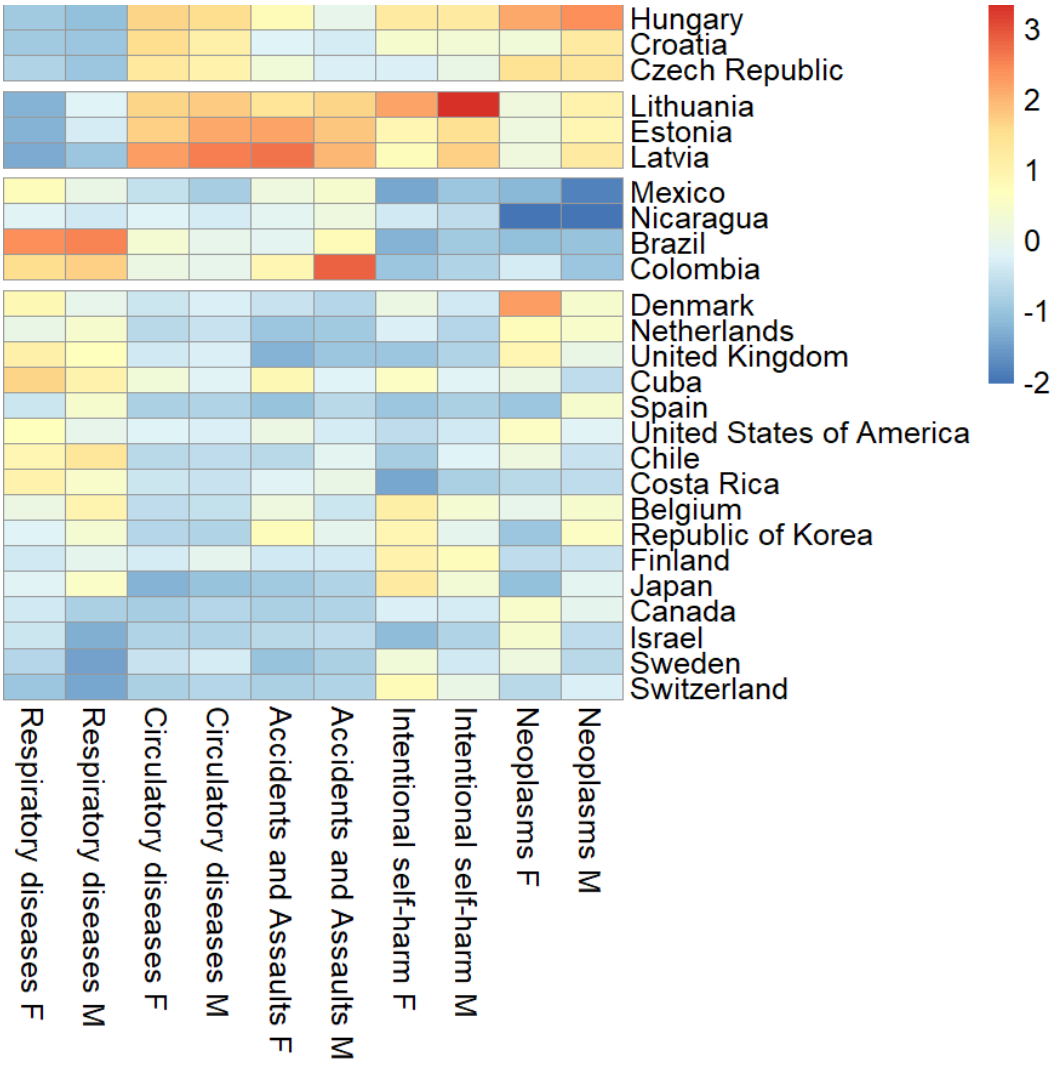


Figure 15: Heatmap of the standardized ASDRs of the 2001 data. Four clusters are given as separate blocks.

5.2 Clustering for the year 2018 data

The dendrogram for the hierarchical clustering of the 2018 data is presented in Figure 16. Based on the dendrogram, the most plausible number of clusters for the 2018 data is three or four. The substantial increase in merge height when moving from three to two clusters indicates that a two-cluster solution combines countries that are rather dissimilar into the same cluster. This interpretation is supported by the results from `NbClust`, where a clear majority of indices (18 out of 30) favor a three-cluster solution when evaluated over a range of two to eight clusters.

The cluster structures for two, three, and four clusters are presented in Figures 17, 18, and 19, respectively. As in 2001, the two-cluster solution, shown in Figure 17, separates Eastern European countries from the remaining countries. Elevated ASDRs due to circulatory diseases and neoplasms continue to characterize the Eastern European cluster, although relative declines are observed for accidents and assaults and intentional self-harm compared to 2001.

In the three-cluster solution shown in Figure 18, several Latin American countries form a distinct cluster, which separates them from the cluster containing a broader group of countries. This separation is consistent with the notable decrease in the merge height observed when increasing the number of clusters from two to three (see Figure 16). The Latin American cluster consists of Brazil, Colombia, Mexico, and Cuba, while Chile, Costa Rica, and Nicaragua remain within the larger cluster. In other words, the Latin American cluster in 2018 consists of the same countries as in 2001, except that Cuba replaces Nicaragua.

Introducing a fourth cluster further partitions the largest cluster with a broader group of countries, as illustrated in Figure 19. Of the two newly formed clusters, the one containing the Republic of Korea is characterized by much higher ASDRs due to intentional self-harm compared to the other newly formed cluster. This discrepancy is particularly evident among females. Notably, the Republic of Korea exhibits exceptionally high ASDRs due to female intentional self-harm. In contrast, the other newly formed cluster containing Nicaragua displays relatively low ASDRs due to intentional self-harm, highlighting a clear distinction between these clusters.

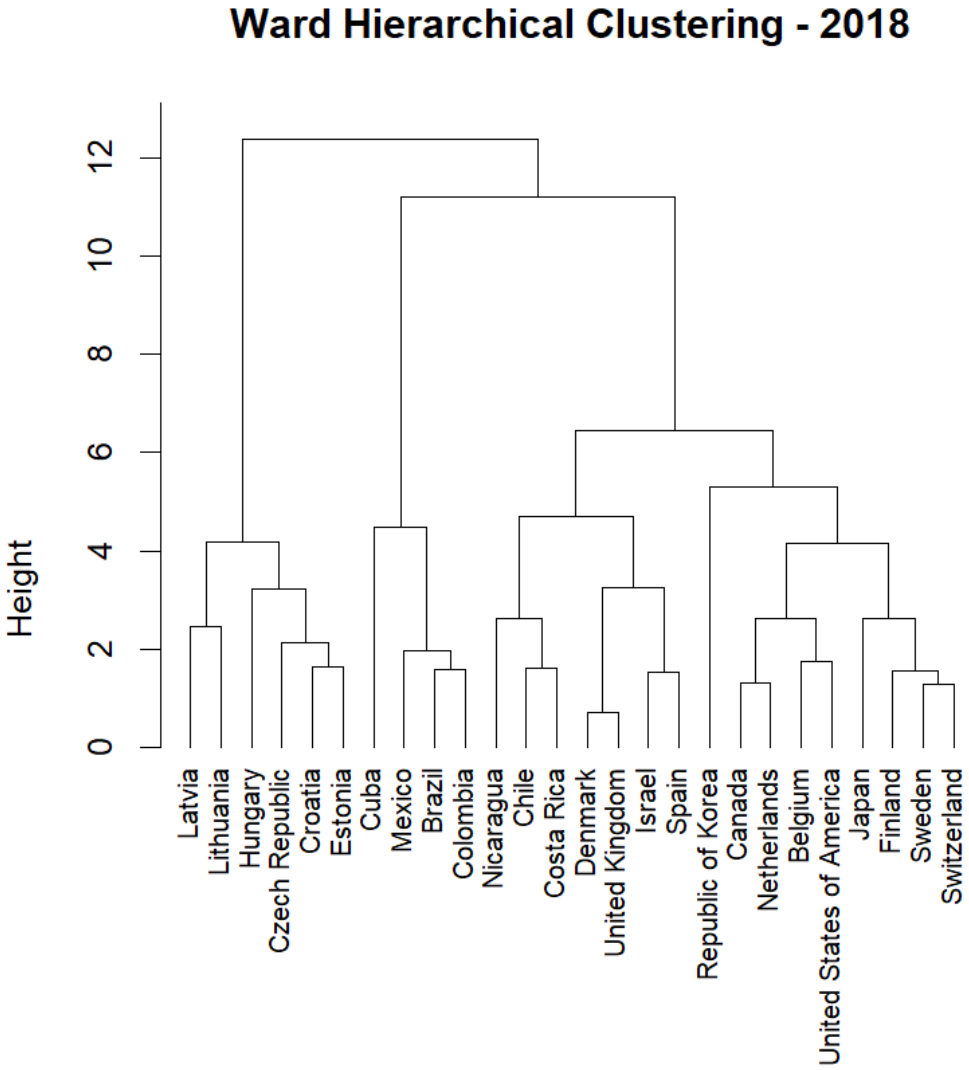


Figure 16: Dendrogram of the clustering for the year 2018 data.

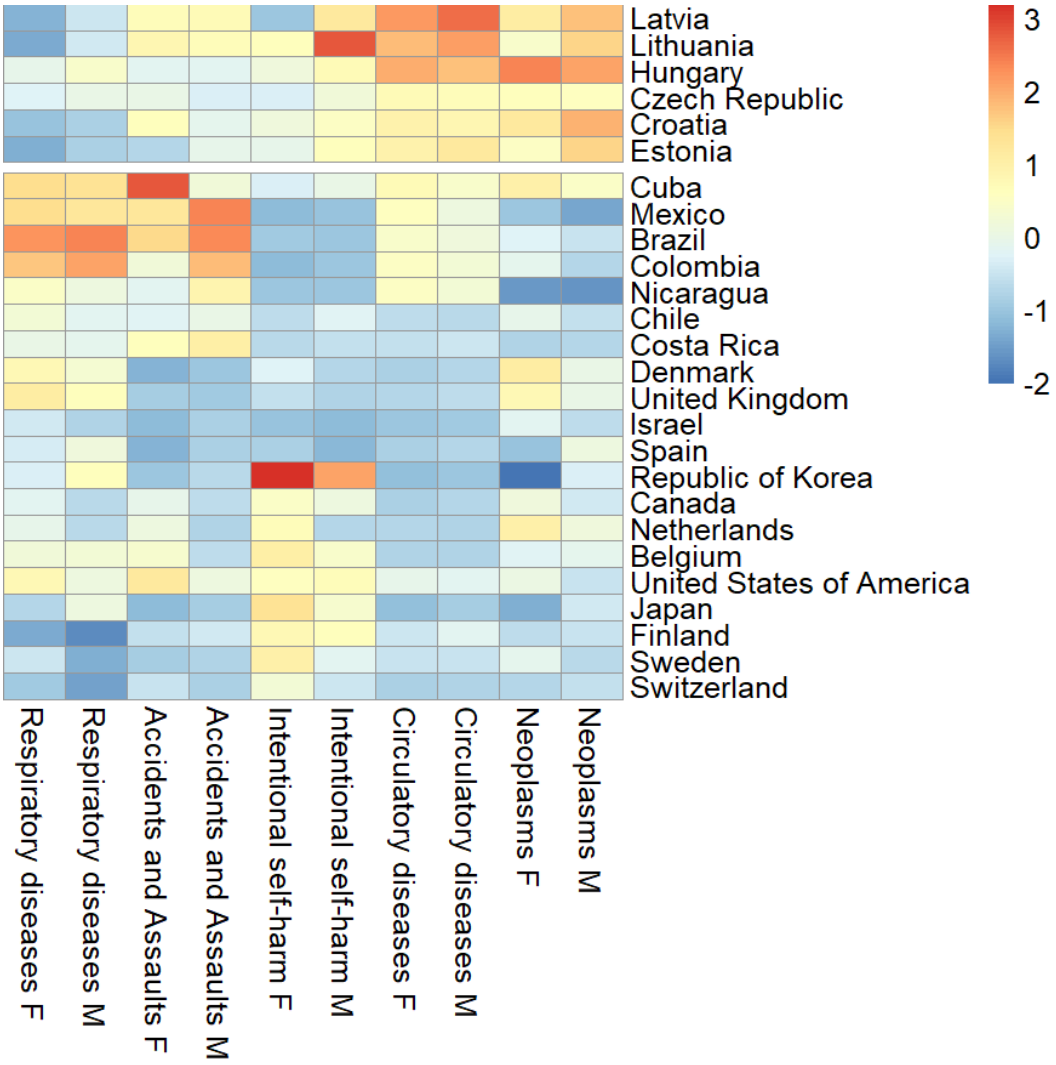


Figure 17: Heatmap of the standardized ASDRs of the 2018 data. Two clusters are given as separate blocks.

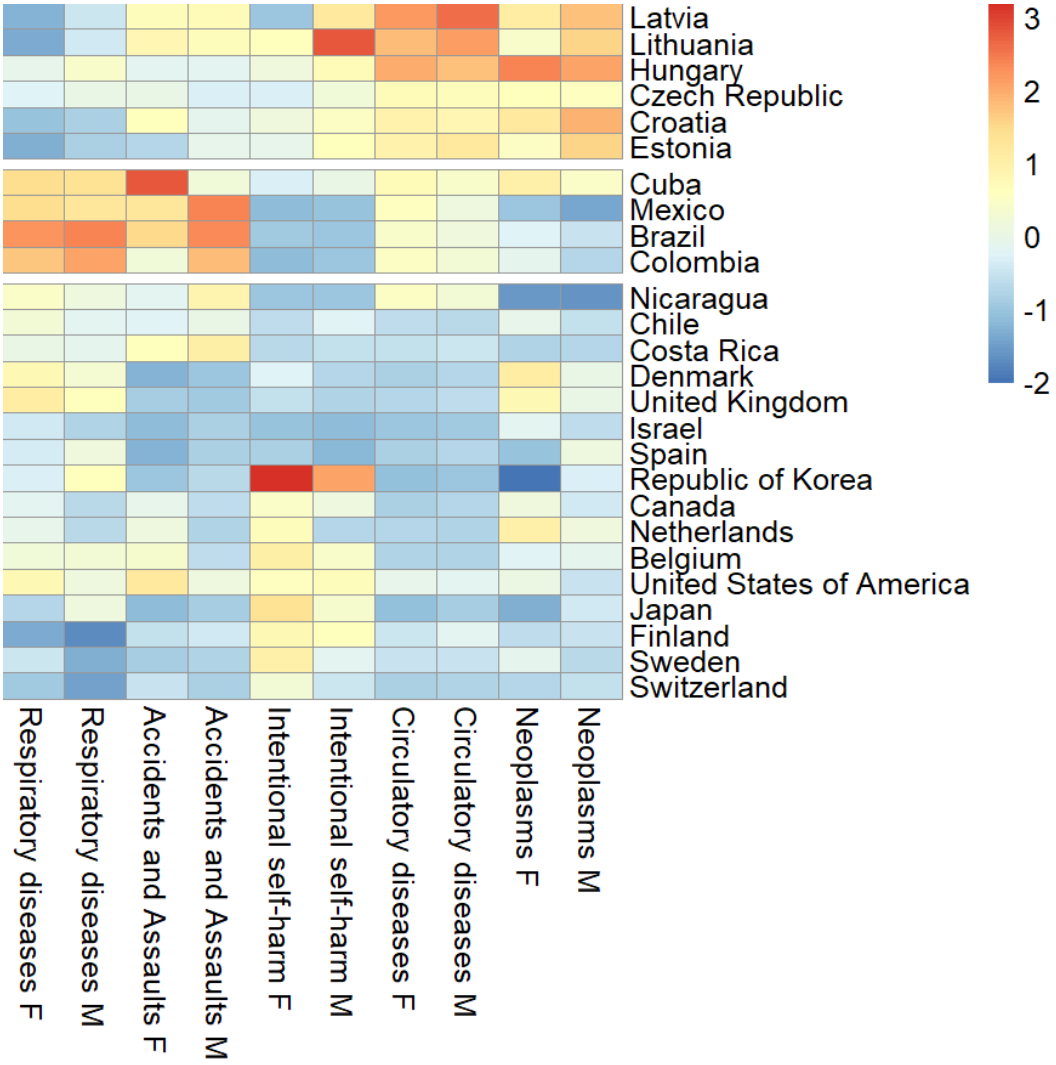


Figure 18: Heatmap of the standardized ASDRs of the 2018 data. Three clusters are given as separate blocks.

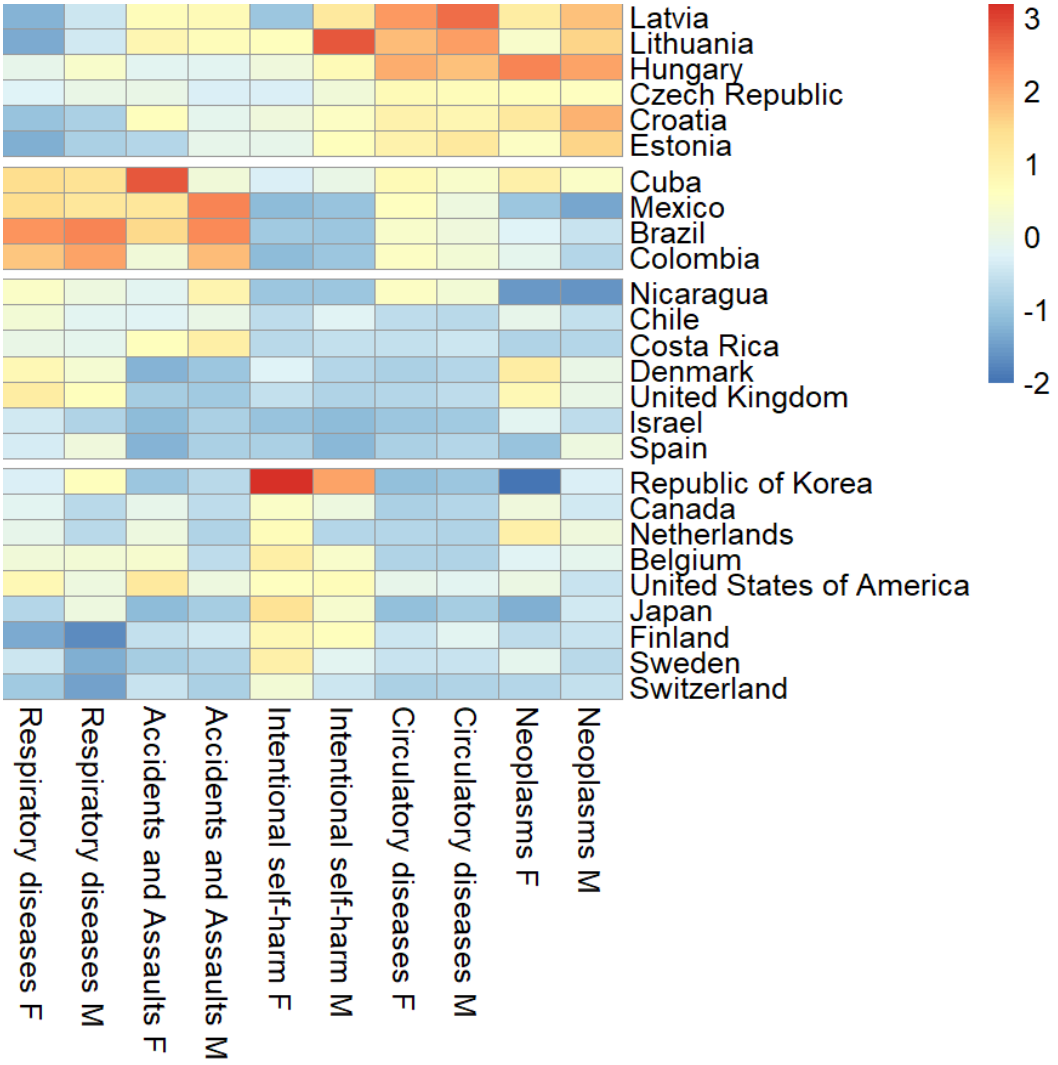


Figure 19: Heatmap of the standardized ASDRs of the 2018 data. Four clusters are given as separate blocks.

5.3 Clustering for the year 2021 data

The dendrogram for the 2021 data clustering is presented in Figure 20. Visual inspection of the dendrogram indicates that the most reasonable range for the number of clusters is between two and five. Consistent with this visual assessment, `NbClust`, evaluated for two to eight clusters, proposes three clusters as the most supported solution, with 12 indices favoring this choice.

With two clusters, shown in Figure 21, the countries are divided such that Eastern European countries, Latin American countries, and the United States form one cluster (hereafter LatEastUSA), while the remaining countries form the second cluster (REST). This clustering structure differs from those observed in 2001 and 2018, where the two-cluster solution separated Eastern Europe from all other countries. In 2021, countries in the LatEastUSA cluster generally had higher ASDRs due to circulatory and/or respiratory diseases than those in the REST cluster.

Introducing a third cluster divides the LatEastUSA cluster into two distinct clusters, as illustrated in Figure 22. The other of these new clusters consists primarily of Eastern European countries together with Cuba (EastCu), while the other includes the remaining Latin American countries and the United States (LatUSA). The main distinguishing feature between these two clusters is the lower ASDRs observed in the LatUSA cluster, due to neoplasms, compared with the EastCu cluster. Additionally, ASDRs due to intentional self-harm are on average lower in the LatUSA cluster than in the EastCu cluster, further contributing to their separation.

When a fourth cluster is introduced, the Republic of Korea separates from the REST cluster, as shown in Figure 23. This separation is driven by Korea's distinct mortality profile, characterized by substantially higher ASDRs due to intentional self-harm and lower ASDRs due to female neoplasms compared to the other countries in the REST cluster.

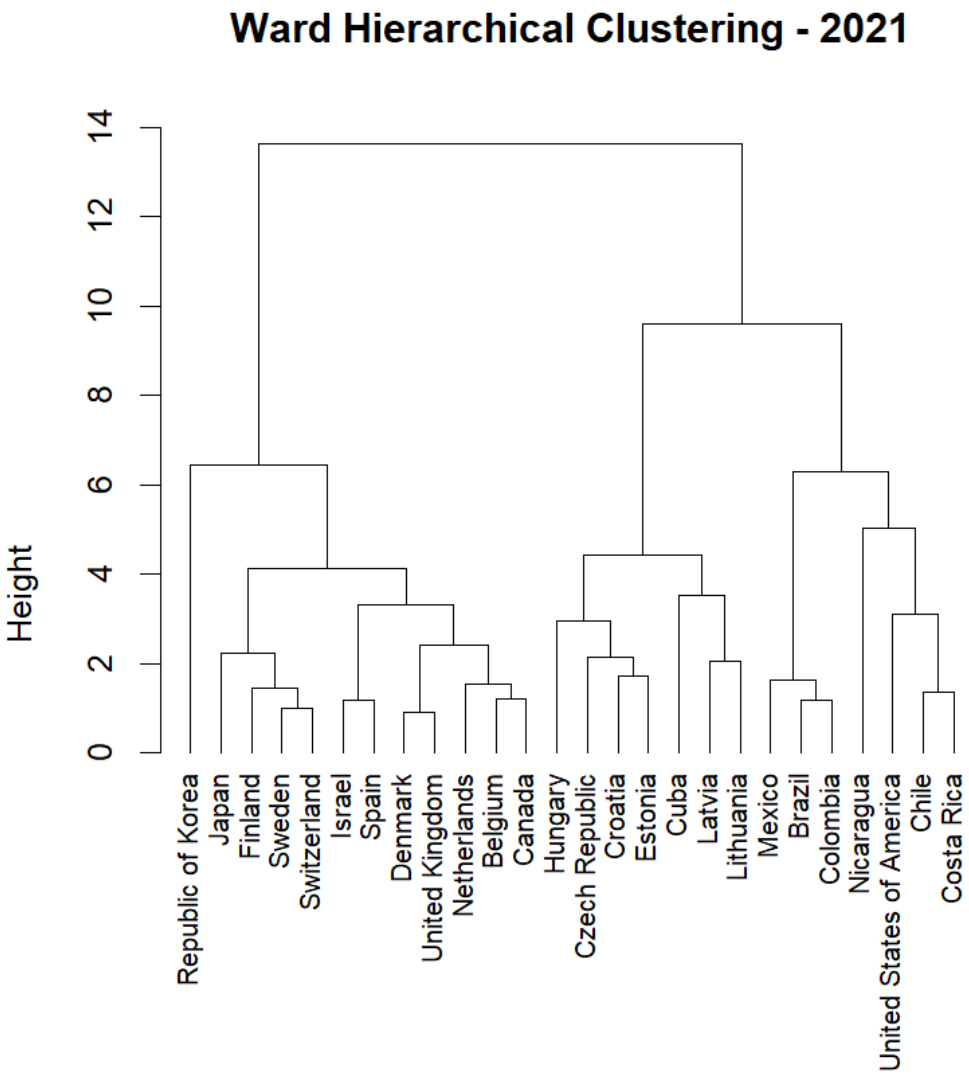


Figure 20: Dendrogram of the clustering for the year 2021 data.

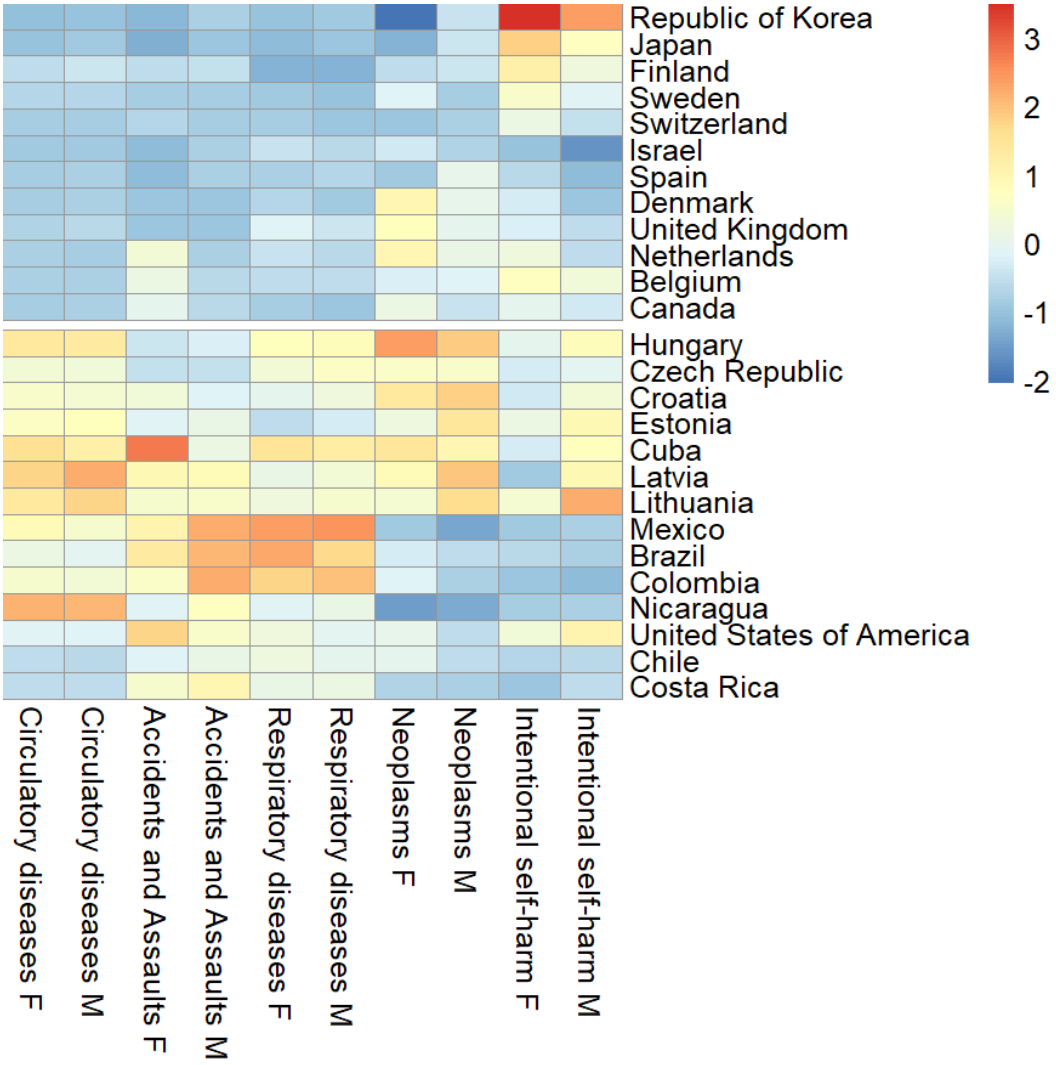


Figure 21: Heatmap of the standardized ASDRs of the 2021 data. Two clusters are given as separate blocks.

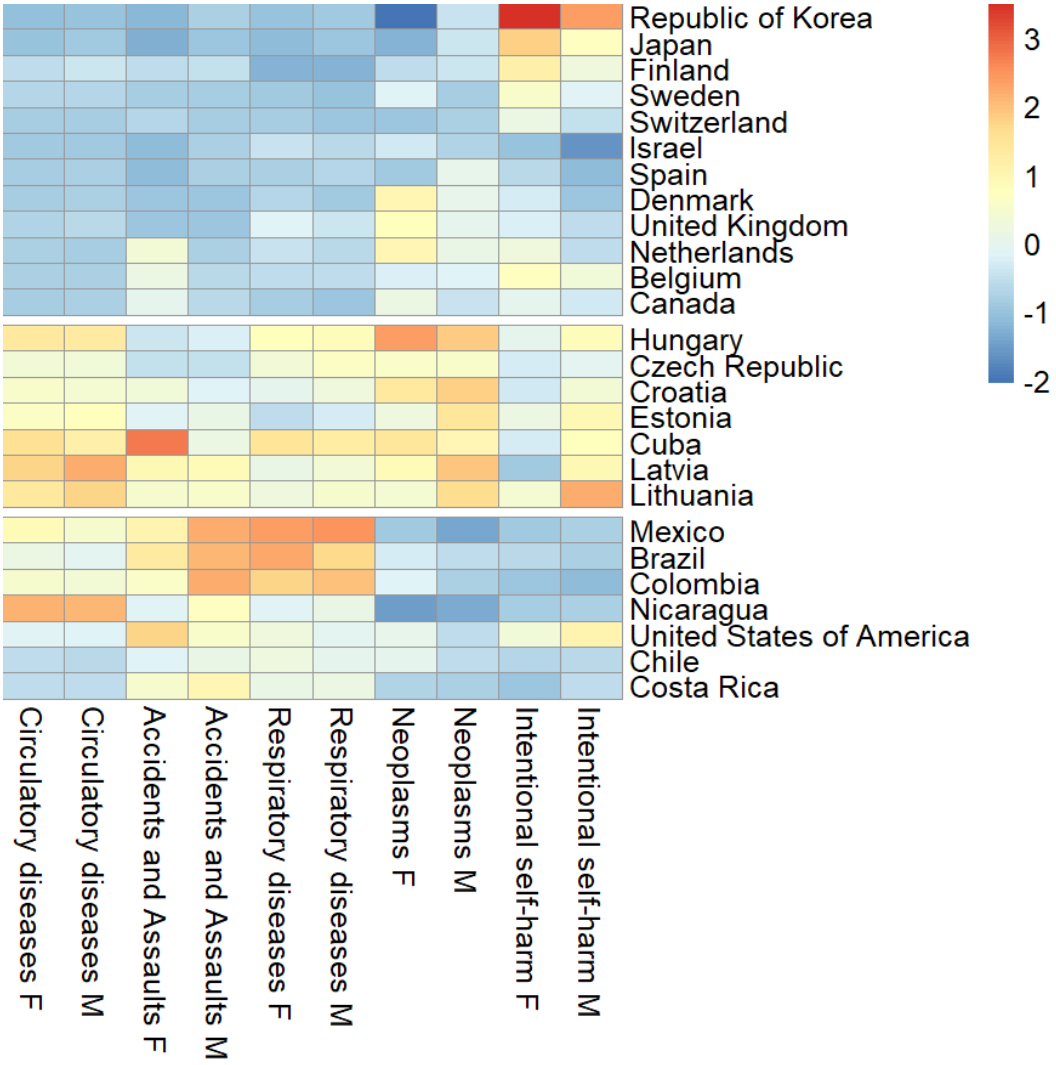


Figure 22: Heatmap of the standardized ASDRs of the 2021 data. Three clusters are given as separate blocks.

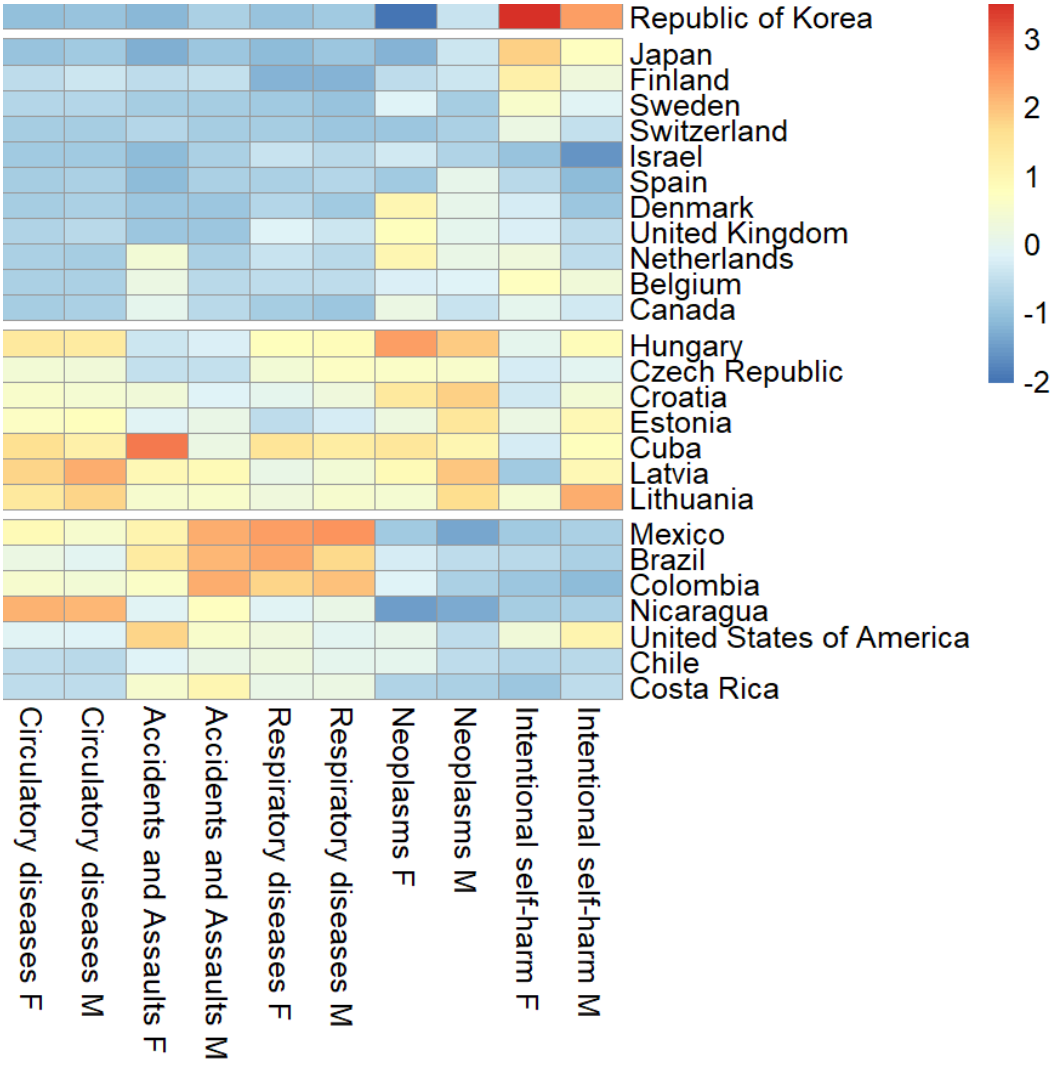


Figure 23: Heatmap of the standardized ASDRs of the 2021 data. Four clusters are given as separate blocks.

5.4 Mortality clusters along HDI data

To further assess and compare the identified clusters from the mortality data, we examine the Human Development Index (HDI) data for each analyzed country and year, along with the clusters. HDI, introduced in Section 2.4, is a composite index that combines life expectancy, education, and gross national income (GNI) per capita using a geometric mean. The HDI data used in this analysis were obtained from UNDP and OWID (2025).

The following sections present HDI values by country for the two-, three-, and four-cluster solutions in 2001, 2018, and 2021. Within each cluster, countries are ordered in decreasing HDI. Importantly, HDI is not used in the clustering procedure itself; rather, it serves as an external validation variable to assess whether the mortality-based clusters align with broader patterns of socioeconomic development, thereby strengthening the analysis. The clusters analyzed in this chapter correspond to those presented in Sections 5.1–5.3.

5.4.1 Clusters and HDI data for the year 2001

The clustering results for the year 2001 data with two, three, and four clusters, together with the country-specific HDI heatmaps, are presented in Figures 24, 25, and 26, respectively. Within each cluster, countries are ordered by HDI in descending order.

In the two-cluster solution, the Eastern European cluster exhibits relatively homogeneous HDI values, whereas the cluster comprising the remaining countries displays a substantially wider range of HDI levels. This indicates that the mortality-based clustering partially aligns with the similarities in the level of socioeconomic development.

In the three-cluster solution, the Latin American cluster includes the countries with the lowest HDI values in the dataset: Nicaragua, Colombia, and Brazil. Cuba is an exception in this configuration, as it remains in the larger cluster despite having a lower HDI than Mexico, which belongs to the Latin American cluster.

When a fourth cluster is introduced, the division of the Eastern European cluster does not follow HDI patterns in a straightforward manner. Estonia, which has the second-highest HDI among Eastern European countries, is grouped with Latvia and Lithuania, both of which rank among the three lowest in terms of HDI within the Eastern European cluster. Conversely, Croatia, despite having the second-lowest HDI among Eastern European countries, is clustered with the Czech Republic and Hungary, which are among the three highest.

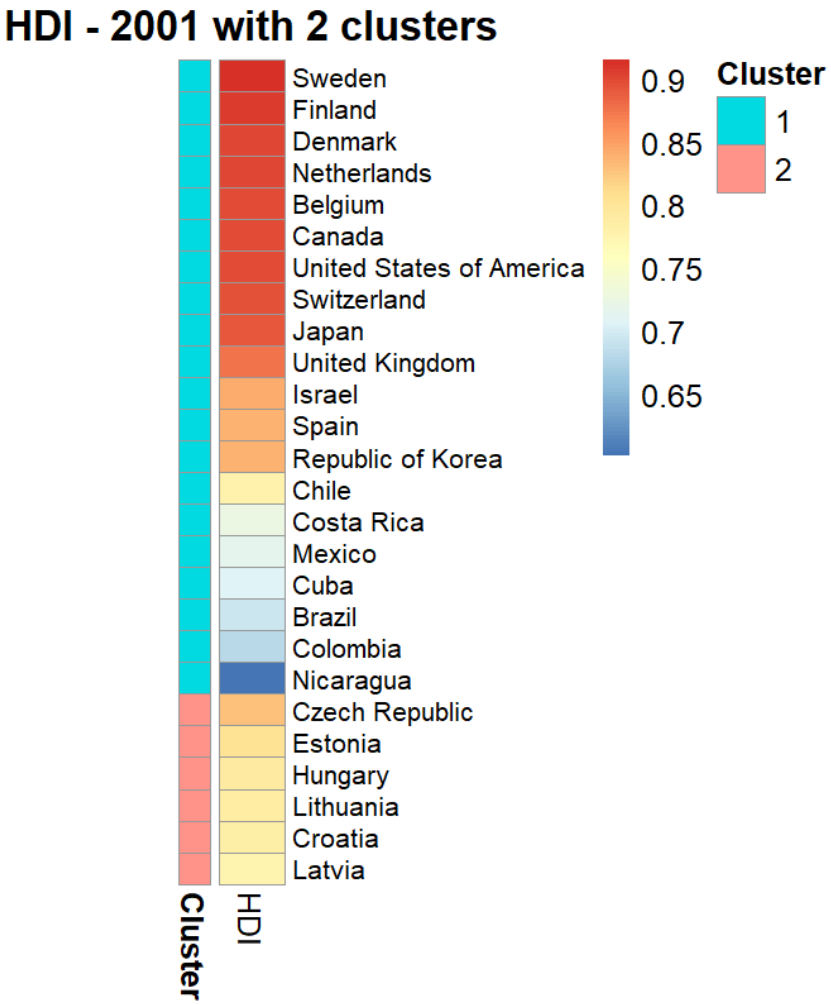


Figure 24: Heatmap of HDI in 2001. The two-cluster solution for the year 2001 mortality data is marked with separate blocks.

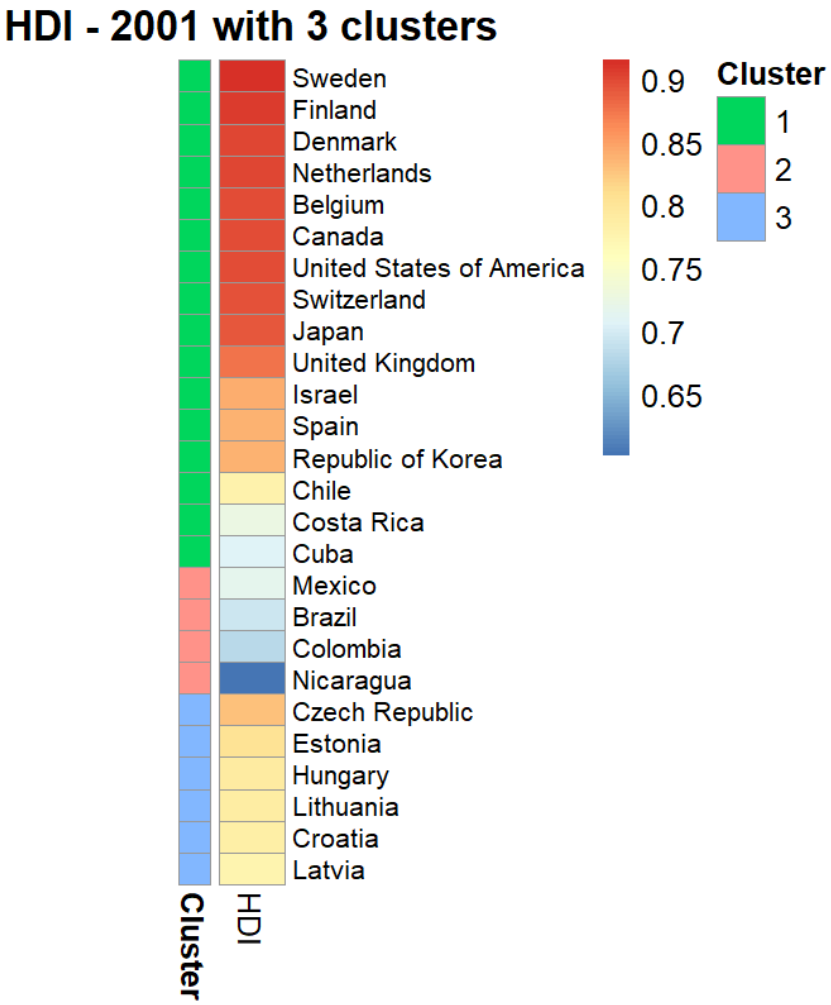


Figure 25: Heatmap of HDI in 2001. The three-cluster solution for the year 2001 mortality data is marked with separate blocks.

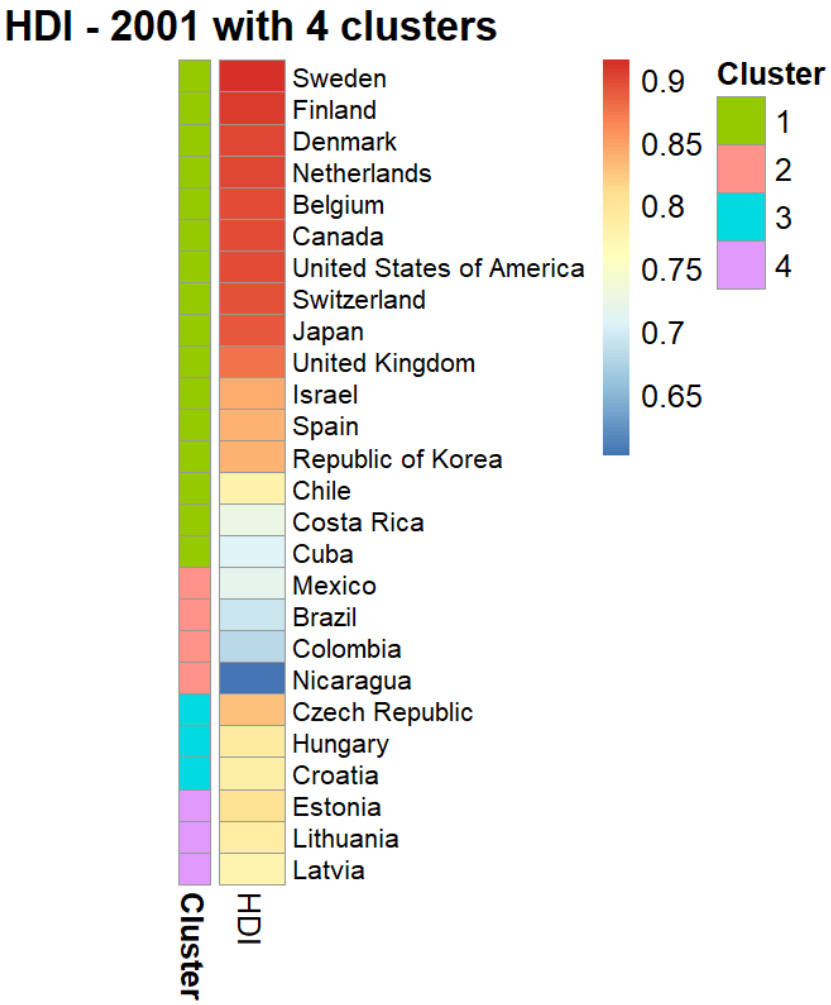


Figure 26: Heatmap of HDI in 2001. The four-cluster solution for the year 2001 mortality data is marked with separate blocks.

5.4.2 Clusters and HDI data for the year 2018

The clustering results for the year 2018 data with two, three, and four clusters, together with the country-specific HDI heatmaps, are presented in Figures 27, 28, and 29, respectively. Within each cluster, countries are ordered by HDI in descending order.

In the two-cluster solution, the change in the order of Eastern European countries by HDI is the main difference compared to the two-cluster solution for 2001. Hungary now exhibits the lowest HDI within this cluster, while Lithuania and Latvia have also surpassed Croatia. Changes are also observed in the larger cluster, where Switzerland has replaced Sweden as the country with the highest HDI. Overall, HDI values have increased across all countries between 2001 and 2018, with Lithuania experiencing the largest absolute increase (0.102) and Japan the smallest (0.029).

In the three-cluster solution, the Latin American cluster is notably more homogeneous in terms of HDI than in 2001, largely due to the exclusion of Nicaragua from this cluster. In fact, the difference between the maximum and minimum HDI value within the Latin American cluster is 0.016, compared to 0.044 in the Eastern European cluster and 0.272 in the largest cluster. This increased homogeneity suggests a closer alignment between human development and mortality-based clustering among Latin American countries in 2018.

When the number of clusters is increased to four, the largest cluster splits into two groups that differ considerably in terms of HDI variability. Of these two clusters, the one containing Switzerland is highly homogeneous with respect to HDI, whereas the other remains heterogeneous, ranging from Nicaragua at the lower end to Denmark at the higher end. Consequently, in the four-cluster configuration, three clusters exhibit relatively homogeneous HDI values, while one cluster remains heterogeneous.

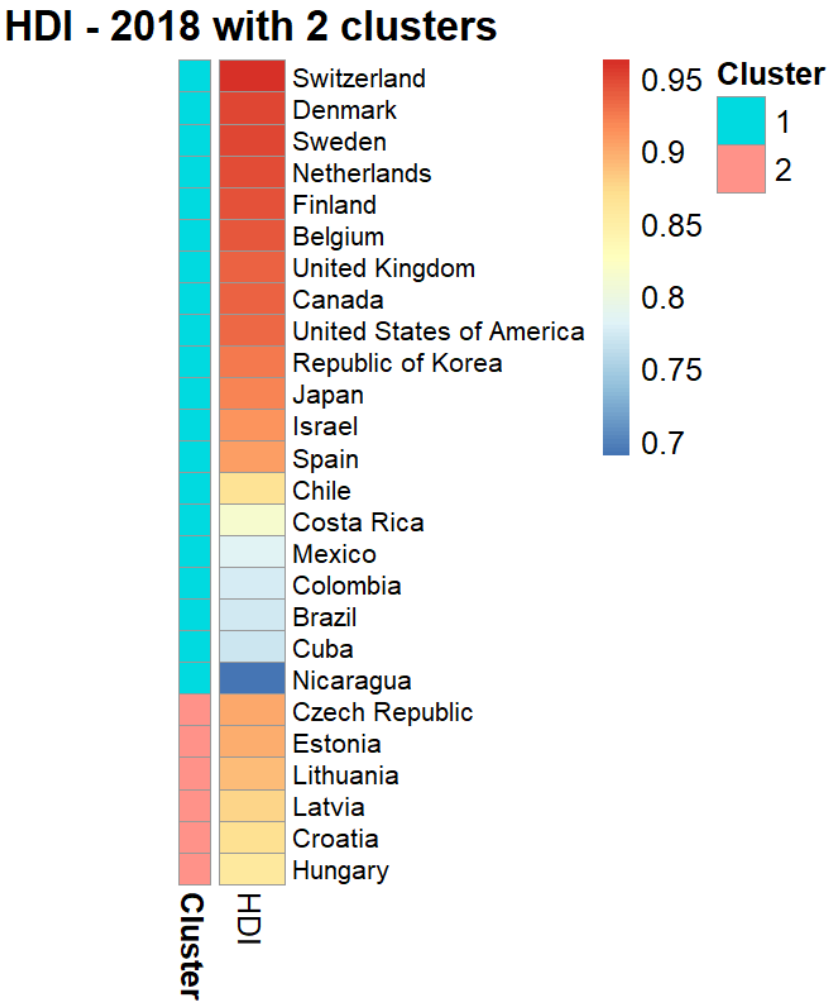


Figure 27: Heatmap of HDI in 2018. The two-cluster solution for the year 2018 mortality data is marked with separate blocks.

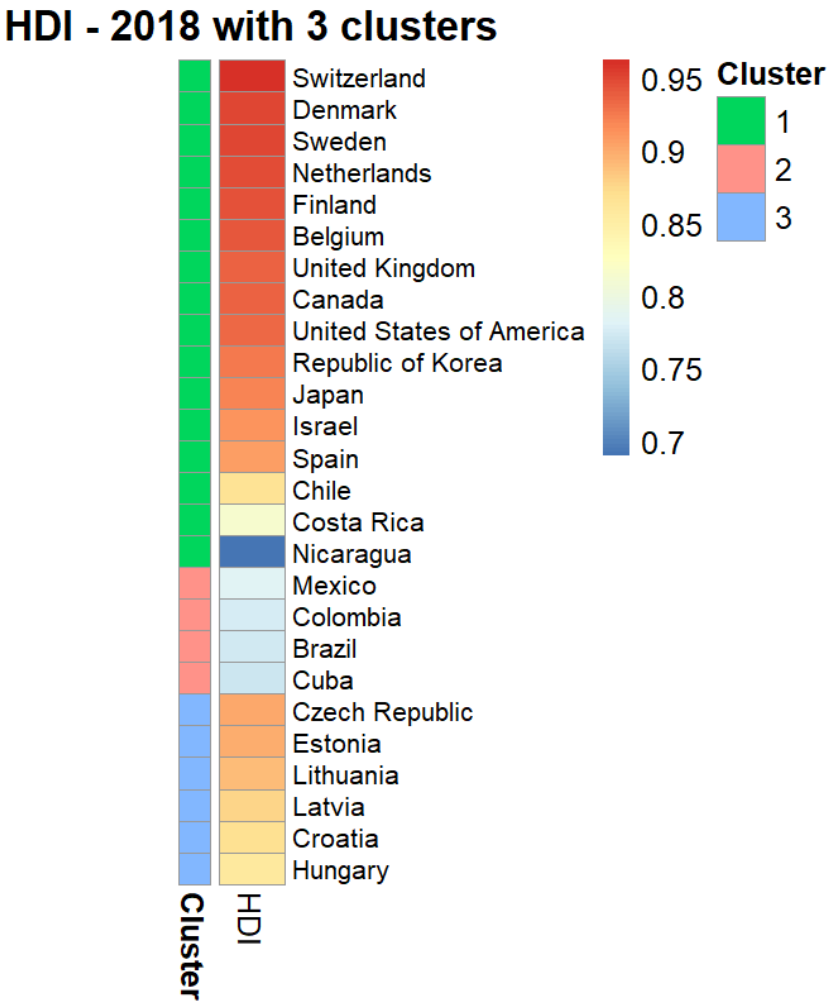


Figure 28: Heatmap of HDI in 2018. The three-cluster solution for the year 2018 mortality data is marked with separate blocks.

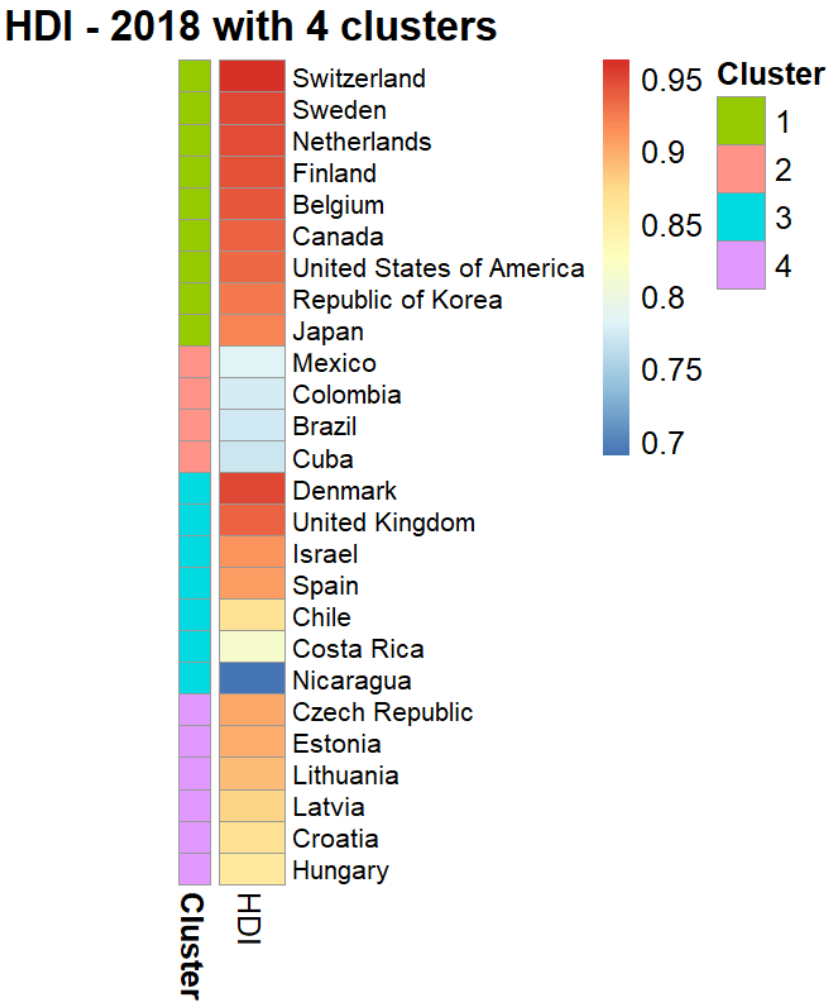


Figure 29: Heatmap of HDI in 2018. The four-cluster solution for the year 2018 mortality data is marked with separate blocks.

5.4.3 Clusters and HDI data for the year 2021

The clustering results for the year 2021 data with two, three, and four clusters, together with the country-specific HDI heatmaps, are presented in Figures 30, 31, and 32, respectively. Within each cluster, countries are ordered by HDI in descending order.

In the two-cluster solution, the clustering aligns closely with HDI levels. One cluster consists of countries with higher HDI, while the other contains countries with lower HDI values. The only notable exception to this is the United States, which has a higher HDI than Spain and Israel but is grouped into the lower-HDI cluster. Despite this exception, HDI can be seen as a strong explanatory factor for the two-cluster solution.

In addition, an examination of changes in HDI from 2018 to 2021 reveals that, unlike for the period from 2001 to 2018, HDI has declined in several countries. Countries with a negative HDI trend over this period include Cuba, Mexico, Colombia, the USA, Nicaragua, Brazil, Hungary, Latvia, Lithuania, and Chile. All of these countries are included in the lower-HDI cluster, highlighting the developmental similarities within the cluster.

When the number of clusters is increased to three, the lower-HDI cluster is further divided. In this split, the Eastern European countries, which exhibit relatively homogeneous HDI levels, together with Cuba, form a separate cluster. The remaining cluster contains countries with varying HDI values, ranging from 0,682 in Nicaragua to 0,921 in the United States. This heterogeneity indicates that HDI alone does not sufficiently explain the formation of this cluster. In the four-cluster solution, the Republic of Korea separates from the high-HDI cluster.

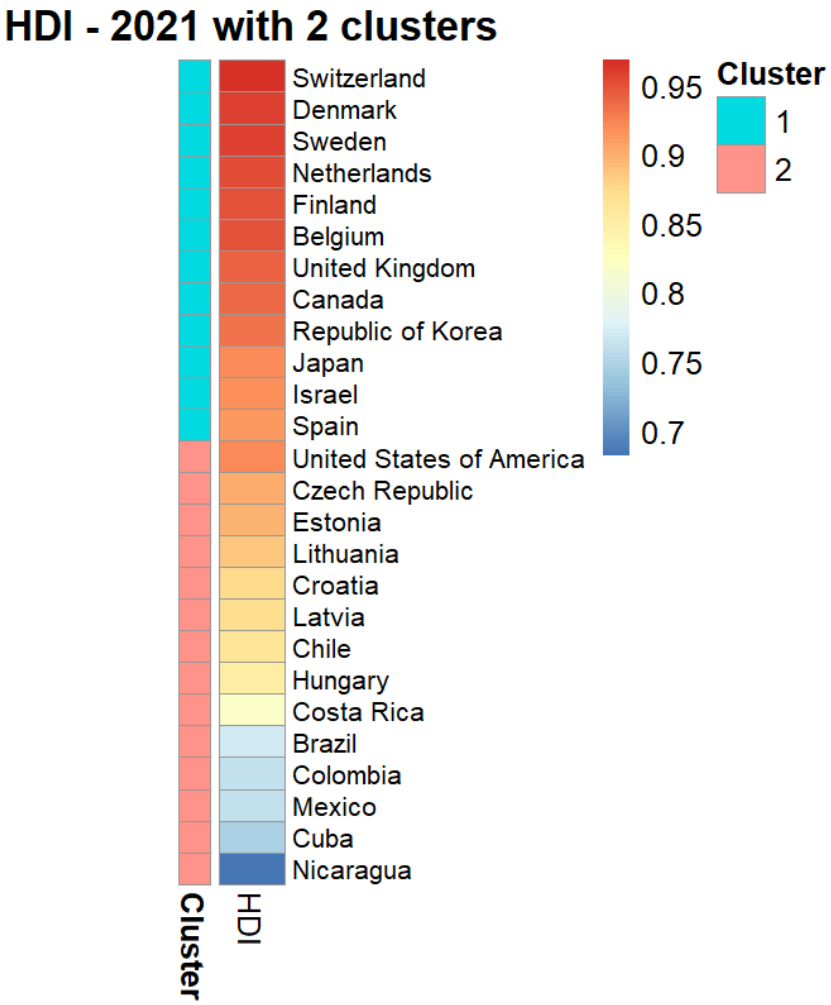


Figure 30: Heatmap of HDI in 2021. The two-cluster solution for the year 2021 mortality data is marked with separate blocks.

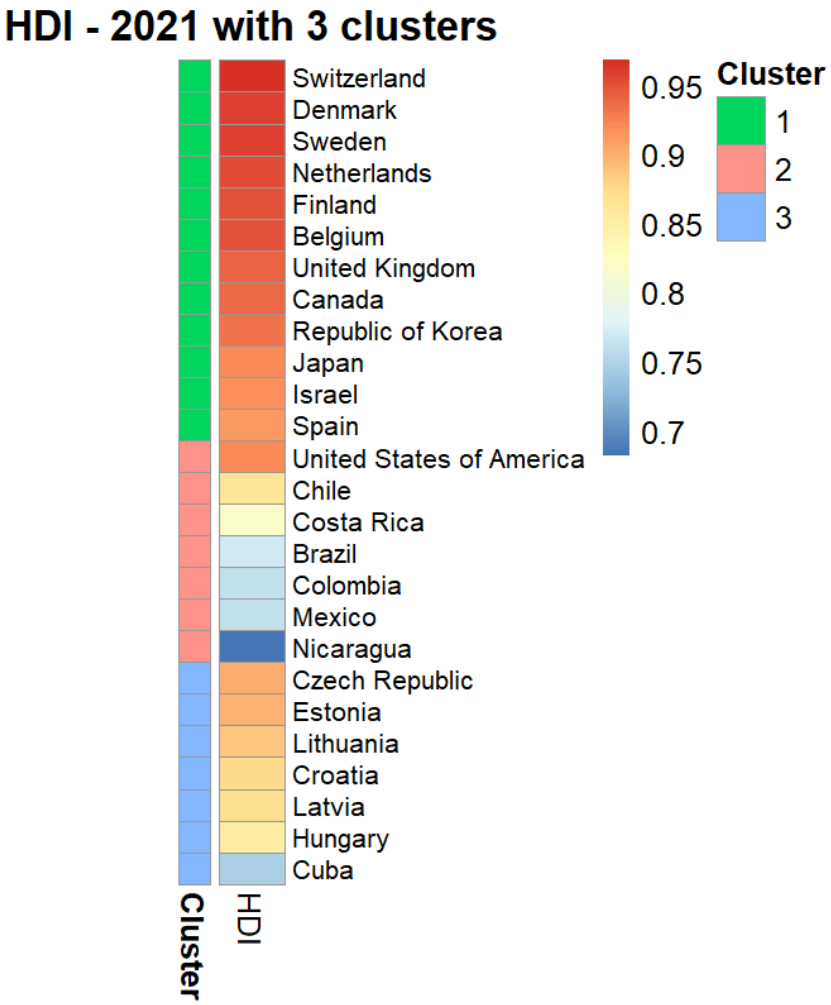


Figure 31: Heatmap of HDI in 2021. The three-cluster solution for the year 2021 mortality data is marked with separate blocks.

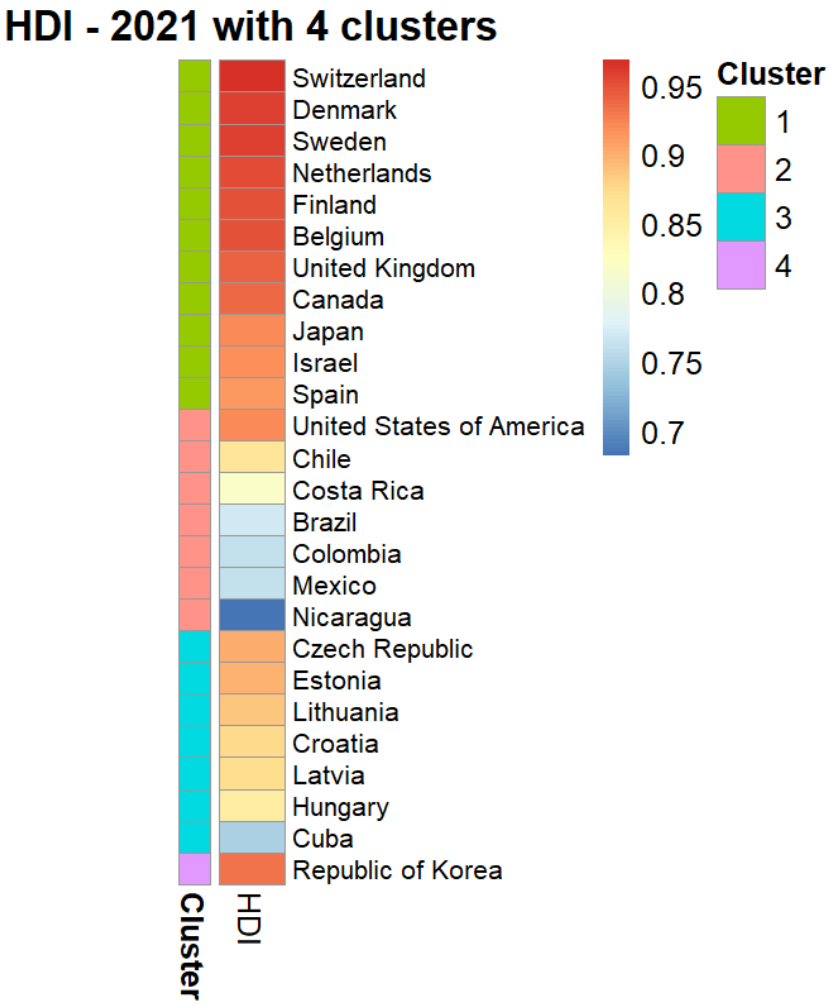


Figure 32: Heatmap of HDI in 2021. The four-cluster solution for the year 2021 mortality data is marked with separate blocks.

6 Discussion

The clustering results show that geographically aligning mortality patterns can be identified from the analyzed data. In particular, Eastern European countries consistently clustered together across years and clustering solutions, suggesting similarities in their mortality profiles that distinguish them from other regions. A similar pattern was observed among Latin American countries. These findings indicate that geographically and historically similar areas produce similar mortality patterns.

The comparison of mortality clusters with HDI further suggests that mortality patterns highly align with socioeconomic development. Of the high-HDI countries, Sweden, Finland, Switzerland, the Netherlands, Belgium, Canada, and Japan were clustered together for each considered year and number of clusters. Even though mortality profiles vary across countries, the similarity among countries with comparable HDI levels is evident.

Differences in the clustering between years provide additional insights. For 2001 and 2018, the two-cluster solutions separated Eastern European countries from the rest. In contrast, the 2021 two-cluster solution divided countries into higher- and lower-HDI clusters. The lower-HDI cluster included Eastern European and Latin American countries, as well as the United States. This shift in clustering structure is particularly interesting when the change in ASDRs due to respiratory diseases is examined between 2018 and 2021. Countries experiencing the largest absolute increase in ASDR from respiratory diseases are the countries in the lower-HDI cluster. Given that 2021 is also the first year considered after the start of the COVID-19 pandemic, these results may reflect inter-country differences in pandemic management. However, as the focus of this thesis was not at the disease level but rather on high-level causes of death, further research is required to conclusively determine this relationship.

Several limitations should be considered when interpreting the results. Even though the selected countries were classified by the WHO as having high-quality mortality data, differences in registration coverage and data completeness remain. In particular, several Latin American countries have lower death registration coverage compared to many European countries, with Nicaragua having one of the lowest coverages among the countries considered in this thesis (Mikkelsen et al., 2015). On the other hand, most of the high-HDI countries have almost complete registration coverage. This discrepancy might affect the comparability of cause-specific mortality rates between countries. In addition, this study did not take into account all causes of mortality, which might affect the clusters. For example, elevated infant and maternal mortality rates, relevant for some lower-HDI countries, might not be completely accounted for. These factors should be considered when interpreting the clustering results.

7 Conclusions

This thesis examines the structure of cause-specific mortality patterns across countries using hierarchical clustering applied to ICD-10-coded data from the WHO Mortality Database. Agglomerative hierarchical clustering with Ward's method was conducted on standardized age-standardized death rates (ASDRs) for five high-level causes of death: neoplasms, accidents and assaults, diseases of the circulatory system, and diseases of the respiratory system. The analysis was performed using data from 26 countries for three years: 2001, 2018, and 2021. The obtained clusters were also analyzed using the Human Development Index (HDI) as an external factor to determine how the clusters align with the level of socioeconomic development of the countries.

Across all three years, the clustering results demonstrate interpretable patterns where countries with geographical and developmental similarities are grouped together. Two-, three-, and four-cluster solutions were considered and examined for each of the three years to see how the clusters develop over time. While the clusters varied across years, several themes were identified. In particular, Eastern European countries frequently formed a distinct cluster characterized by relatively high ASDRs due to neoplasms and diseases of the circulatory system. Similarly, Latin American countries tended to cluster together, reflecting shared mortality profiles.

The alignment between the clusters and HDI was a notable finding of this study. Although HDI was not used in the clustering process, countries within the same mortality cluster often had similar HDI levels. Latin American countries consistently had the lowest HDI values among the analyzed countries, while Eastern European countries exhibited moderate, relatively homogeneous HDI levels. Also, countries with the highest HDI were almost consistently clustered together across the considered years. This alignment emphasizes the close relationship between socioeconomic development and mortality profiles.

The two-cluster solutions are the most interesting when comparing clustering results between the three considered years. For 2001 and 2018, Eastern European countries are separated from the rest of the countries into their own cluster. However, for 2021, the two-cluster solution separates the United States, Latin American, and Eastern European countries into their own cluster. The countries in this block have the highest absolute increase in the ASDRs due to respiratory diseases. The year 2021 is also the first year analyzed after the COVID-19 pandemic, and thus, a potential explanation for this phenomenon. However, the role of COVID-19 is not studied in this thesis.

Overall, this thesis demonstrates that hierarchical clustering of cause-specific mortality data yields meaningful and interpretable clusters that reflect both geographical and developmental similarities. Cluster patterns, particularly in cause-specific mortality data, can serve as tools for comparative health analysis across countries and for guiding policies on the future development of health care systems.

References

Omar B Ahmad, Cynthia Boschi-Pinto, Alan D Lopez, Christopher JL Murray, Rafael Lozano, and Mie Inoue. Age standardization of rates: a new WHO standard. *World Health Organization*, 2001.

Rana Ahmed, Ryan Robinson, and Kevin Mortimer. The epidemiology of noncommunicable respiratory disease in sub-Saharan Africa, the Middle East, and North Africa. *Malawi Medical Journal*, 29(2):203–211, 2017.

Anna-Maria Aksan and Shankha Chakraborty. Life expectancy across countries: Convergence, divergence and fluctuations. *World Development*, 168, 2023.

Chandrakanth Are, Shireen Rajaram, Madhuri Are, Hemanth Raj, Benjamin O Anderson, Ramesh Chaluvarya Swamy, Manavalan Vijayakumar, Tianqiang Song, Manoj Pandey, James A Edney, et al. A review of global cancer burden: trends, challenges, strategies, and a role for surgeons. *Journal of surgical oncology*, 107 (2):221–226, 2013.

Rose G Bender, Sarah B Sirota, Lucien R Swetschinski, Regina-Mae V Dominguez, Amanda Novotney, Eve E Wool, et al. Global, regional, and national incidence and mortality burden of non-COVID-19 lower respiratory infections and aetiologies, 1990–2021: a systematic analysis from the Global Burden of Disease study 2021. *The Lancet Infectious Diseases*, 24(9):974–1002, 2024.

Susanna Bennett, Kathryn A Robb, Tiago C Zortea, Adele Dickson, Cara Richardson, and Rory C O'Connor. Male suicide risk and recovery factors: A systematic review and qualitative metasynthesis of two decades of research. *Psychological Bulletin*, 149(7-8):371, 2023.

Paul M Bisca, Vu Chau, Paolo Dudine, Raphael A Espinoza, Jean-Marc Fournier, Pierre Guérin, Niels-Jakob Hansen, and Jorge Salas. *Violent Crime and Insecurity in Latin America and the Caribbean: A Macroeconomic Perspective*. International Monetary Fund, 2024.

Clarke B Blackadar. Historical review of the causes of cancer. *World journal of clinical oncology*, 7(1):54–86, 2016.

Farnçois Boissier de Sauvages. *Nosologia methodica sistens morborum classes, genera et species, juxta Sydenhami mentem et botanicorum ordinem*. Amsterdam: Frères De Tournes, 1763.

Freddie Bray, Mathieu Laversanne, Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Isabelle Soerjomataram, and Ahmedin Jemal. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 74(3):229–263, 2024.

Malika Charrad, Nadia Ghazzali, Véronique Boiteau, and Azam Niknafs. NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *Journal of Statistical Software*, 61(6):1–36, 2014. URL https://www.jstatsoft.org/v61/i06/.

Daniel Cobos Muñoz, Carla Abouzahr, and Don de Savigny. The 'Ten CRVS Milestones' framework for understanding civil registration and vital statistics systems. *BMJ global health*, 3(2), 2018.

Daniel Cobos Muñoz, Don de Savigny, Renee Sorchik, Khin S Bo, John Hart, Viola Kwa, Xavier Ngomituje, Nicola Richards, and Alan D Lopez. Better data for better outcomes: the importance of process mapping and management in CRVS systems. *BMC medicine*, 18(1):67, 2020.

Goodarz Danaei, Stephen Vander Hoorn, Alan D Lopez, Christopher JL Murray, and Majid Ezzati. Causes of cancer in the world: comparative risk assessment of nine behavioural and environmental risk factors. *The Lancet*, 366:1784–1793, 2005.

Derek De Beurs, Eiko I Fried, Karen Wetherall, Seonald Cleare, Daryl B O'Connor, Eamonn Ferguson, Ronan E O'Carroll, and Rory C O'Connor. Exploring the psychology of suicidal ideation: A theory driven network analysis. *Behaviour research and therapy*, 120:103419, 2019.

Don De Savigny, Ian Riley, Daniel Chandramohan, Frank Odhiambo, Erin Nichols, Sam Notzon, Carla AbouZahr, Raj Mitra, Daniel Cobos Muñoz, Sonja Firth, et al. Integrating community-based verbal autopsy into civil registration and vital statistics (CRVS): system-level considerations. *Global health action*, 10(1):1272882, 2017.

Mariachiara Di Cesare, Pablo Perel, Sean Taylor, Chodziwadziwa Kabudula, Honor Bixby, Thomas A Gaziano, Diana V McGhie, Jeremiah Mwangi, Borjana Pervan, Jagat Narula, et al. The heart of the world. *Global heart*, 19(1):11, 2024.

Majid Ezzati, Ziad Obermeyer, Ioanna Tzoulaki, Bongani M Mayosi, Paul Elliott, and David A Leon. Contributions of risk factors and medical care to cardiovascular mortality trends. *Nature Reviews Cardiology*, 12(9):508–530, 2015.

Lisa M Force, Jonathan M Kocarnik, Miranda L May, Kayleigh Bhangdia, Andrew Crist, Louise Penberthy, Natalie Pritchett, Alistair Acheson, Lee Deitesfeld, Hasan Aalruz, et al. The global, regional, and national burden of cancer, 1990–2023, with forecasts to 2050: a systematic analysis for the Global Burden of Disease Study 2023. *The Lancet*, 406(10512):1565–1586, 2025.

Juanita A Haagsma, Nicholas Graetz, Ian Bolliger, et al. The global burden of injury: incidence, mortality, disability-adjusted life years and time trends from the Global Burden of Disease study 2013. *Injury prevention*, 22(1):3–18, 2016.

Wolfgang Härdle and Léopold Simar. *Applied multivariate statistical analysis, 4th edition*. Springer, 2015.

Melonie Heron. Deaths: Leading Causes for 2016. *National vital statistics reports: from the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System*, 67(6):1–77, 2018.

JA Hirsch, G Nicola, G McGinty, RW Liu, RM Barr, MD Chittle, and L Manchikanti. ICD-10: history and context. *American Journal of Neuroradiology*, 37(4):596–599, 2016.

Selen Y Işikhan and Dilek Güleç. The clustering of world countries regarding causes of death and health risk factors. *Iranian Journal of Public Health*, 47(10):1520, 2018.

Spencer L James, Lydia R Lucchesi, Catherine Bisignano, Chris D Castle, Zachary V Dingels, Jack T Fox, Erin B Hamilton, Nathaniel J Henry, Kris J Krohn, Zichen Liu, et al. The global burden of falls: global, regional and national estimates of morbidity and mortality from the Global Burden of Disease Study 2017. *Injury prevention*, 26(Suppl 2):i3–i11, 2020.

Angur Mahmud Jarman. Hierarchical cluster analysis: Comparison of single linkage, complete linkage, average linkage and centroid linkage method. *Georgia Southern University*, 29:90240, 2020.

Bindu Kalesan, Matthew E Mobily, Olivia Keiser, Jeffrey A Fagan, and Sandro Galea. Firearm legislation and firearm mortality in the USA: a cross-sectional, state-level study. *The Lancet*, 387(10030):1847–1855, 2016.

J Nicholas Kassebaum, Megha Aurora, Ryan M Barber, et al. Global, regional, and national disability-adjusted life-years (DALYs) for 315 diseases and injuries and healthy life expectancy (HALE), 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet*, 388:1603–58, 2016.

Leonard Kaufman and Peter J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, New York, 1990. ISBN 0-471-73578-7.

Uzma Rahim Khan, Junaid A Razzak, and Martin Gerdin Wärnberg. Global trends in adolescents' road traffic injury mortality, 1990–2019. *Archives of disease in childhood*, 106(8):753–757, 2021.

Vinay Kumar, Abul K. Abbas, and Nelson Fausto. *Robbins and Cotran Pathologic Basis of Disease*. Elsevier Health Sciences, Philadelphia, 7 edition, 2005.

Daniel T Lackland, Edward J Roccella, Anne F Deutsch, Myriam Fornage, Mary G George, George Howard, Brett M Kissela, Steven J Kittner, Judith H Lichtman, Lynda D Lisabeth, et al. Factors influencing the decline in stroke mortality: a statement from the American Heart Association/American Stroke Association. *Stroke*, 45(1):315–353, 2014.

Ainhoa-Elena Léger and Stefano Mazzuco. What can we learn from the functional clustering of mortality data? An application to the human mortality database. *European Journal of Population*, 37(4):769–798, 2021.

Marie-Jeanne Lesot, Maria Rifqi, and Hamid Benhadda. Similarity measures for binary and numerical data: a survey. *International Journal of Knowledge Engineering and Soft Data Paradigms*, 1(1):63–84, 2009.

Aleša Lotrič Dolinar, Jože Sambt, and Simona Korenjak-Černe. Clustering EU countries by causes of death. *Population Research and Policy Review*, 38(2): 157–172, 2019.

T Soni Madhulatha. An overview on clustering methods. *arXiv preprint arXiv:1205.1117*, 2012.

J John Mann, Christina A Michel, and Randy P Auerbach. Improving suicide prevention through evidence-based strategies: a systematic review. *American journal of psychiatry*, 178(7):611–624, 2021.

Colin Mathers, Gretchen Stevens, Dan Hogan, Wahyu Retno Mahanani, and Jessica Ho. Global and Regional Causes of Death: Patterns and Trends, 2000–15, 2017. URL http://europepmc.org/books/NBK525280.

Vivek Mehta, Seema Bawa, and Jasmeet Singh. Analytical review of clustering techniques and proximity measures. *Artificial Intelligence Review*, 53(8):5995–6023, 2020.

Shanthi Mendis, Pekka Puska, and Bo Norrving. *Global atlas on cardiovascular disease prevention and control.* World Health Organization in collaboration with the World Heart Federation and the World Stroke Organization., 2011.

Lene Mikkelsen, David E Phillips, Carla AbouZahr, Philip W Setel, Don De Savigny, Rafael Lozano, and Alan D Lopez. A global assessment of civil registration and vital statistics systems: monitoring data quality and progress. *The Lancet*, 386 (10001):1395–1406, 2015.

Iwao Milton Moriyama, Ruth M Loy, Alastair Hamish Tearloch Robb-Smith, Harry Michael Rosenberg, and Donna L Hoyert. *History of the statistical classification of diseases and causes of death*. National Center for Health Statistics, 2011.

Christopher J L Murray, Alan D Lopez, Jeremy T Barofsky, Chloe Bryson-Cahn, and Rafael Lozano. Estimating population cause-specific mortality fractions from in-hospital mortality: validation of a new method. *PLoS Medicine*, 4(11):e326, 2007.

Fionn Murtagh and Pedro Contreras. Algorithms for hierarchical clustering: an overview. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 2 (1):86–97, 2012.

Peter Jongho Na, Jeonghyun Shin, Ha Rim Kwak, Jaewon Lee, Dylan J Jester, Piumee Bandara, Jim Yong Kim, Christine Y Moutier, Robert H Pietrzak, Maria A Oquendo, et al. Social determinants of health and suicide-related outcomes: a review of meta-analyses. *JAMA psychiatry*, 2025.

Mohsen Naghavi, Laurie B. Marczak, Michael Kutz, Katya Anne Shackelford, Megha Arora, Molly Miller-Petrie, Miloud Taki Eddine Aichour, Nadia Akseer, Rajaa M. Al-Raddadi, Khurshid Alam, Suliman A. Alghnam, Carl Abelardo T. Antonio, Olatunde Aremu, Amit Arora, Mohsen Asadi-Lari, Reza Assadi, Tesfay Mehari Atey, et al. Global Mortality From Firearms, 1990-2016. *JAMA*, 320(8):792–814, 08 2018. ISSN 0098-7484. doi: 10.1001/jama.2018.10060. URL https://doi.org/10.1001/jama.2018.10060.

Mohsen Naghavi, Kanyin Liane Ong, Amirali Aali, Hazim S Ababneh, Yohannes Habtegiorgis Abate, Cristiana Abbafati, Rouzbeh Abbasgholizadeh, Mohammadreza Abbasian, Mohsen Abbasi-Kangevari, Hedayat Abbastabar, et al. Global burden of 288 causes of death and life expectancy decomposition in 204 countries and territories and 811 subnational locations, 1990–2021: a systematic analysis for the Global Burden of Disease Study 2021. *The Lancet*, 403(10440): 2100–2132, 2024.

National Cancer Institute. What Is Cancer? https://www.cancer.gov/about-cancer/understanding/what-is-cancer. Accessed: 2025-10-30.

Official Statistics of Finland (OSF). Causes of Death [e-publication]. Quality Description: Causes of Death 2020. http://stat.fi/til/ksyyt/2020/ksyyt_2020_2021-12-10_laa_001_en.html, 2020. ISSN 1799-5078. Helsinki: Statistics Finland. [Accessed: 15 October 2025].

Lucia Otero Varela, Chelsea Doktorchik, Natalie Wiebe, Hude Quan, and Catherine Eastwood. Exploring the differences in ICD and hospital morbidity data collection features across countries: an international survey. *BMC Health Services Research*, 21(1):308, 2021.

Liyuan Pu, Li Wang, Ruijie Zhang, Tian Zhao, Yannan Jiang, and Liyuan Han. Projected global trends in ischemic stroke incidence, deaths and disability-adjusted life years from 2020 to 2030. *Stroke*, 54(5):1330–1339, 2023.

Nannan Qian, Chengcheng Lu, Taohua Wei, Wenming Yang, Han Wang, Huaizhen Chen, Jun Li, Sihuan Zhu, Weiqi Wang, and Ningshu Shao. Epidemiological trends and forecasts in stroke at global, regional and national levels. *Journal of Stroke and Cerebrovascular Diseases*, page 108347, 2025.

Syeda Amna Rizvi, Muhammad Umair, and Muhammad Aamir Cheema. Clustering of countries for COVID-19 cases based on disease prevalence, health systems and environmental indicators. *Chaos, Solitons & Fractals*, 151:111240, 2021.

Saeid Safiri, Ata Mahmoodpoor, Ali-Asghar Kolahi, Seyed Aria Nejadghaderi, Mark JM Sullman, Mohammad Ali Mansournia, Khalil Ansarin, Gary S Collins, Jay S Kaufman, and Morteza Abdollahi. Global burden of lower respiratory infections during the last three decades. *Frontiers in public health*, 10:1028525, 2023.

Ailiana Santosa, Stig Wall, Edward Fottrell, Ulf Högberg, and Peter Byass. The development and experience of epidemiological transition theory over four decades: a systematic review. *Global health action*, 7(1):23574, 2014.

Shweta Sharma, Neha Batra, et al. Comparative study of single linkage, complete linkage, and ward method of agglomerative clustering. In *2019 international conference on machine learning, big data, cloud and parallel computing (COMITCon)*, pages 568–573. IEEE, 2019.

Ali Seyed Shirkhorshidi, Saeed Aghabozorgi, and Teh Ying Wah. A comparison study on similarity and dissimilarity measures in clustering continuous data. *PloS one*, 10(12):e0144059, 2015.

Lindsey A Torre, Rebecca L Siegel, Elizabeth M Ward, and Ahmedin Jemal. Global cancer incidence and mortality rates and trends—an update. *Cancer epidemiology, biomarkers & prevention*, 25(1):16–27, 2016.

UNDP. *Human Development Report 1990*. Oxford University Press, New York, 1990.

UNDP and OWID. Human Development Index – UNDP. https://archive.ourworldindata.org/20251209-133038/grapher/human-development-index.html, 2025. Human Development Report (2025), with minor processing by Our World in Data.

United Nations, Department of Economic and Social Affairs, Population Division. World Population Prospects 2024, Online Edition. Online database, 2024. URL https://population.un.org/wpp/downloads?folder=Standard%20Projections&group=Population. Accessed: 2025-11-10.

Haidong Wang, Katherine R Paulson, Spencer A Pease, Stefanie Watson, Haley Comfort, Peng Zheng, Aleksandr Y Aravkin, Catherine Bisignano, Ryan M Barber, Tahiya Alam, et al. Estimating excess mortality due to the COVID-19 pandemic: a systematic analysis of COVID-19-related mortality, 2020–21. *The Lancet*, 399 (10334):1513–1536, 2022.

Yang Wang and Jinfeng Wang. Modelling and prediction of global non-communicable diseases. *BMC public health*, 20(1):822, 2020.

Bethany A West, Rose A Rudd, Erin K Sauber-Schatz, and Michael F Ballesteros. Unintentional injury deaths in children and youth, 2010–2019. *Journal of safety research*, 78:322–330, 2021.

Christopher P Wild, Carolina Espina, Linda Bauld, Bernardo Bonanni, Hermann Brenner, Karen Brown, Joakim Dillner, David Forman, Ellen Kampman, Mef Nilbert, et al. Cancer prevention Europe. *Molecular oncology*, 13(3):528–534, 2019.

World Health Organization. *International statistical classification of diseases and related health problems. - 10th revision, Sixth edition, 2019.*, volume 2. World Health Organization, 2019.

World Health Organization. SCORE for Health Data Technical Package: Global Report on Health Data Systems and Capacity, 2020. Technical report, World Health Organization, Geneva, Switzerland, 2021. Licence: CC BY-NC-SA 3.0 IGO.

World Health Organization. *World Health Statistics 2023: Monitoring Health for the SDGs, Sustainable Development Goals*. World Health Organization, Geneva, 2023. URL https://cdn.who.int/media/docs/default-source/gho-documents/world-health-statistic-reports/2023/world-health-statistics-2023_20230519_.pdf. Licence: CC BY-NC-SA 3.0 IGO.

World Health Organization. WHO Methods and Data Sources for Country-Level Causes of Death 2000–2021. Global Health Estimates Technical Paper WHO/DDI/DNA/GHE/2024.2, Department of Data and Analytics (DNA), Division of Data, Analytics and Delivery for Impact (DDI), World Health Organization, Geneva, Switzerland, May 2024. URL https://cdn.who.int/media/docs/default-source/gho-documents/global-health-estimates/ghe2021_cod_methods.pdf?sfvrsn=dca346b7_1.

World Health Organization. WHO Mortality Database. Data repository, 2025a. URL https://www.who.int/data/data-collection-tools/who-mortality-database. Accessed: 2025-10-15; Available at https://www.who.int/data/data-collection-tools/who-mortality-database.

World Health Organization. Suicide Worldwide in 2021: Global Health Estimates, 2025b. Licence: CC BY-NC-SA 3.0 IGO.

World Health Organization. Unintentional injuries — WHO Mortality Database. Online visualisation platform, 2025c. URL https://platform.who.int/mortality/themes/theme-details/topics/topic-details/MDB/unintentional-injuries. Accessed: 2025-11-21.

Rui Xu and Donald Wunsch. Survey of clustering algorithms. *IEEE Transactions on neural networks*, 16(3):645–678, 2005.

Paul Siu Fai Yip, Yan Zheng, and Clifford Wong. Demographic and epidemiological decomposition analysis of global changes in suicide rates and numbers over the period 1990–2019. *Injury prevention*, 28(2):117–124, 2022.