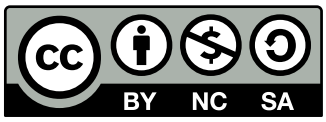
Aalto University
School of Science

Master's Programme in Mathematics and Operations Research

On Cluster Structures of Cancer Incidence and Mortality Data Over Time in Finland

Tommi Huhtinen

Master's Thesis
2025



Aalto University
School of Science

Author Tommi Huhtinen

Title On Cluster Structures of Cancer Incidence and Mortality Data Over Time in Finland

Degree programme Mathematics and Operations Research

Major Systems and Operations Research

Supervisor Prof. Pauliina Ilmonen

Advisor Prof. Pauliina Ilmonen

Date 12 August 2025 **Number of pages** 70+80 **Language** English

Abstract

The cancer burden is increasing globally. In 2022, nearly 20 million new cancer cases were diagnosed, and by 2050, this figure has been estimated to increase by 77%, or over 35 million new cases. Part of this development can be explained by the expected increase in population during the same time frame from 8.0 billion to 9.7 billion. In addition, the share of people aged 65 years and above is expected to increase from 10% to 16%, and aging, after all, is among the main risk factors causing cancer. However, our current way of living and environment are associated with many other risk factors causing cancer, including an unbalanced diet, obesity, physical inactivity, smoking, alcohol use, microplastics, and xenoestrogens.

In this thesis, cluster structures of cancer incidence and mortality data from 1962 to 2022 in Finland are identified and analyzed. The analysis is divided between females and males, and different age groups, ranging from 20-29 to 70-79 years. Both unstandardized and standardized data is used. To identify the cluster structures, an agglomerative hierarchical clustering algorithm is utilized, combined with a tailored proximity measure and the average linkage method. After employing the clustering algorithm, the resulting cluster structures are described, differences in the cluster structures between different subgroups determined by age and gender are evaluated, and whether hormone-related cancers emerge in the same cluster due to the Western lifestyle is considered. To support the analysis, a description of the Western lifestyle, as well as the associations between its components and cancer, are also provided as part of the thesis.

In terms of results of the thesis, it was observed that in many cases there is one large cluster containing many different cancers, while in the remaining clusters, there is often only one, sometimes two cancers per cluster. As expected, it appeared that differences in the scale affected the resulting clustering structures a bit in the case of unstandardized data. It was also discovered that hormone-related cancers, such as breast, cervical, and prostate cancer often formed clusters of their own, contrary to forming joint clusters with other hormone-related cancers. In addition, lung and tracheal cancer was observed in many cases, both among females and males, form a cluster of its own, suggesting possible changes in smoking behavior.

Keywords Agglomerative hierarchical clustering, cancer incidence, cancer mortality, functional data analysis, Western lifestyle

Aalto-yliopisto
Perustieteiden korkeakoulu

Tekijä Tommi Huhtinen

Työn nimi Klusterirakenteista syöpäilmaantuvuus ja -kuolleisuusdatassa yli ajan Suomessa

Koulutusohjelma Matematiikka ja operaatiotutkimus

Pääaine Systeemi- ja operaatiotutkimus

Työn valvoja Prof. Pauliina Ilmonen

Työn ohjaaja Prof. Pauliina Ilmonen

Päivämäärä 12.8.2025 **Sivumäärä** 70+80 **Kieli** englanti

Tiivistelmä

Maailmanlaajuinen syöpäkuorma kasvaa. Vuonna 2022 diagnosoitiin lähes 20 miljoonaa uuttaa syöpätapausta, ja vuoteen 2050 mennessä tämän luvun on ennustettu kasvavan 77%:lla, joka vastaa yli 35 miljoonaa uutta tapausta. Osa tästä kehityksestä voidaan selittää odotetulla väestönkasvulla 8,0 miljardista 9,7 miljardiin samalla aikavälillä. 65-vuotiaiden ja vanhempien osuuden väestöstä odotetaan lisäksi kasvavan 10%:sta 16%:iin, ja ikääntyminen onkin yksi merkittävimmistä syövän riskitekijöistä. Nykyinen elämäntapamme ja ympäristömme altistavat monille muille syövän riskitekijöille. Näitä ovat epätasapainoinen ruokavalio, lihavuus, liikkumattomuus, tupakointi, alkoholin käyttö, mikromuovit ja ksenoestrogeenit.

Tässä diplomityössä tunnistetaan ja analysoidaan klusterirakenteita Suomen syöpäilmaantuvuus ja -kuolleisuusdatassa vuodesta 1962 vuoteen 2022. Analyysi tehdään naisille ja miehille sekä eri ikäryhmille, jotka käsittävät ikäryhmät 20-29-vuotiaista 70-79-vuotiaisiin. Työssä käytetään sekä standardoimatonta että standardoitua dataa. Klusterirakenteiden tunnistamiseksi hyödynnetään kasaavaa hierarkkista klusterointialgoritmia, johon on yhdistetty räätälöity etäisyysmitta sekä keskiarvoinen linkitysmenetelmä. Klusterirointialgoritmin avulla tunnistettuja klusterirakenteita kuvaillaan, eroja eri alaryhmien, jotka määräytyvät sukupuolen ja ikäryhmän mukaan, arvioidaan sekä tutkitaan, erottuvatko hormonaaliset syövät omaksi klusterikseen länsimaisen elämäntavan seurauksena. Analyysin tueksi työhön sisältyy kuvaus länsimaisesta elämäntavasta sekä sen osatekijöiden ja syövän välisistä yhteyksistä.

Huomattiin, että monessa tapauksessa muodostuu yksi iso klusteri, joka sisältää monta eri syöpää, kun taas jäljelle jäävät klusterit muodostuvat usein yhdestä syövästä, joskus kahdesta. Odotusten mukaisesti vaikutti siltä, että erot skaalassa vaikuttivat tuloksiin hieman, kun tarkasteltiin standardoimatonta dataa. Hormonaalisten syöpien, kuten rinta-, eturauhasen ja kohdunkaulan syöpä, kohdalla huomattiin, että nämä syövät muodostivat usein kukin oman klusterinsa eivätkä yhteisiä klustereita muiden hormonaalisten syöpien kanssa. Lisäksi keuhko- ja henkitorven syövän huomattiin usein, niin naisten kuin miesten tapauksessa, muodostavan oman klusterinsa, mikä saattaa johtua muutoksista tupakoinnissa.

Avainsanat Funktionaalinen data-analyysi, kasaava hierarkkinen klusterointi, länsimainen elämäntapa, syöpäesiintyvyys, syöpäkuolleisuus

Preface

First and foremost, I would like to thank Professor Pauliina Ilmonen for her role both as the advisor and supervisor of this thesis. Not only did she provide me with this interesting topic but also enabled a smooth thesis process from start to finish.

Second, I would like to thank my friends, family, and everyone else who has provided their support for me not only during the thesis process but also more broadly throughout my studies. I am extremely grateful for having you in my life.

Helsinki, 6 August 2025

Tommi Huhtinen

Contents

Abbreviations

AICR	American Institute for Cancer Research
FDA	Functional Data Analysis
MAD	Median Absolute Deviation
WCRF	World Cancer Research Fund

1 Introduction

1.1 Cancer as a disease

Cells are the most fundamental building blocks of all living things [1], and a human body, for example, consists of around 30 trillion cells [2]. Individual cells often specialize in a certain function, and by cooperating with other specialized cells they are able to form larger entities, such as organs and, eventually, living organisms [1]. In addition, in their nuclei in structures known as chromosomes, cells contain the genetic material of creatures, that is, genes [3]. The genes can be viewed as instructions on how to construct a particular protein based on a given chain of amino acids [1], [2]. In this way, the cell is able to maintain its unique character when it grows and splits [1].

During the lifetime of a cell, the genes it contains might experience changes, or mutations [3]. Usually, one gene mutation is not enough to result in a cancer [3], [4], but the required number of changes is roughly half a dozen [2], [5]. In addition, the human body has developed a system to repair these changes [3]. However, mutations to the so-called cancer genes, oncogenes and tumor suppressor genes, can be especially harmful, as they might break the repair system [2] - [4]. In a normal setting, these genes participate in supporting and regulating cell growth [2]. When mutated, however, the oncogenes can cause excessive growth of the cell, while the tumor suppressing genes might become inactive, thereby failing to prevent disproportionate growth and promote cancer [2], [3]. Then, after thousands of divisions of the cancer cell and in most cases years of time, a solid tumor can be seen with an X-ray or felt by hand [3].

It is worth mentioning that instead of one disease, cancer refers to a large group of diseases [3], [6]. The abnormal growth of the cell tissue is common for them [3], [6]. Cancer can start to develop in almost any tissue or organ in the body and is named according to the place where it started [6]. Cancer may also spread to other tissues or organs through a process called metastasis [2], [6]. This is dangerous, as the spreading cancer might disrupt a vital organ [2] and hence, is a major cause of death from cancer [6]. Metastasis is characteristic for malignant tumors, also known as cancer, but there exist also other forms of tumors called benign tumors that are localized, grow slowly, and are usually not deathly [3].

1.2 Global cancer burden

As of 2021, cancer was the second most common cause of death after cardiovascular diseases [7]. According to estimates presented in [8], there were 19,964,811 new incidences and 9,736,779 deaths caused by cancer in 2022, when all cancers and all sexes were considered. Given the total of 62,278,628 deaths in that year [9], a bit less than one sixth of the deaths was caused by cancer. For females, the three cancers causing the most deaths were breast, lung, and colorectal cancer, while for males they were lung, liver, and colorectal cancer [8]. In terms of incidence, the authors discovered the leading three cancers for females matched the three cancers causing the most deaths, whereas for males, prostate cancer, causing the fifth most deaths, together with lung and colorectal cancer resulted in the most diagnosed cancer cases.

As for liver cancer, it caused the fifth most cancer cases among males in 2022 [8].

By 2050, the global incidence burden from cancer has been estimated to increase 77% from the 2022 figure, which translates into over 35 million new cancer cases, assuming that the current incidence rate remains constant and given the projected population growth and aging changes [8]. That is, the population is expected to increase from 8.0 billion in 2022 to 9.7 billion in 2050 [10], and the share of people aged 65 years and above is expected to increase from 10% in 2022 to 16% in 2050 [11]. Hence, there are going to be not only more people who might get a cancer but also more people who are more likely to develop a cancer, as aging is among the main risk factors causing cancer [12] - [14].

Aging, however, is not the only risk factor. Other risk factors include lifestyle factors such as nutrition, smoking, alcohol use, and physical inactivity [15], [16], obesity [17], medical treatments, such as diagnostic X-rays [18], hormone drugs, and drugs suppressing the immune system, naturally occurring exposures such as ultraviolet radiation, radon, and infectious agents, [16], pollution [15], [16], and microplastics [19]. Keeping in mind that the above is only a subset of all known carcinogens, or substances or exposures that may lead to cancer [16], exposure to one or multiple carcinogens appears inevitable in our current way of living.

1.3 Research question and structure of the thesis

One hypothesis behind the thesis is that the cancer incidence and mortality rates have not remained constant over time, even after aging of the population has been taken into account. This assumes that there has been a change in exposure to different carcinogens due to changes in the lifestyle or in the environment we live in. On the other hand, it should be noted that increased knowledge about cancer and carcinogens, improved treatments, more frequent screening, and changes in the threshold for diagnosing cancer are factors that could also affect these rates.

To study the hypothesis, Finnish cancer incidence and mortality data from 1962 to 2022 is utilized, retrieved from [20]. In addition, the incidence and mortality rates of some cancers are expected to develop similarly to each other, but not across all cancers. In particular, the interest is to uncover any distinguishable cluster structures in the data. Moreover, due to variations in hormonal activity by age and gender, any possible differences in the cluster structures between different gender and age groups are also examined. Lastly, as the Western lifestyle is present in Finland, it is studied whether hormone-related cancers form a cluster structure of their own. The study of the aforementioned aspects translates into the following, threefold research question:

- What kind of cluster structures can be identified from the cancer incidence and mortality data over time in Finland?
- How do the cluster structures differ from each other for different subgroups determined by age and gender?
- Do hormone-related cancers emerge in the same cluster due to the Western lifestyle?

To answer this threefold research question, the thesis is organized as follows. Section 2 first characterizes the Western lifestyle, especially with regard to habits and substances attributed to it and that are possibly related to cancer. Then, empirical evidence is presented on the relationship between these habits and substances and cancer. In Section 3, the Finnish cancer incidence and mortality data utilized in the thesis is introduced in more detail. Moreover, the most common cancers in Finland are identified. In Section 4, the clustering methodology, applied later in Section 5, is discussed. Section 6 contains the main contribution to the existing literature, that is, the research question described above is answered. Lastly, in Section 7, the results and conclusions of the thesis are summarized.

2 Characterization of the Western lifestyle

Multiple different features, including materialism, individualism [21], [22], utilitarianism, humanism, liberalism, and democracy [22], are used to characterize the Western lifestyle and culture. In addition, more practical aspects, such as certain dietary choices, are also attributed to the Western lifestyle [23], [24]. In this Section, different lifestyle factors of the Western lifestyle are discussed, in particular from the point of view of substances and habits related to the Western lifestyle that possibly cause cancer. Then, in Sections 2.1.1 - 2.1.7, empirical evidence is reviewed regarding the relationship between the identified substances and habits and cancer.

As mentioned above, one feature of the Western lifestyle is the so-called Western diet, also known as the Western dietary pattern [25]. Compared to the preagricultural era, this diet consists of a large intake of carbohydrate and fat and fewer protein intake at the macronutrient level [24]. In addition, the use of dairy products, refined grains, sugars, and vegetable oils [24], processed foods [25], and salt [24], [25] has increased as part of the Western diet. The increase in the use of these products has resulted in an unbalanced diet containing an excess amount of energy [25], sodium, and omega-6 fatty acids, and an insufficient supply of omega-3 fatty acids, fruits, vegetables, fiber, potassium, as well as other minerals and vitamins [24], [25].

In addition to unbalanced nutrition, the Western diet has been associated with weight gain and obesity [23], [26]. On the other hand, research suggests that the Western diet is not solely responsible for the trend of increasing body weight, but lack of physical activity and sedentary lifestyle, also attributed to Western societies, might promote obesity as well [27]. Furthermore, obesity may lead to a self-perpetuating cycle of less physical activity, thereby lowering the energy consumption of the body and increasing the likelihood of gaining weight [28].

Smoking and alcohol use are two habits that are not unique to the Western societies, but are present throughout the world [29] - [31]. On the other hand, during the early 2000s and 2010s, European countries, for instance, are among those in which smoking has had one of the highest prevalence compared to the global levels, both among females and males [29], [30]. The habit of smoking itself in the Western countries seems to demonstrate a slightly inconsistent trend. That is, in some countries, such as Sweden, Norway, and the United States, smoking prevalence has decreased both among females and males at statistically significant annualized rate from 1980 to 2012, while in other countries, including Austria, France, and Portugal among females, and Serbia and Croatia among males, it has increased at statistically significant rate during the same time period [30]. Similarly to smoking, alcohol use has been the most common in the European and American regions during the early 2000s in terms of total alcohol per capita, with approximately 60% of the population in the respective regions classified as current drinkers in 2019 [31].

The environment in which we live also contains substances whose intake is not necessarily under our control, including microplastics and xenoestrogens. Microplastics, defined as plastic particles of a size of less than 5 millimeters [19], are found not only in items such as seafood, sugar, honey, salt, alcohol, both bottled and tap water, but also in different outdoor and indoor environments, for example in the form of air

we breathe in [32]. While bearing in mind that the list above is not an exhaustive list of microplastic sources, the authors of [32] estimate that due to the consumption of these items alone, an average American adult is exposed to between 98,000 and 121,000 microplastic particles annually, depending on the gender. In addition, xenoestrogens, or foreign chemicals that mimic estrogens [33], have been found in both tap water and food, the latter due to accumulation in the food chain, as some of the xenoestrogens are highly persistent in the environment [34]. Xenoestrogens have become more common in the living environment, especially during the last decades [35] due to their use in certain drugs, industrial products [34], [35], and cosmetics, for example [33], [35].

2.1 Empirical evidence on substances and habits causing cancer

2.1.1 Diet

Table 1 presents items related to the Western diet as well as associated cancer types and the estimated risk of developing a particular cancer. Items whose consumption has either increased or decreased as part of the Western diet are considered in this context. For all items, except omega-6 and omega-3 fatty acids and salt, the associated cancer types and the corresponding effect on the risk of developing cancer have been obtained from the publication of the World Cancer Research Fund/American Institute for Cancer Research (WCRF/AICR) [36]. The scale of risk of developing cancer used in the publication is as follows: 'substantial effect on risk unlikely', 'limited – suggestive increases risk', 'probable increases risk', 'convincing increases risk, limited' – 'suggestive decreases risk', 'probable decreases risk', and 'convincing decreases risk'. Table 1 utilizes the same scale. The associated cancer types and risk effects of omega-6 and omega-3 fatty acids, as well as salt, are obtained from two meta-analyses: the first two from [37] and the latter from [38]. In terms of low-quality evidence, the authors of [37] found that an increased amount omega-3 might slightly increase the risk of developing prostate cancer, but due to the quality of the evidence, this finding is excluded from Table 1.

As somewhat expected, most of the associated cancers are related to the digestive system, such as colorectal, stomach, and pancreatic cancer. However, two items, namely dairy products and processed meat, can be distinguished from the rest of the items in this regard. That is, dairy products are associated with a suggestively increased risk of prostate cancer, while processed meat suggestively increases the risk of developing lung cancer, in addition to a few other cancers. On the other hand, also other items, such as greater intake of fruits and non-starchy vegetables are in part associated with cancers not related to the digestive system, but their effect on the risk of developing cancer is decreasing rather than increasing. It can also be observed, when the items of Table 1 are considered at an overall level, that in most cases the effect on the risk of developing cancer is suggestive rather than probable or convincing.

Item	Associated cancer types	Effect on risk
Dairy products	Colorectum	Probably decreases risk
	Premenopausal breast	Suggestively decreases risk
	Prostate	Suggestively increases risk
Refined grains	None identified	None identified
Sugar sweetened drinks	None identified	None identified
Processed meat	Nasopharynx	Suggestively increases risk
	Oesophagus (squamous cell carcinoma)	Suggestively increases risk
	Lung	Suggestively increases risk
	Stomach	Suggestively increases risk
	Pancreas	Suggestively increases risk
	Colorectum	Convincingly increases risk
Salt	Stomach	Probably increases risk
Omega-6 fatty acids	None identified	None identified
Omega-3 fatty acids	None identified	None identified
Fruits, low intake	Stomach	Suggestively increases risk
	Colorectum	Suggestively increases risk
Fruits, greater intake	Oesophagus (squamous cell carcinoma)	Suggestively decreases risk
	Lung	Suggestively decreases risk
Non-starchy vegetables, low intake	Colorectum	Suggestively increases risk
Non-starchy vegetables, greater intake	Mouth, pharynx, larynx	Suggestively decreases risk
	Nasopharynx	Suggestively decreases risk
	Oesophagus (adenocarcinoma)	Suggestively decreases risk
	Oesophagus (squamous cell carcinoma)	Suggestively decreases risk
	Lung	Suggestively decreases risk
	Premenopausal breast	Suggestively decreases risk
	Postmenopausal breast	Suggestively decreases risk
Fiber	Colorectum	Probably decreases risk

Table 1: Associated cancer types and the effect of items, whose consumption has either increased or decreased as part of the Western diet, on the risk of developing cancer.

2.1.2 Obesity

Table 2, constructed using the publication of the WCRF/AICR [36], displays the associated cancer types and the corresponding effect of adult body fatness on the risk of developing cancer. Although being overweight and obese are not exactly two identical things, Table 2 can be considered as indicative for obesity as well, as obesity refers to a simply higher body mass index than overweight [17]. As Table 2 shows, there is a great variety of cancers associated with adult body fatness, most of them having probably or convincingly increased risk of developing the corresponding cancer. Interestingly, adult body fatness appears to have a probably decreased risk of developing premenopausal breast cancer. The protective effect of adult body fatness against premenopausal breast cancer has been attributed to hormonal changes caused by body fatness [39].

Item	Associated cancer types	Effect on risk
Adult body fatness	Mouth, pharynx, larynx	Probably increases risk
	Oesophagus (adenocarcinoma)	Convincingly increases risk
	Stomach	Probably increases risk
	Pancreas	Convincingly increases risk
	Gallbladder	Probably increases risk
	Liver	Convincingly increases risk
	Colorectum	Convincingly increases risk
	Premenopausal breast	Probably decreases risk
	Postmenopausal breast	Convincingly increases risk
	Ovary	Probably increases risk
	Endometrium	Convincingly increases risk
	Cervix uteri	Suggestively increases risk
	Prostate	Probably increases risk
	Kidney	Convincingly increases risk

Table 2: Associated cancer types and the effect of adult body fatness on the risk of developing cancer.

2.1.3 Physical inactivity

Although physical inactivity has been associated with obesity and vice versa, as discussed above, there are some cancers that have been specifically associated with physical inactivity. Research [40] shows that there is a strong inverse relationship between physical activity and breast cancer, both pre- and postmenopausal, although the relationship is stronger for the latter. A similar, strong inverse relationship has also been observed between physical activity and colorectal cancer [40]. In addition, the author finds the results on the relationship between physical activity and lung cancer, and physical activity and prostate cancer, are inconsistent. In terms of the former relationship, also WCRF/AICR finds that physical activity only suggestively

decreases the risk of lung cancer [36], while in terms of the latter relationship, future research [41] has found the relationship between physical activity and prostate cancer inconclusive too. Some evidence however exists indicating that vigorous rather than moderate activity would be needed to reduce the risk of prostate cancer [40], [41]. Lastly, there is limited evidence that increased physical activity may reduce the risk of ovarian and endometrial cancer [40], as well as adenocarcinoma and squamous cell carcinoma [36]. Nevertheless, an active lifestyle has been estimated to reduce the overall cancer risk by up to 46% [40].

2.1.4 Smoking

Smoking is the most important alterable risk factor for cancer for those who smoke [36]. As Table 3, whose contents are extracted from [42], shows, the cancer risk from smoking coincides especially with the risk of developing lung cancer. That is, on average, smokers are 15-30 times more likely to develop lung cancer than non-smokers. In addition, smokers have a tenfold average risk of developing laryngeal cancer compared to non-smokers, while the average relative risk of developing oral cavity, oro- and hypopharyngeal cancer due to smoking is four- to fivefold. It can also be observed that not all cancer types associated with smoking are related to the respiratory system, but there is also an association with cancers of the pancreas and cervix uteri, as well as myeloid leukemia, for example.

Cancer type	**Average relative risk**
Lung	15-30
Larynx	10
Oral cavity	4-5
Oro- and hypopharynx	4-5
Oesophagus (squamous cell carcinoma or other)	2-5
Pancreas	2-4
Nasal cavity, paranasal sinuses	1.5-2.5
Oesophagus (adenocarcinoma)	1.5-2.5
Nasopharynx	1.5-2.5
Cervix uteri	1.5-2.5
Stomach	1.5-2.0
Liver	1.5-2.0
Kidney	1.5-2.0
Myeloid leukemia	1.5-2.0

Table 3: Associated cancer types and the respective average relative risk of smoking. The risk of oro- and hypopharyngeal and laryngeal cancers corresponds to a range of relative risks after adjusting for alcohol use [42].

2.1.5 Alcohol use

Where the effect of the items part of the Western diet on the risk of developing different cancers is suggestive for the most part, as discussed in Section 2.1.1, the same cannot be said about alcohol. That is, as Table 4, based on the publication of the WCRF/AICR [36], shows, alcohol not only convincingly increases the risk of developing mouth, pharyngeal, laryngeal, squamous cell carcinoma of oesophagus, liver, colorectal, and postmenopausal breast cancer, but also probably the risk of stomach and premenopausal breast cancer, as well as suggestively the risk of lung and pancreatic cancer. In fact, research even suggests that alcohol use is the main source of increased risk of cancers of the head and neck area, and in overall terms, one of the most notable cancer causes in addition to smoking, chronic infections, and obesity [43]. Consequently, the authors of [36] find that any level of alcohol use increases the risk of developing at least some cancer.

Item	Associated cancer types	Effect on risk
Alcohol	Mouth, pharynx, larynx	Convincingly increases risk
	Oesophagus (squamous cell carcinoma)	Convincingly increases risk
	Lung	Suggestively increases risk
	Stomach	Probably increases risk
	Pancreas	Suggestively increases risk
	Liver	Convincingly increases risk
	Colorectum	Convincingly increases risk
	Premenopausal breast	Probably increases risk
	Postmenopausal breast	Convincingly increases risk

Table 4: Associated cancer types and the effect of alcohol use on the risk of developing cancer.

2.1.6 Microplastics

Water and some food items contain microplastics [32]. For humans, microplastics are primarily absorbed through the digestive system [19]. Most of the cancers research has associated with microplastics are also related to the digestive system [19], [44]. These associated cancers are stomach [19], pancreatic [44], colon, and liver cancer [19], [44]. In addition to cancers related to the digestive system, cancers of the skin [44] lung, prostate, and breast have also been associated with microplastics [19], [44]. Because there are multiple cancers associated with microplastics, and due to the great variety of different microplastics, microplastics as a whole may have multiple different roles in developing cancer. Microplastics can, for instance, promote cancer initiation, growth, aggressiveness, immunosuppression, and metastasis [44]. However, as the authors of [44] note, current research on the role of microplastics in cancer development lacks human-based studies and evidence, as most of the studies so far have been executed in in-vitro environments.

2.1.7 Xenoestrogens

In part, xenoestrogens help to understand the association between the items discussed above, such as alcohol and smoking, and cancer. That is, both alcohol and smoking, in addition to other features, yield estrogen-related changes in the human body, and consequently, have been associated with breast and lung cancer, respectively [35]. In addition, the authors of [35] also find xenoestrogens might help explain gender differences in the incidence of certain cancers, including kidney and pancreatic cancers. An increased risk of developing a brain tumor, through transplacental exposure to xenoestrogens, is suggested by the authors as well.

On the other hand, xenoestrogens have been associated with cancers of the reproductive system, including testicular, ovarian [35], and cervical cancer [33]. In particular, the recent and significant increase in the incidence of testicular cancer cannot be explained by smoking, physiological stress, genetic changes, or cryptorchidism, while some studies, although limited in number, have demonstrated a possible relationship between testicular cancer and disturbances in estrogen levels [35]. In addition, in terms of cervical cancer, research suggests that estrogens and xenoestrogens could promote the development of cervical cancer, along with other known risk factors [33].

3 Cancer incidence and mortality data over time in Finland

In this Section, the Finnish cancer incidence and mortality data, obtained from the Finnish Cancer Registry [20], is analyzed. In particular, the most common cancers in Finland in 2022 are identified in Section 3.1. Then, the incidence and mortality of these cancers are studied over time, from 1962 to 2022, in Sections 3.2 and 3.3, respectively. To account for the aging of the Finnish population [45], both incidence and mortality are considered as rates per 100,000 person years instead of absolute figures. Furthermore, the population has been divided into age groups of 10 years, ranging from 20-29 to 70-79 years. Ages below 20 and over 79 years are disregarded, as childhood cancers seldom have lifestyle or environmental causes [46] and aging as such, as described above, is a major risk factor for cancer. Since the focus is more on the effects of the Western lifestyle, these choices are made.

The rates per 100,000 person years and age groups of 10 years are both provided by the Cancer Registry, and are used as such. The cancer statistic application of the Finnish Cancer Registry [20] allows for applying a geographical filter, but no such filter is applied and the data is treated on a country level. All available cancer types are extracted with a few exceptions. Namely, basal cell carcinomas of the skin and genitals, non-invasive neoplasms of cervix uteri, vagina, and vulva, carcinoma in situ of the breast as well as borderline tumor of the ovary are left out of the analysis. These cancer types are excluded as they either represent early stages of the cancer, have low malignant potential, or rarely spread to other parts of the body [47] - [52].

3.1 Most common cancers in Finland by age group and gender

Figure 1 displays the diagnosed cancer cases among females across the different age groups in Finland in 2022. For each age group, diagnosed cancer cases of the 10 most common cancer types are presented separately, while the rest of the cancer types, given that any exists within a certain age group, are classified as 'other' and presented together under this common label.

As a first observation from Figure 1, it is evident that breast cancer is the most common individual cancer type in all age groups, except the age group of 20-29 years, for which melanoma of the skin is the most common cancer type. Second, it appears that the older the age group, the higher the share of 'other' cancer cases compared to any individual cancer type. Correspondingly, for the age group of 70-79 years, 'other' cancers correspond to the most number of diagnoses, while for the other age groups, either breast cancer or melanoma of the skin are the most common. Third, it can also be observed that the older the age group, the greater the overall number of diagnoses. As a consequence of these two observations, it seems older females are merely more likely to get cancer, regardless of the cancer type. If individual cancer types were examined, colon, lung, and tracheal cancers would be the most common after breast cancer in the age groups of 60-69 and 70-79 years. For the younger age groups, 40-49

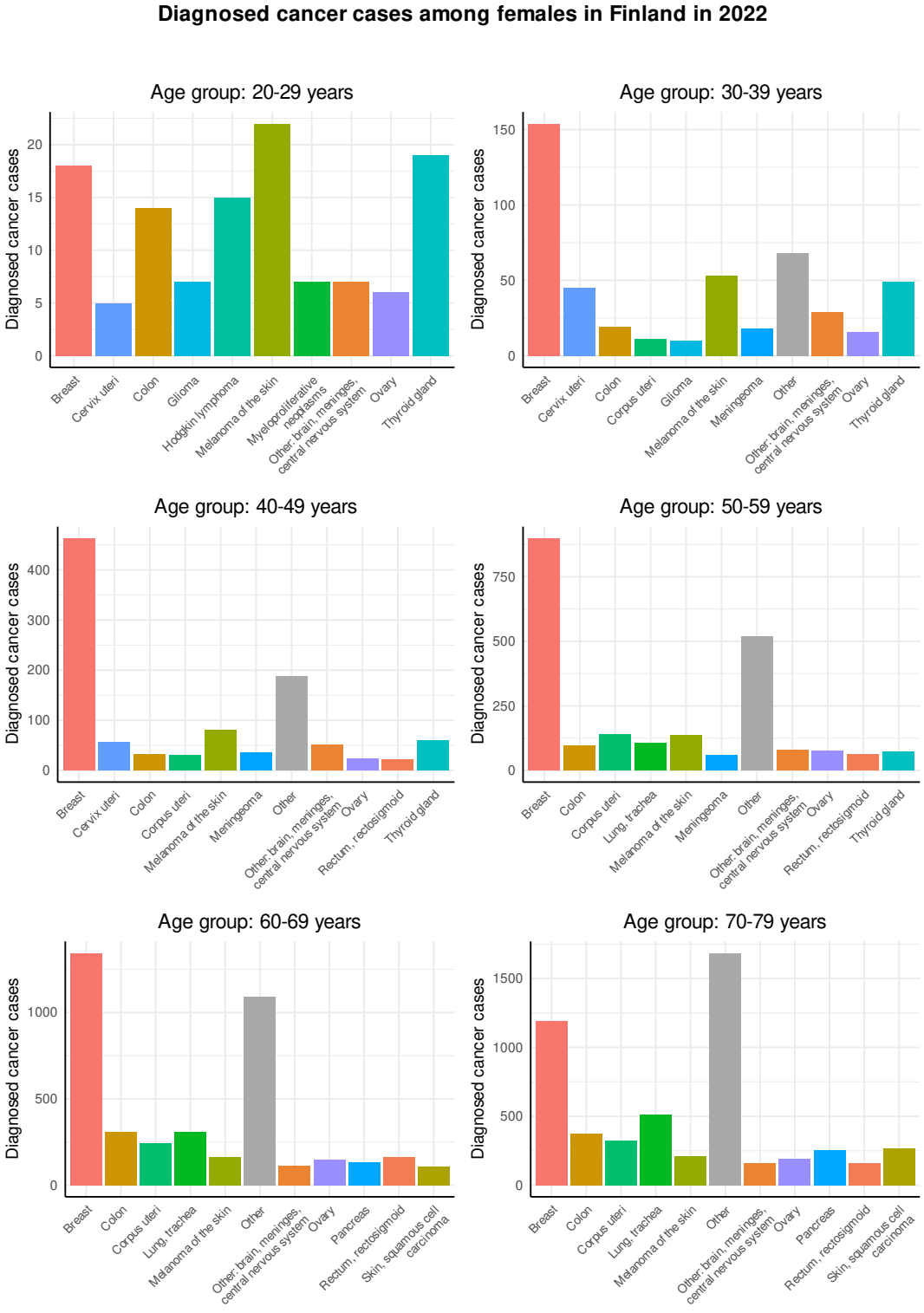


Figure 1: The number of cancer cases among females across different age groups in Finland in 2022.

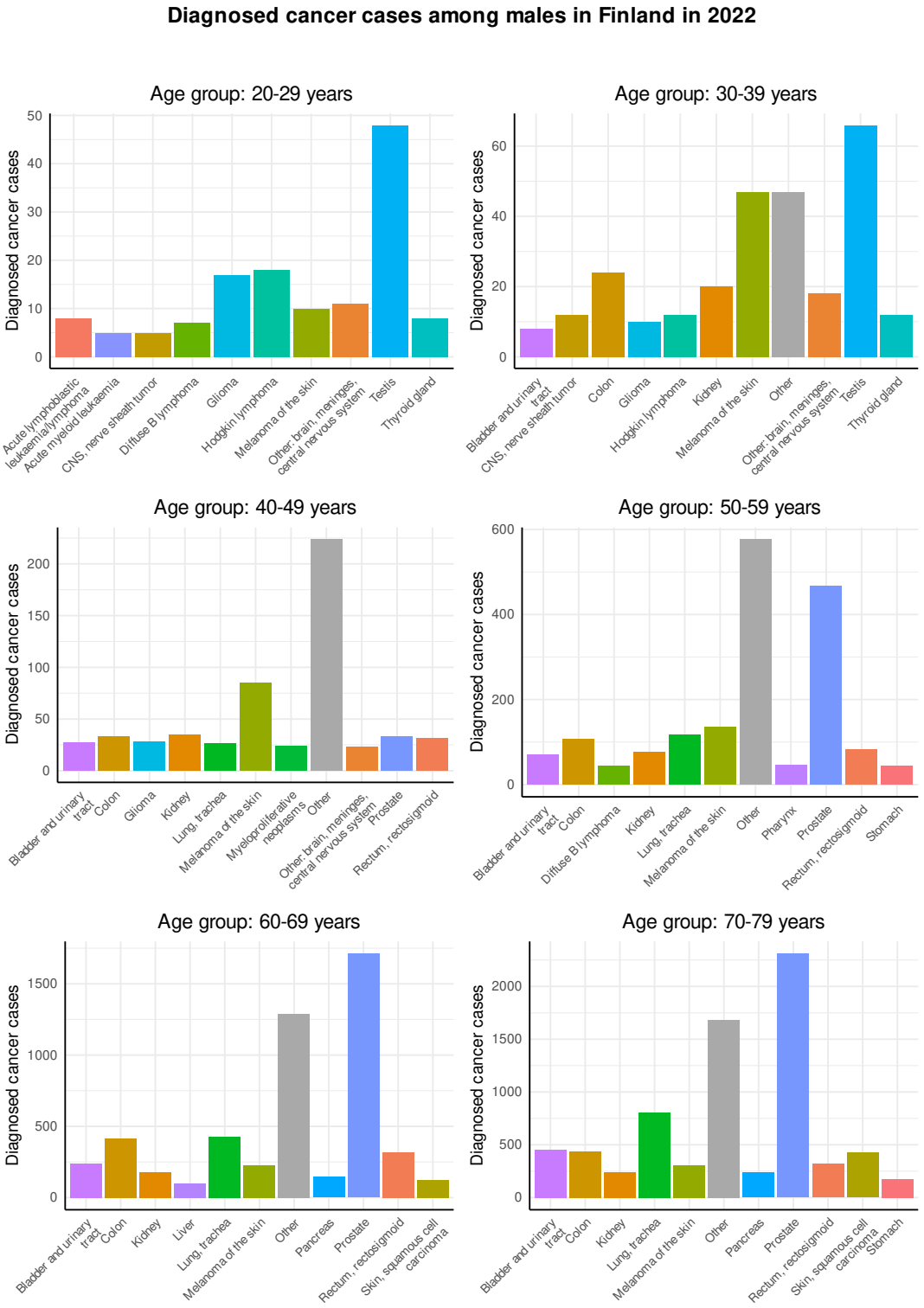


Figure 2: The number of cancer cases among males across different age groups in Finland in 2022.

years and under, thyroid gland cancer and melanoma of the skin are be the most frequent ones in addition to breast cancer.

Figure 2 shows the diagnosed cancer cases among males in Finland in 2022. Most notably, testicular cancer is the most common cancer type in the age groups of 20-29 and 30-39 years, 'other' cancers in the age groups of 40-49 and 50-59 years, and prostate in the age groups of 60-69 and 70-79 years. That is, even though the number of 'other' cancer cases exceeded the number of any individual cancer type in the oldest age group for females, the prostate cancer is such a common cancer type among 70-79-year-old males that the same does not happen. When it comes to the other observations from Figure 2, the following can be noticed. First, similar to females, the older the age group, the greater the overall number of cancer diagnoses. Another similarity between females and males is that colon, lung, and tracheal cancers appear to be relatively common among the age groups of 60-69 and 70-79 years. On the other hand, bladder and urinary tract cancers are also relatively common in these age groups, whereas for females, they did not occur among the 10 most common cancer types in any of the age groups. Additionally, cancers related to the digestive system, including pharynx, stomach, liver, colon, rectum, and rectosigmoid, appear to be on a general level a bit more common among males than females.

3.2 Development of the incidence of the most common cancers over time

Figures 3 and 4 visualize the incidence rate per 100,000 person years of the most common cancers among females in Finland, identified in Section 3.1, as time series from 1962 to 2022. Visualizations are done separately for different age groups. In all age groups, except for the age group of 20-29 years, breast cancer stands out from the other most common cancer types and has a distinctly higher incidence rate than the other cancers during the period examined. Moreover, in all age groups, including the age group of 20-29 years, the incidence rate of breast cancer has increased somewhat linearly from 1962 to 2022. There are, however, a couple of exceptions to this linear trend, one of them being roughly constant rate of 150 incidences per 100,000 person years in the age group of 40-49 years from early 1990s to 2022. Additionally, in the age group of 50-59 years, there appears to be a higher incidence rate increment in the mid-1980s than what would be compatible with the linear trend. Start of the national breast cancer screening program around the same time [52] might explain the increment. Lastly, in all of the age groups of 40-49 years and older, there seems to be a slightly downward pointing trend within the last 7-8 years of the investigated time period. This can be a true phenomenon due to multiple possible reasons or a matter of how carefully breast cancer is screened and diagnosed.

There are also other observations to be made beyond breast cancer. To begin with, in the age group of 20-29 years, the incidence rate per 100,000 person years of melanoma of the skin increased sharply in the mid-2010s. Also, the rate of thyroid gland cancer in this age group appears to have alternately increased and decreased over intervals of approximately 5 years. On the other hand, given that the overall scale

Incidence rate per 100,000 person years of the most 10 common cancers among females in Finland from 1962 to 2022: age groups 20-29, 30-39, and 40-49 years

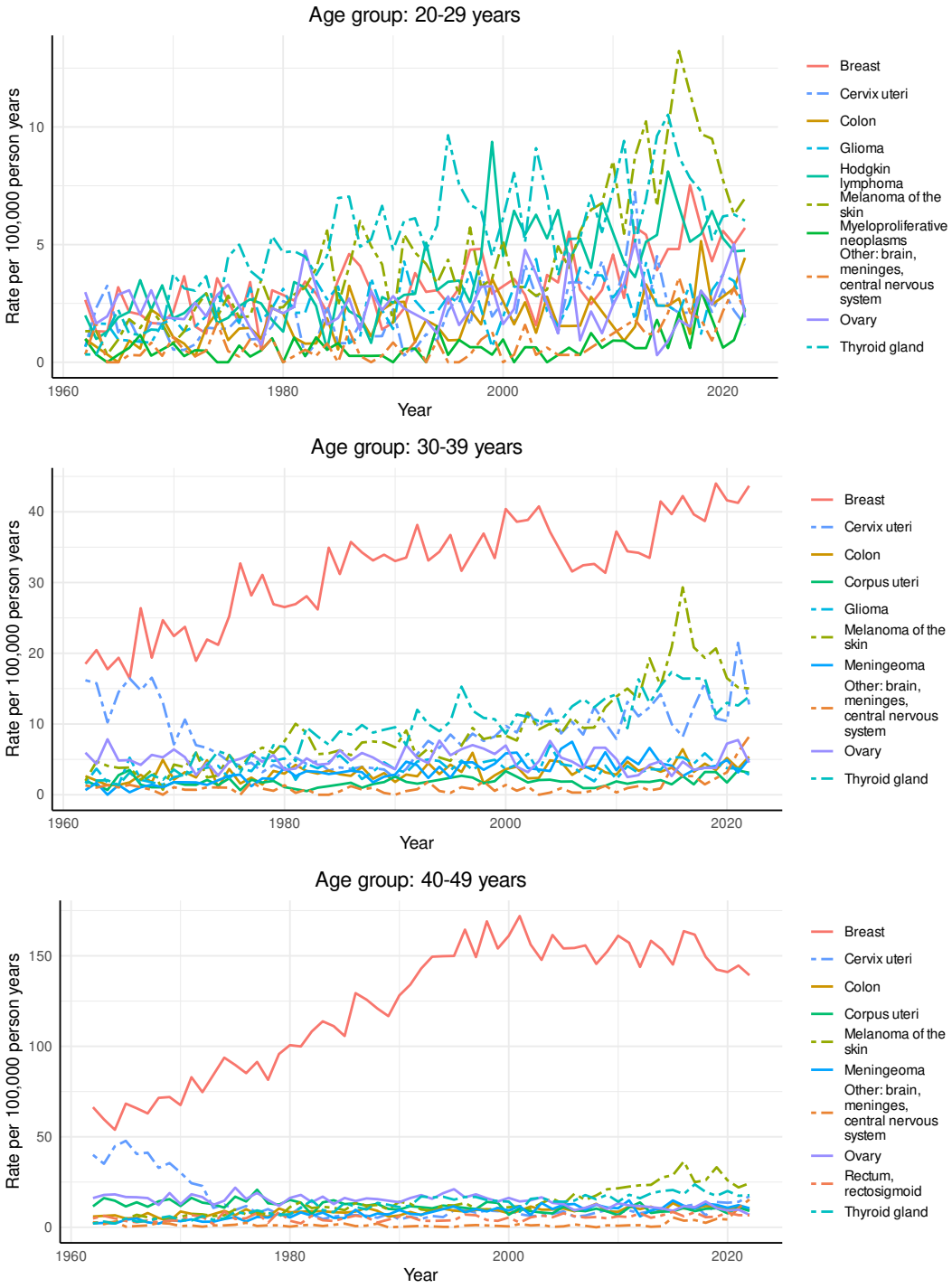


Figure 3: Incidence rate per 100,000 person years of the most common cancers among females in Finland from 1962 to 2022 for the age groups 20-29, 30-39, and 40-49 years.

Incidence rate per 100,000 person years of the most 10 common cancers among females in Finland from 1962 to 2022: age groups 50-59, 60-69, and 70-79 years

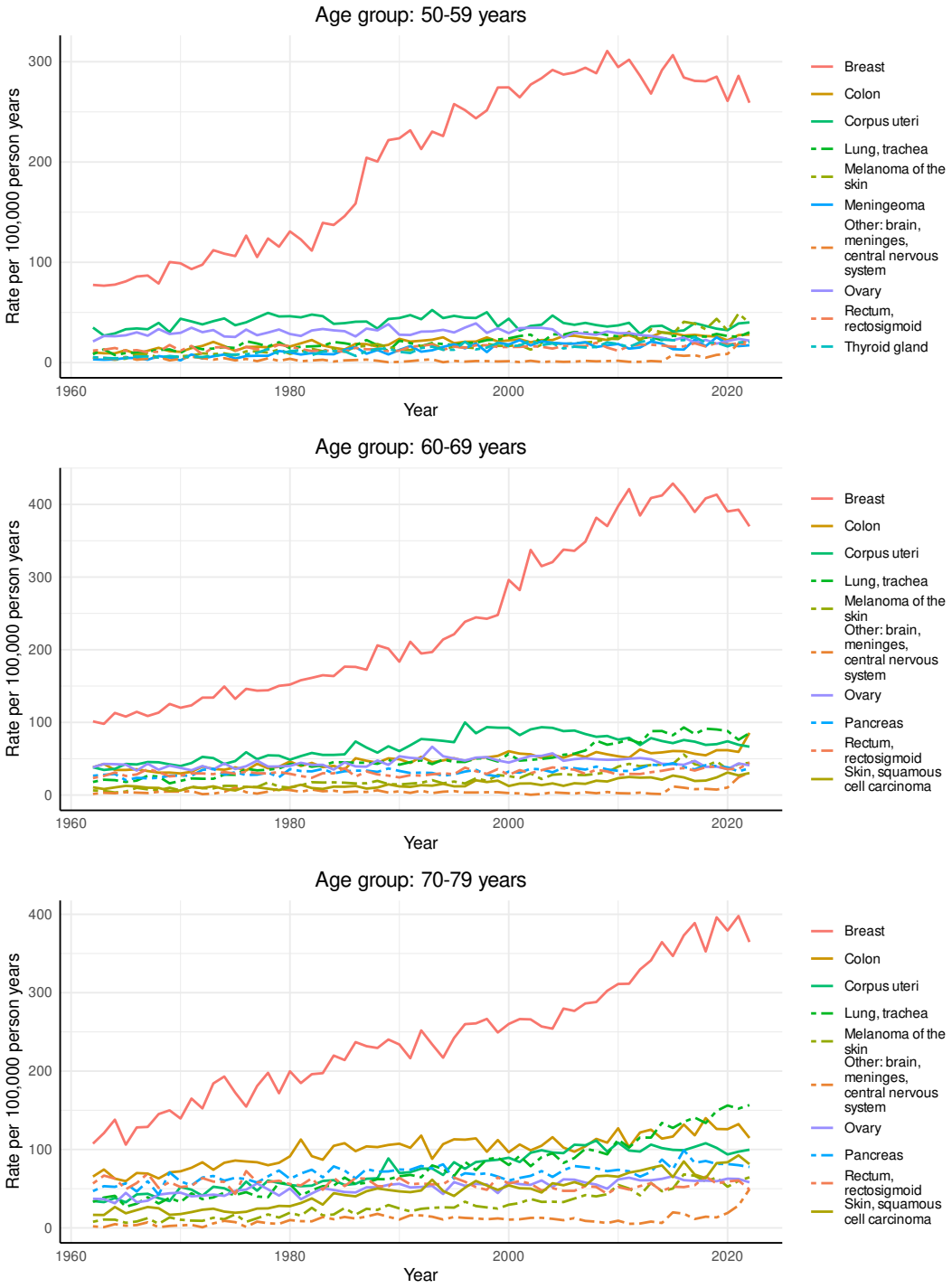


Figure 4: Incidence rate per 100,000 person years of the most common cancers among females in Finland from 1962 to 2022 for the age groups 50-59, 60-69, and 70-79 years.

of the y-axis is relatively small, and correspondingly the scale of the changes, the sharp increases and decreases one can observe are most likely plain randomness. The 10-year moving average graph of the same time series, shown by Figure A1 presented in Appendix A, confirms that the underlying trend seems to be linearly increasing instead of fluctuating. In the age group of 30-39 years, there is a much larger increase in the incidence rate of melanoma of the skin around the same time in the mid-2010s. Furthermore, since the same kind of increase can also be observed in the age groups of 40-49 and 50-59 years as well as in the 10-year moving average graphs displayed by Figures A1 and A2 presented in Appendix A, there might be some underlying trend that has caused the rates of melanoma of the skin to increase during the last 10 years of the period examined.

In addition to melanoma of the skin and breast cancer, there are two other cancer types that have interesting incidence rates over time. The first is the cancer of the cervix uteri, whose incidence rate clearly decreased in the age groups of 30-39 and 40-49 years during the 1960s and early 1970s, thanks to a nationwide screening program [53], and has remained relatively constant since then in the age group of 40-49 years. However, in the age group of 30-39 years, the incidence rate appears to have increased since the 1990s. The second is lung and tracheal cancer, whose incidence rate has increased in the age groups of 60-69 and 70-79 years, becoming the third and second most common cancer in terms of the incidence rate in these age groups, respectively. Interestingly, in the age group of 70-79 years, almost all other cancer types appear to obey a linearly increasing trend as well, while in the other groups, incidence rates of most of the cancers seem to be roughly constant during the time period studied.

The development of the incidence rate per 100,000 person years of the most common cancers among males in Finland from 1962 to 2022 is shown by Figures 5 and 6, and the corresponding 10-year moving average graphs by Figures A3 and A4 presented in Appendix A. Similarly to females, the most common cancers referred to here are the cancers identified in Section 3.1. Most notably, in the age groups of 20-29 and 30-39 years, the incidence rate of testicular cancer increased considerably in the very late 1990s in the first group and early 2000s in the latter group. As for the other age groups, instead of testis, the incidence rate of prostate cancer increased around the same time, most clearly in the age groups of 50-59, 60-69, and 70-79 years. In the latter two age groups, the increase occurred even a bit earlier, around the early 1990s. A third cancer type, whose incidence rate has increased during the investigated period, appears to be melanoma of the skin, especially in the age groups from 30-39 to 50-59 years.

Unlike the increases in the incidence rates of melanoma of the skin, testicular, and prostate cancer, there have also been decreases in the incidence rates of some other cancers, such as lung, trachea, and stomach. In particular, in the age groups of 40-49 years and older, the cancer of the lung and trachea has decreased, from being clearly the most common or the second most common cancer type in terms of the incidence rate per 100,000 person years in the 1960s to not considerably, if at all, more common than the other 10 most common cancers. In addition, in the age groups of 50-59 and 70-79 years, the incidence rate of stomach cancer has also decreased at the same time as the cancer of the lung and trachea, and to an even lower rate.

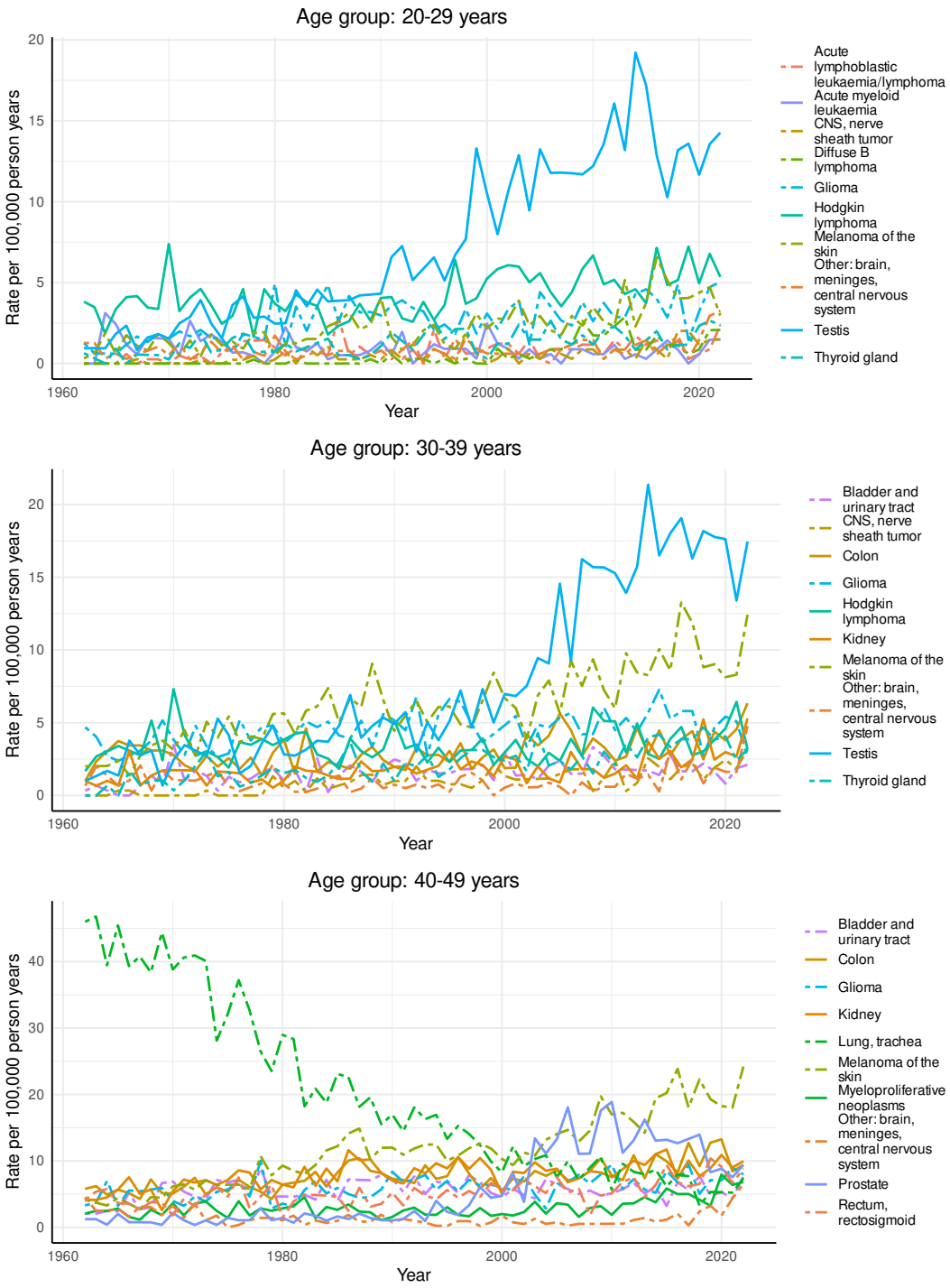


Figure 5: Incidence rate per 100,000 person years of the most common cancers among males in Finland from 1962 to 2022 for the age groups 20-29, 30-39, and 40-49 years.

Figure 6: Incidence rate per 100,000 person years of the most common cancers among males in Finland from 1962 to 2022 for the age groups 50-59, 60-69, and 70-79 years.

3.3 Development of the mortality of the most common cancers over time

After studying the incidence rates of the most common cancers, a similar kind of analysis is done for the same cancers in terms of mortality per 100,000 person years. Figures 7 and 8 illustrate the mortality per 100,000 person years of the most common cancers among females in Finland as time series from 1962 to 2022. Again, the corresponding 10-year moving average graphs are included into the Appendix A. As somewhat expected, based on the relatively high incidence rates displayed by Figures 3 and 4, breast cancer corresponds to one of the highest mortality per 100,000 person years in all age groups and throughout the studied time period. The difference in mortality per 100,000 person years compared to other cancer types, however, is smaller than one might expect based on the difference between the incidence rate of breast cancer and other cancers. It turns out that the aim of the national breast cancer screening program is to detect cancer as early as possible, even so that small early-stage tumors that would not have caused symptoms during a person's lifetime are also diagnosed [52]. On the other hand, early diagnoses allow for more effective treatment [54] and therefore reduced mortality [52], [54].

In terms of other cancer types than breast cancer, few observations arise. First, in the age groups 50-59, 60-69, and 70-79 years, lung and tracheal cancer appears to have a relatively high mortality, in particular starting from the early 2000s. Moreover, in the latter two age groups, the mortality has been around the same or even surpassed that of breast cancer from 2010 onward. On the other hand, as Figure 4 showed, the incidence rate of lung and tracheal cancer has increased in these age groups, too, but not even close to the incidence rate of breast cancer. Similarly, the mortality from pancreatic cancer appears rather high, given that the incidence of the cancer was not particularly high. Therefore, these two observations suggest that lung, tracheal, and pancreatic cancers are relatively deadly. This hypothesis is supported by the fact that lung cancer is one of the three cancers causing the most deaths globally, both among females and males [8], and research describing pancreatic cancer as difficult to diagnose until an advanced stage, thus limiting possible treatment options and, furthermore, increasing mortality [55].

To conclude this Section, the mortality per 100,000 person years of the most common cancers among males in Finland is studied over time. The corresponding time series, spanning from 1962 to 2022, are visualized by Figures 9 and 10. For most of the examined period, testicular cancer is surprisingly not the cancer type having the highest mortality in the age groups of 20-29 years and 30-39 years, but thyroid gland cancer is. The surprise comes from the fact that testicular cancer has a relatively high incidence rate in these age groups, as Figure 5 illustrated. The same kind of observation applies to prostate cancer as well. That is, in terms of the incidence rate, it is having undeniably the highest rate in the age groups from 50-59 to 70-79 years from the late 1990s onward, but the mortality per 100,000 person years of the cancer does not appear to be high at all in the age groups of 50-59 and 60-69 years, and moderately high in the age group of 70-79 years. These observations could, for example, imply that testicular and prostate cancers are not as dangerous as other cancers, or are relatively

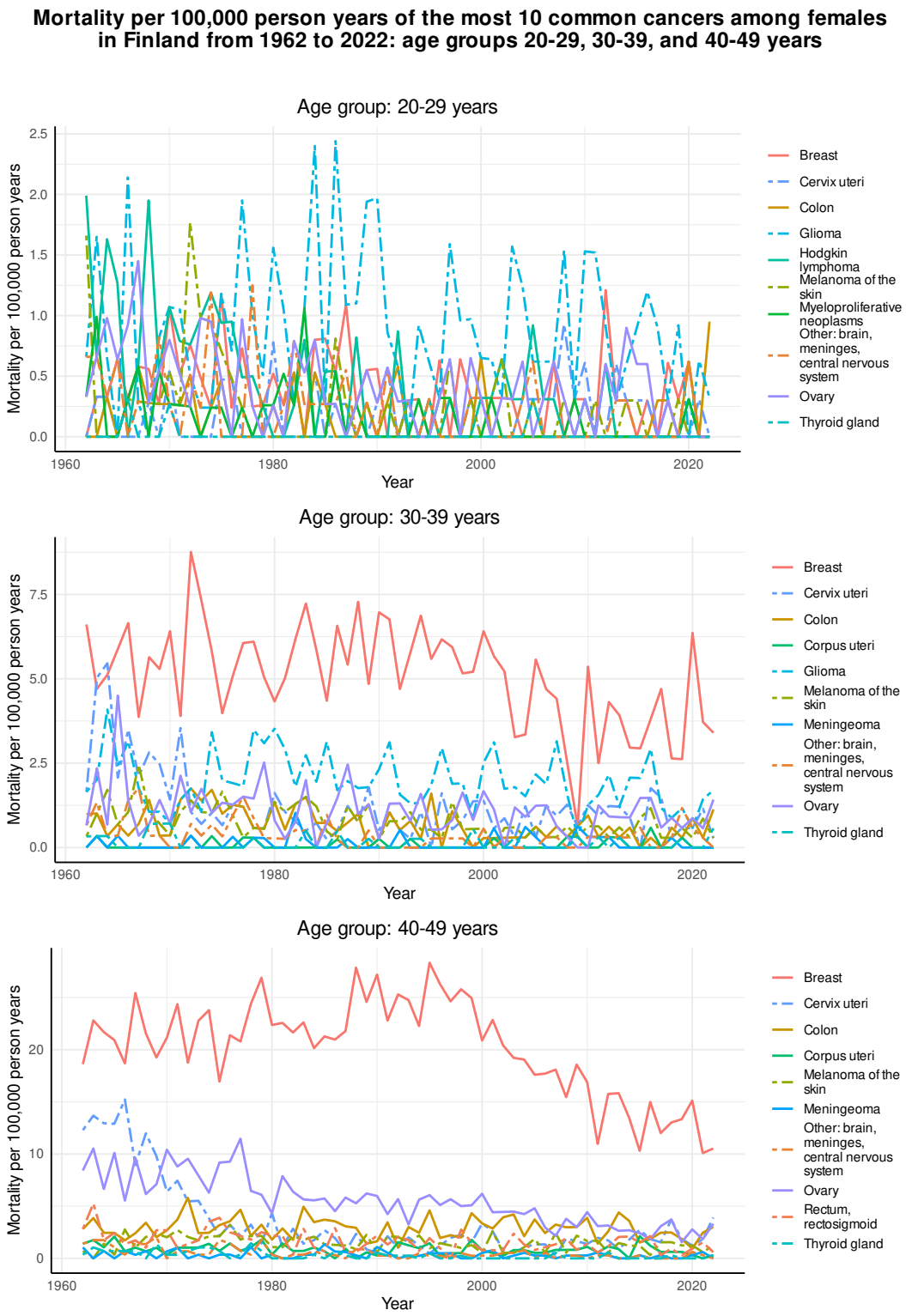


Figure 7: Mortality per 100,000 person years of the most common cancers among females in Finland from 1962 to 2022 for the age groups 20-29, 30-39, and 40-49 years.

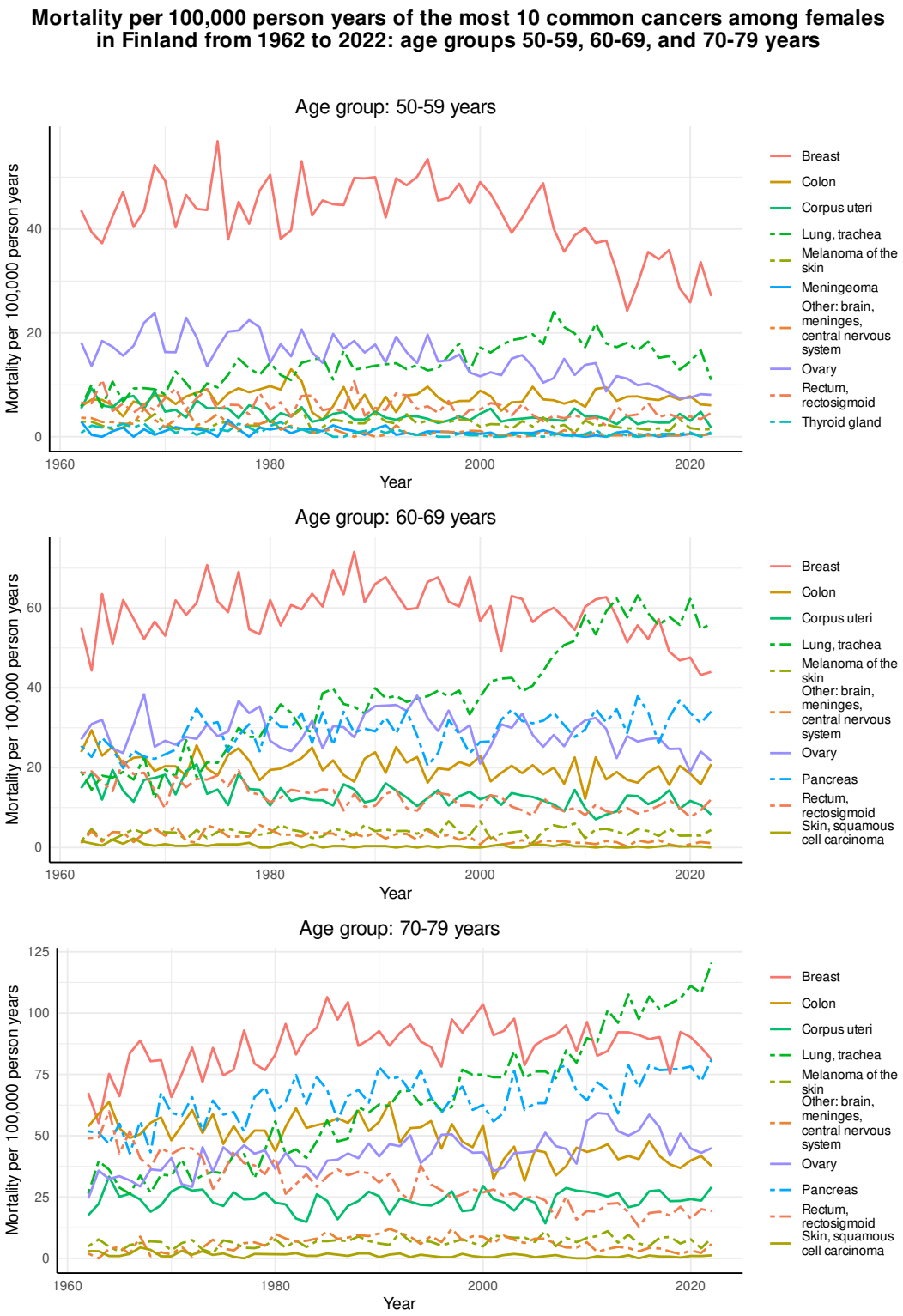


Figure 8: Mortality per 100,000 person years of the most common cancers among females in Finland from 1962 to 2022 for the age groups 50-59, 60-69, and 70-79 years.

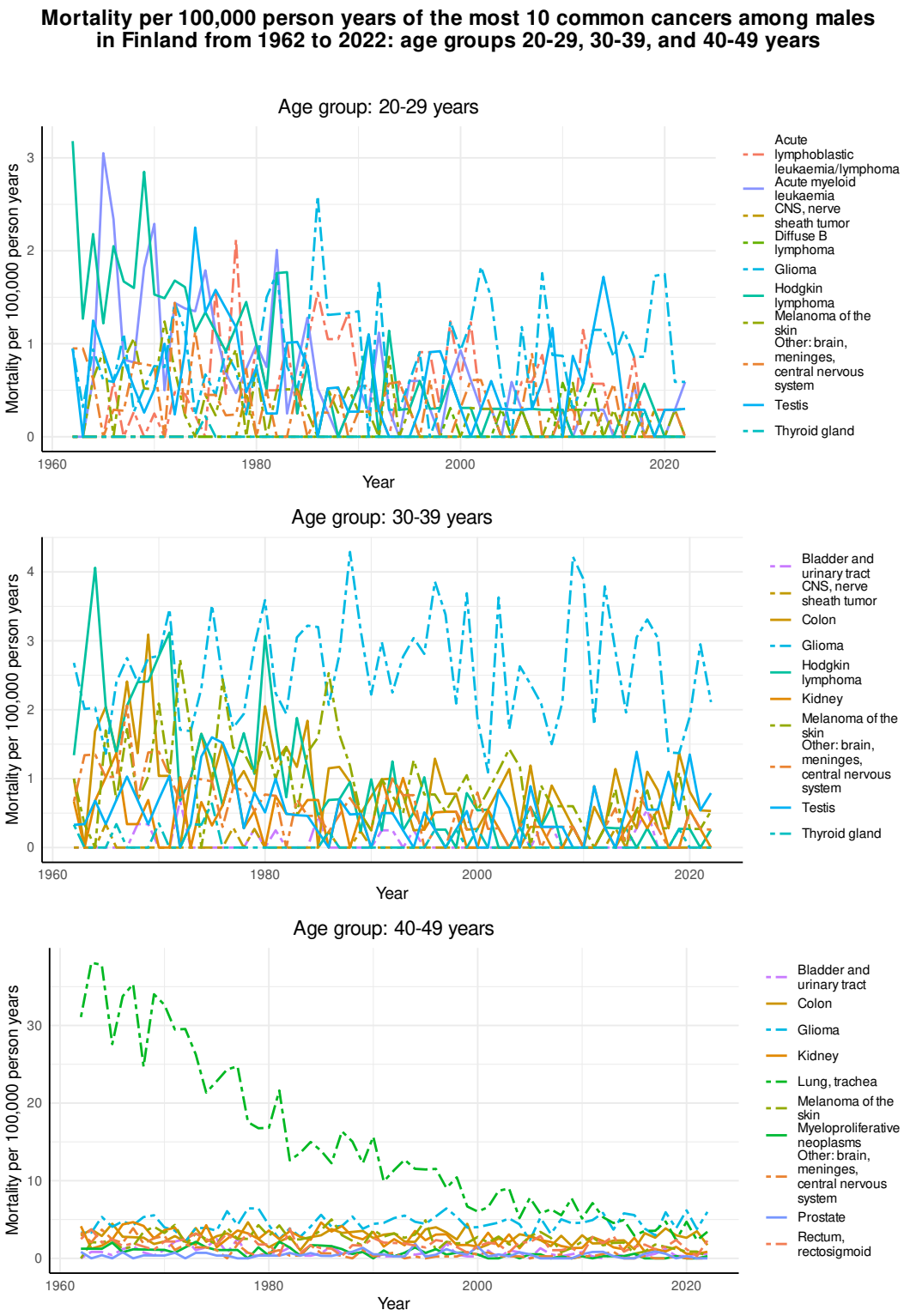


Figure 9: Mortality per 100,000 person years of the most common cancers among males in Finland from 1962 to 2022 for age the groups 20-29, 30-39, and 40-49 years.

Figure 10: Mortality per 100,000 person years of the most common cancers among males in Finland from 1962 to 2022 for the age groups 50-59, 60-69, and 70-79 years.

easier to treat.

Finally, it can be observed that cancer of the lung and trachea is relatively deadly among males, similar to females. This is visible especially in the age groups from 40-49 to 70-79 years. It is still worth keeping in mind that the incidence rate of this cancer is also high in these age groups, and the shape of the mortality time series reminds of that of the incidence rate of the cancer in the respective age groups. Another thing common with females is that pancreatic cancer also seems to correspond to relatively high mortality per 100,000 person years in the age groups of 60-69 and 70-79 years, although it is not that common in terms of the incidence rate.

4 Clustering of functional data

One of the research questions of the thesis reads: "What kind of cluster structures can be identified from the cancer incidence and mortality data over time in Finland?". A valid question is, however, what is an appropriate method for finding these cluster structures, and what is even meant with clustering in the first place. The latter topic is discussed in Section 4.1, while the former is the subject of Section 4.2.

4.1 Basics of clustering

When it comes to identifying groups of objects, or clusters, in a multivariate dataset, two different cases may emerge [56]. In some cases, there is an interest to determine whether some natural clusters exist in a given dataset, whereas in the other cases the interest is to allocate objects into a set of pre-existing groups [56]. The latter is also known as classification [57]. Cluster analysis, on the other hand, refers to the former case, that is, the set of tools for identifying the natural groups of objects from the dataset [56]. With the help of these tools, the fundamental objective is to identify clusters that exhibit maximal within-group homogeneity, while simultaneously ensuring that the disparities between the separate groups are maximized as well [56], [57].

To achieve this objective, cluster analysis is composed of two central steps: the choice of the proximity measure and the choice of the group-building algorithm [56]. As the authors express, the proximity measure is needed to quantify the proximity, or distance, between each pair of objects in the dataset. Furthermore, the closer the objects are in terms of the chosen proximity measure, the more homogeneous they are. In practice, there are many different proximity measures, and the nature of the data guides the choice of an appropriate proximity measure [56], [58]. For example, for continuous data, L_p-distances, $p \geq 1$, induced by the respective L_p-norms, are a common choice [56]. However, it should be noted that the use of L_p-distances assumes the same scale among the variables, and in the absence of this, the data should be standardized before applying the proximity measure [56].

Once the pairwise distances between objects in the dataset have been computed using the chosen proximity measure, a subsequent task is to assign the objects to groups in a way that maximizes the between-cluster distances and minimizes the pairwise within-cluster distances [56]. There are two main types of group-building, or clustering, algorithms to perform this task: hierarchical and partitioning algorithms [56]. Hierarchical algorithms can be split to divisive and agglomerative techniques [57], [58] . Especially the latter, agglomerative hierarchical algorithms, are widely used in practice [56]. These start from the most fine-grained partition possible, that is, each object forms its own cluster, compute the between-cluster distances, and then join the two closest clusters into one [56]. This process is repeated until every object in the dataset is grouped in a single cluster. For the divisive hierarchical algorithms, the process is the opposite; that is, they start from one large cluster containing all objects and then split the groups apart, forming a finer partition of the dataset after each step [56].

The main difference between hierarchical and partitioning algorithms is how permanent the assignment of an object to a certain group is [56]. That is, in the context of hierarchical algorithms, the assignment is always permanent, whereas in partitioning techniques, the assignment might change during the clustering procedure. Furthermore, partitioning algorithms utilize an initial clustering, after which the algorithm aims to reach a certain score by exchanging objects between groups [56]. For example, the k-means clustering algorithm aims to partition the dataset into k clusters by minimizing the within-cluster sum of squared distances, which are centered around the respective cluster means [57].

4.2 Shape sensitive clustering of functional data

Functional Data Analysis (FDA) concerns the analysis and theory of data that may be regarded as a collection of observed continuous functions, including curves, images, and shapes [57]. Since the Finnish cancer incidence and mortality data can be viewed as such a collection, incorporating some FDA theory into the thesis is reasonable. In particular, when it comes to the choice of the clustering method, the contributions of the FDA community should be integrated with those of the clustering community to ensure as appropriate an outcome as possible.

To support this integration and, more broadly, connect the two research communities, a review article on clustering methods for functional data was published in 2023 [59]. Most importantly, the article provides a comprehensive and structured review of clustering techniques for functional data, including a three-tier categorization of the existing clustering techniques. The three tiers of the categorization are based on the dimensionality of the input, the type of the clustering algorithm, and the application strategy of the chosen algorithm [59]. In addition to providing the categorization, the authors emphasize that further advances in the field could be achieved by customizing state-of-the-art clustering algorithms to specific use cases, including curve smoothing and registration where appropriate. They also mention that there is rarely sufficient justification behind the choice of a particular clustering algorithm.

Therefore, to arrive at a customized algorithm for the clustering problem at hand and also to provide some justification for the chosen algorithm, the three-tier categorization [59] is used as a systematic framework to guide the choice. First, in terms of the first tier of the categorization, that is, dimensionality of the input, the question is whether to use a finite- or infinite-dimensional input. As such, the cancer incidence and mortality data represent discrete, finite-dimensional observations of the underlying continuous functions corresponding to cancer incidence and mortality. Thus, one could either try to estimate these underlying functions through curve smoothing or simply work with the discrete observations. For the purposes of the thesis, the latter is chosen for two reasons. First, it is believable that the broader trends behind cancer incidence and mortality are observable even without curve smoothing and despite possible random fluctuations. Second, some fluctuations may reflect real phenomena that affect cancer incidence and mortality, which can be relevant from the point of view of the thesis.

In terms of the second tier of the categorization, many different clustering algorithms for finite-dimensional functional data are employed in the literature, including centroid-

based clustering, hierarchical clustering, and spectral clustering [59]. Since the number of clusters for the cancer incidence and mortality data is not necessarily known beforehand, centroid-based clustering algorithms, such as k-means, do not appear to be a reasonable choice. Additionally, the authors of [59] also demonstrate that the agglomerative hierarchical clustering algorithm, combined with an appropriate linkage method, outperforms not only k-means but also density-based clustering, subspace clustering, and the Gaussian mixture model in their experiment. The linkage method dictates which clusters should be merged as the algorithm proceeds, for example, should the merging be made based on the minimum, average, or maximum distance between objects in different clusters (single, average, and complete linkage, respectively) [58]. The average linkage appears to perform best when one or two large clusters are expected in the resulting clustering structure and the data is non-periodic, while Ward's method is suggested for periodical data [58]. As no periodicity is expected among the cancer incidence and mortality data, average linkage is the preferred choice in this context.

The third tier of the categorization [59] concerns the application strategy of the chosen clustering algorithm. This means, in practice, how to address possible phase and/or amplitude variations in the data. Accounting for phase variations becomes important if the overall shape of the curves matters more than the exact timing of the amplitude changes [59]. In the literature, the technique of aligning curves in different phases is known as curve registration, curve alignment, or time warping [59]. However, it can be argued that curve registration is not relevant in the context of the thesis. The reason is simply that the timing of the amplitude changes in cancer incidence and mortality data is important. For example, if the amount of a carcinogen increased in the environment and the incidence of some cancers subsequently increased, these cancers should be assigned to the same cluster. In contrast, if curve registration was performed, this kind of information could be lost.

Having conceptually justified the choice of agglomerative hierarchical clustering algorithm with average linkage, a detailed mathematical formulation follows. First, a proximity measure shall be defined that determines the distance between two curves on the basis of their overall shape, considering both phase and amplitude. To construct this measure, two components are needed: a one-dimensional distance measure $d(\cdot,\cdot)$ that quantifies the pointwise distance between two curves, and a one-dimensional location measure $T(\cdot)$ that captures the baseline incidence rate or mortality of a given cancer. Combining these two components in a meaningful way results in the following definition.

Definition 4.2.1. Let $X(t)$ and $Y(t)$ be continuous functions over the time interval $[T_0, T_K]$. Then the proximity measure $D(X,Y)$ between X and Y is defined as:

$$D(X,Y) = \frac{1}{T_K - T_0} \int_{T_0}^{T_K} (d(X(t),Y(t)) - d(T(X),T(Y)))dt, \tag{1}$$

where $d(\cdot,\cdot)$ is some one-dimensional distance measure and $T(\cdot)$ is some one-dimensional location measure.

The proximity measure (1) captures the overall similarity in shape between two curves by comparing the pointwise deviations, while excluding the effect of possible differences in the baseline rates. Furthermore, the term $1/(T_K - T_0)$ is used to normalize the measure by taking into account the length of the given time interval.

In the case of the Finnish cancer incidence and mortality data, the underlying continuous functions representing incidence and mortality over time are not directly observed. Instead, the functional dataset at hand is a collection of discrete yearly observations sampled at equal intervals. Therefore, a discrete approximation of (1) shall be formulated. This can be achieved by utilizing the Riemann sum approximation [60]:

$$\int_a^b f(x)dx \approx \sum_i^n f(x_i)\Delta x, \tag{2}$$

where $\Delta x = (b-a)/n$ and n is large enough. Assuming the number of observations, k, of the yearly cancer incidence and mortality rates is large enough, (2) can be applied to (1), yielding the following:

$$\begin{aligned} D(X,Y) &= \frac{1}{T_K - T_0}\int_{T_0}^{T_K}(d(X(t),Y(t)) - d(T(X),T(Y)))dt \\ &\approx \frac{1}{T_K - T_0}\sum_{i=1}^{k}(d(x_i,y_i) - d(T(X),T(Y)))\frac{T_K - T_0}{k} \\ &= \frac{1}{k}\sum_{i=1}^{k}(d(x_i,y_i) - d(T(X),T(Y))) \end{aligned} \tag{3}$$

where, $X = [x_1, x_2, ..., x_k]$ and $Y = [y_1, y_2, ..., y_k]$ represent collections of k yearly observations of either cancer incidence or mortality rate.

So far $d(\cdot,\cdot)$ and $T(\cdot)$ have been treated as some general distance and location measure, respectively. Two possible choices for both are presented next.

Example 4.2.2. Let $d(x_i, y_i) = |x_i - y_i|$, for $i = 1, ..., k$, and $T(X) = \text{Med}(X)$, where

$$\text{Med}(X) = \begin{cases} x_{(k+1)/2} & \text{if } k \text{ is odd} \\ \frac{x_{(k/2)} + x_{(k+1)/2}}{2} & \text{if } k \text{ is even} \end{cases} \tag{4}$$

is the median of the k observations in the collection X [56]. Then the discrete proximity measure (3) becomes:

$$D(X,Y) = \frac{1}{k}\sum_{i=1}^{k}(|x_i - y_i| - |\text{Med}(X) - \text{Med}(Y)|).$$

Example 4.2.3. Let $d(x_i, y_i) = (x_i - y_i)^2$, for $i = 1, ..., k$, and $T(X) = \bar{x}$, where $\bar{x} = (\sum_{i=1}^{k} x_i)/k$ is the mean of the k observations. Then the discrete proximity measure (3) becomes:

$$D(X,Y) = \frac{1}{k}\sum_{i=1}^{k}((x_i - y_i)^2 - (\overline{x} - \overline{y})^2).$$

The choices for $d(\cdot,\cdot)$ and $T(\cdot)$ presented in Example 4.2.3 are preferred over those presented in Example 4.2.2 for two reasons. First, choosing $d(x_i, y_i) = (x_i - y_i)^2$ instead of $d(x_i, y_i) = |x_i - y_i|$ helps to discern large differences in the shapes of the two curves. Second, choosing $T(X) = \overline{x}$ instead of $T(X) = \text{Med}(X)$ helps to capture information along all points in the curve and thereby its overall shape, while the median only considers an individual data point.

After choosing $d(x_i, y_i) = (x_i - y_i)^2$, for $i = 1, ..., k$, and $T(X) = \overline{x}$, an important property of the proximity measure (3) can be shown.

Lemma 4.2.4. Let $d(x_i, y_i) = (x_i - y_i)^2$, for $i = 1, ..., k$, and $T(X) = \overline{x}$, where $\overline{x} = (\sum_{i=1}^{k} x_i)/k$ is the mean of the k observations. Also, let $Y = [y_1, ..., y_k]$, where $y_i = x_i + b$ for $i = 1, ..., k$, and $b \neq 0$ is constant. Then $D(X,Y) = 0$.

Proof. It directly follows that:

$$\begin{aligned}
D(X,Y) &= \frac{1}{k}\sum_{i=1}^{k}(d(x_i, y_i) - d(T(X), T(Y))) \\
&= \frac{1}{k}\sum_{i=1}^{k}((x_i - y_i)^2 - (\overline{x} - \overline{y})^2) \\
&= \frac{1}{k}\sum_{i=1}^{k}((x_i - (x_i + b))^2 - (\overline{x} - (\overline{x} + b))^2) \\
&= \frac{1}{k}\sum_{i=1}^{k}(b^2 - b^2) \\
&= 0.
\end{aligned}$$

□

The practical implication of Lemma 4.2.4 is that the proximity measure (3) is zero when two curves are identical, up to vertical shift by a constant $b \neq 0$.

One might also want to consider scaling the variables x_i, y_i, for $i = 1, ..., k$. The scaling can be performed through the following standardization process:

$$\hat{x}_i = \frac{x_i - T(X)}{\sqrt{s(X)}}, \tag{5}$$

where $\hat{x}_i$ is the standardized variable, $T(X)$ is a one-dimensional location measure, and $s(X)$ is a one-dimensional scatter measure. In practice, there are multiple possible choices for $T(X)$ and $s(X)$, two of which are presented next.

Example 4.2.5. Let $T(X) = \overline{x}$, where $\overline{x} = (\sum_{i=1}^{k} x_i)/k$ is the mean of the k observations, and let $s(X) = (\sum_{i=1}^{k}(x_i - \overline{x})^2)/k$ be the population variance. Then the standardization formula (5) becomes:

$$\hat{x}_i = \frac{x_i - \frac{1}{k}\sum_{i=1}^{k} x_i}{\sqrt{\frac{1}{k}\sum_{i=1}^{k}(x_i - \overline{x})^2}}, \text{ for } i = 1, ..., k.$$

Example 4.2.6. Let $T(X) = \text{Med}(X)$, where $\text{Med}(X)$ is the median of the k observations (4), and let $s(X) = \text{MAD}(X)$, where

$$\text{MAD}(X) = \text{Med}(|x_i - \text{Med}(X)|, ..., |x_k - \text{Med}(X)|)$$

is the Median Absolute Deviation (MAD) [61]. Then the standardization formula (5) becomes:

$$\hat{x}_i = \frac{x_i - \text{Med}(X)}{\sqrt{\text{MAD}(X)}} \text{ for } i = 1, ..., k.$$

Due to the aforementioned concern regarding the use of median in the context of the thesis, and the median-based measures for both location and scatter in Example 4.2.6, the use of mean as the location measure and variance as the scatter measure as presented in Example 4.2.5 appears to be a more appealing choice when it comes to standardizing the variables.

The standardization formula (5) can be applied to both x_i and y_i, for $i = 1, ..., k$. The effect of standardization is subject to testing as, on the one hand, the scale of the variables should affect the resulting clustering structure, but on the other hand, the shape of the curve should still be the primary factor guiding the clustering process, independent of the scale. In addition, standardization of the variables is required for the following property of the proximity measure (3).

Lemma 4.2.7. Let $d(x_i, y_i) = (x_i - y_i)^2$, for $i = 1, ..., k$, $T(X) = \overline{x}$, where $\overline{x} = (\sum_{i=1}^{k} x_i)/k$, and $s(X) = (\sum_{i=1}^{k}(x_i - \overline{x})^2)/k$. Also, let $Y = [y_i, ..., y_k]$, where $y_i = ax_i + b$ for $i = 1, ..., k$, and $a > 0$ and $b \neq 0$ are constants. Lastly, let $\hat{X} = [\hat{x}_i, ..., \hat{x}_k]$ and $\hat{Y} = [\hat{y}_1, ..., \hat{y}_k]$, where the variables $\hat{x}_i$, $\hat{y}_i$, for $i = 1, ..., k$, have been standardized according to the formula (5). Then $D(\hat{X}, \hat{Y}) = 0$.

Proof. By noting

$$\begin{aligned}
\overline{\hat{x}} &= \frac{\overline{x} - \overline{x}}{\frac{1}{k}\sum_{i=1}^{k}(x_i - \overline{x})^2} = 0, \\
\overline{y} &= a\overline{x} + b, \\
\overline{\hat{y}} &= \frac{a\overline{x} + b - (a\overline{x} + b)}{\frac{1}{k}\sum_{i=1}^{k}(ax_i + b - (a\overline{x} + b))^2} = 0,
\end{aligned}$$

and

$$
\begin{aligned}
s(Y) &= \frac{1}{k}\sum_{i=1}^{k}(ax_i + b - \overline{y})^2 \\
&= \frac{1}{k}\sum_{i=1}^{k}(ax_i + b - (a\overline{x} + b))^2 \\
&= \frac{1}{k}\sum_{i=1}^{k}(a(x_i - \overline{x}))^2 \\
&= \frac{a^2}{k}\sum_{i=1}^{k}(x_i - \overline{x})^2 \\
&= a^2 s(X)
\end{aligned}
$$

it directly follows that:

$$
\begin{aligned}
D(\hat{X}, \hat{Y}) &= \frac{1}{k}\sum_{i=1}^{k}(d(\hat{x}_i, \hat{y}_i) - d(T(\hat{X}), T(\hat{Y}))) \\
&= \frac{1}{k}\sum_{i=1}^{k}((\hat{x}_i - \hat{y}_i)^2 - (\overline{\hat{x}} - \overline{\hat{y}})^2) \\
&= \frac{1}{k}\sum_{i=1}^{k}\left(\left(\frac{x_i - \overline{x}}{\sqrt{s(X)}} - \left(\frac{ax_i + b - (a\overline{x} + b)}{\sqrt{s(Y)}}\right)\right)^2 - (0-0)^2\right) \\
&= \frac{1}{k}\sum_{i=1}^{k}\left(\frac{x_i - \overline{x}}{\sqrt{s(X)}} - \left(\frac{a(x_i - \overline{x})}{\sqrt{s(Y)}}\right)\right)^2 \\
&= \frac{1}{k}\sum_{i=1}^{k}\left(\frac{x_i - \overline{x}}{\sqrt{s(X)}} - \left(\frac{a(x_i - \overline{x})}{\sqrt{a^2 s(X)}}\right)\right)^2 \\
&= \frac{1}{k}\sum_{i=1}^{k}\left(\frac{x_i - \overline{x}}{\sqrt{s(X)}} - \left(\frac{a(x_i - \overline{x})}{a\sqrt{s(X)}}\right)\right)^2 \\
&= \frac{1}{k}\sum_{i=1}^{k}\left(\frac{x_i - \overline{x}}{\sqrt{s(X)}} - \left(\frac{(x_i - \overline{x})}{\sqrt{s(X)}}\right)\right)^2 \\
&= 0.
\end{aligned}
$$

□

Lemma 4.2.7 and the associated proof show that when working with the standardized variables, the proximity measure (3) is zero when two curves are identical, up to a scaling by a constant $a > 0$ and vertical shift by $b \neq 0$.

In addition to the proximity measure (3) and the standardization formula (5), a formal definition of the linkage method to be used in the agglomerative hierarchical algorithm shall be provided. As previously noted, many different linkage methods exist, but the chosen linkage method is the average linkage. Its definition, as presented in [58], is given below.

Definition 4.2.8. Let Q and R be clusters or individual objects, and let $D(X, Y)$ be a proximity measure as defined in (3). Then, according to the average linkage method, the distance between the clusters Q and R is defined as:

$$d_A(Q, R) = \frac{1}{|R||Q|} \sum_{X \in R, Y \in Q} D(X, Y), \tag{6}$$

where $|R|$ and $|Q|$ are the number of objects in clusters R and Q, respectively.

To conclude this Section, a pseudo-algorithm of the agglomerative hierarchical clustering algorithm used to cluster the cancer incidence and mortality data is provided. The pseudo-algorithm combines all the different components presented above.

Algorithm Agglomerative hierarchical clustering

1: Standardize the variables (5) (optional)
2: Initialize each object as its own cluster
3: Compute the initial pairwise distances $D(X, Y)$ (3) between all objects
4: **Repeat**
5: Compute the average distances $d_A(Q, R)$ (6) between all cluster pairs
6: Merge the pair of clusters with the smallest $d_A(Q, R)$
7: Update the cluster partition accordingly
8: **Until** All objects are agglomerated into a single cluster

5 Identification of the cluster structures

In this Section, the agglomerative hierarchical clustering algorithm, discussed in more detail in Section 4 above, is used to cluster the cancer incidence and mortality data over time in Finland. The clustering is performed in two steps, first for the incidence data and then for the mortality data, in Sections 5.1 and 5.2, respectively.

5.1 Cluster structures of cancer incidence data over time in Finland

Figure 11 displays the resulting clusters after the agglomerative hierarchical clustering algorithm has been applied to the incidence rates per 100,000 person years of the most common cancers among females aged 20-29 years in Finland from 1962 to 2022. The original time series representing these incidence rates are visualized by Figure 3 in Section 3. The resulting cluster structures are shown for 2, 3, and 4 clusters. That is, the clustering algorithm has been halted once the predefined number of clusters has been reached, and the corresponding results are presented. This is due to the aforementioned reasons: since the number of clusters to be found among the data at hand is not known beforehand with certainty, it is reasonable to experiment with different numbers of resulting clusters.

The first observation arising from Figure 11 is that there is one large cluster containing many different cancers, while in the remaining clusters there is only one cancer per cluster. That is, for example, in the case of 4 resulting clusters, 7 of 10 the most common cancers among females aged 20-29 years form their own cluster, while the other 3 cancers each form a cluster of their own. The same holds also in the case of 3 and 2 resulting clusters. As expected, it seems that the scale affects the resulting clustering structure a bit, as, for example, in the case of 4 resulting clusters, the 3 cancers having the highest incidence rates per 100,000 person years, especially from the year 2000 onward, are excluded from the large cluster. Given that there is no other obvious trend than the modest increase of the incidence rates per 100,000 person years over time, this result appears to be reasonable.

Figure 12 illustrates the resulting clusters after the clustering algorithm has been applied to the same incidence rates per 100,000 person years of the most common cancers among females aged 20-29 years as above, with the exception that the incidence rates have been standardized according to the standardization formula (5) prior to clustering. Also in the case of the standardized data, one large cluster appears to be formed in all 3 cases of either 2, 3, or 4 resulting clusters, while the remaining cancers each form a cluster of their own. However, it can be observed that different cancers are excluded from the large cluster when using standardized versus unstandardized data. On the other hand, similarly to the unstandardized data, the scale of the standardized incidence rates per 100,000 person years seems to guide the clustering process in the sense that cancers having the largest spikes in their standardized incidence rate curves appear to be excluded from the large cluster.

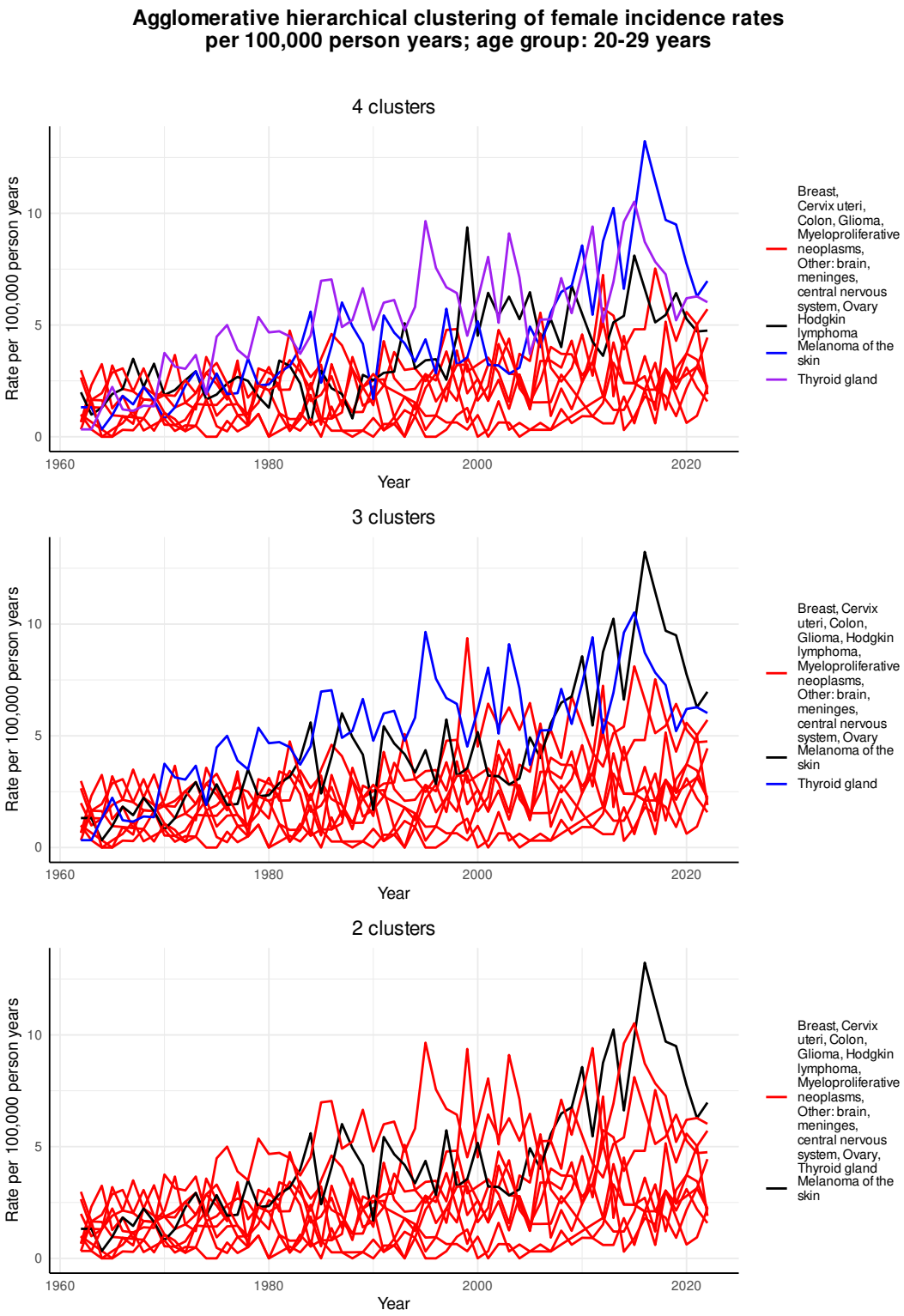


Figure 11: Agglomerative hierarchical clustering applied to the incidence rates per 100,000 person years of the most common cancers among females aged 20-29 years in Finland from 1962 to 2022.

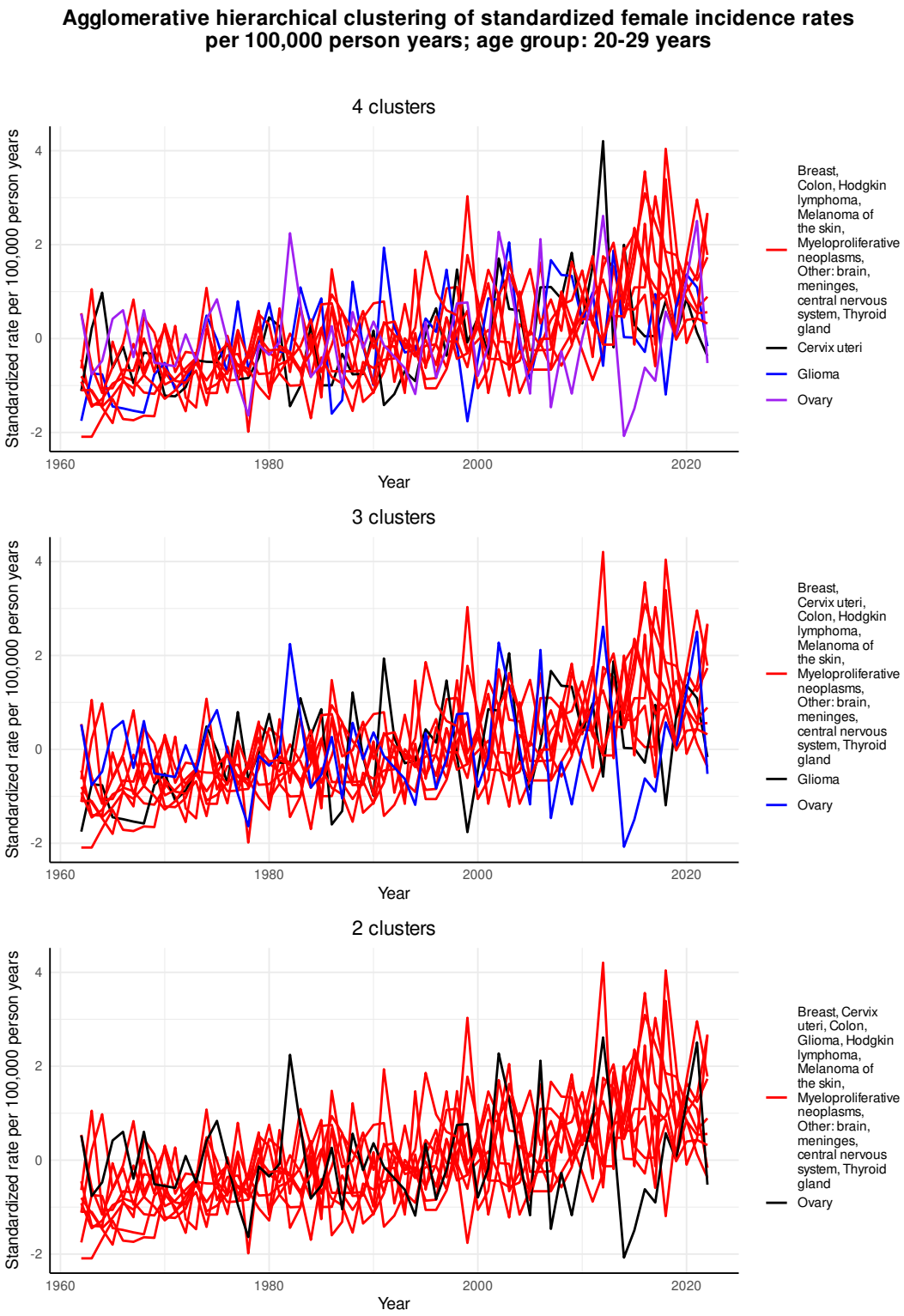


Figure 12: Agglomerative hierarchical clustering applied to the standardized incidence rates per 100,000 person years of the most common cancers among females aged 20-29 years in Finland from 1962 to 2022.

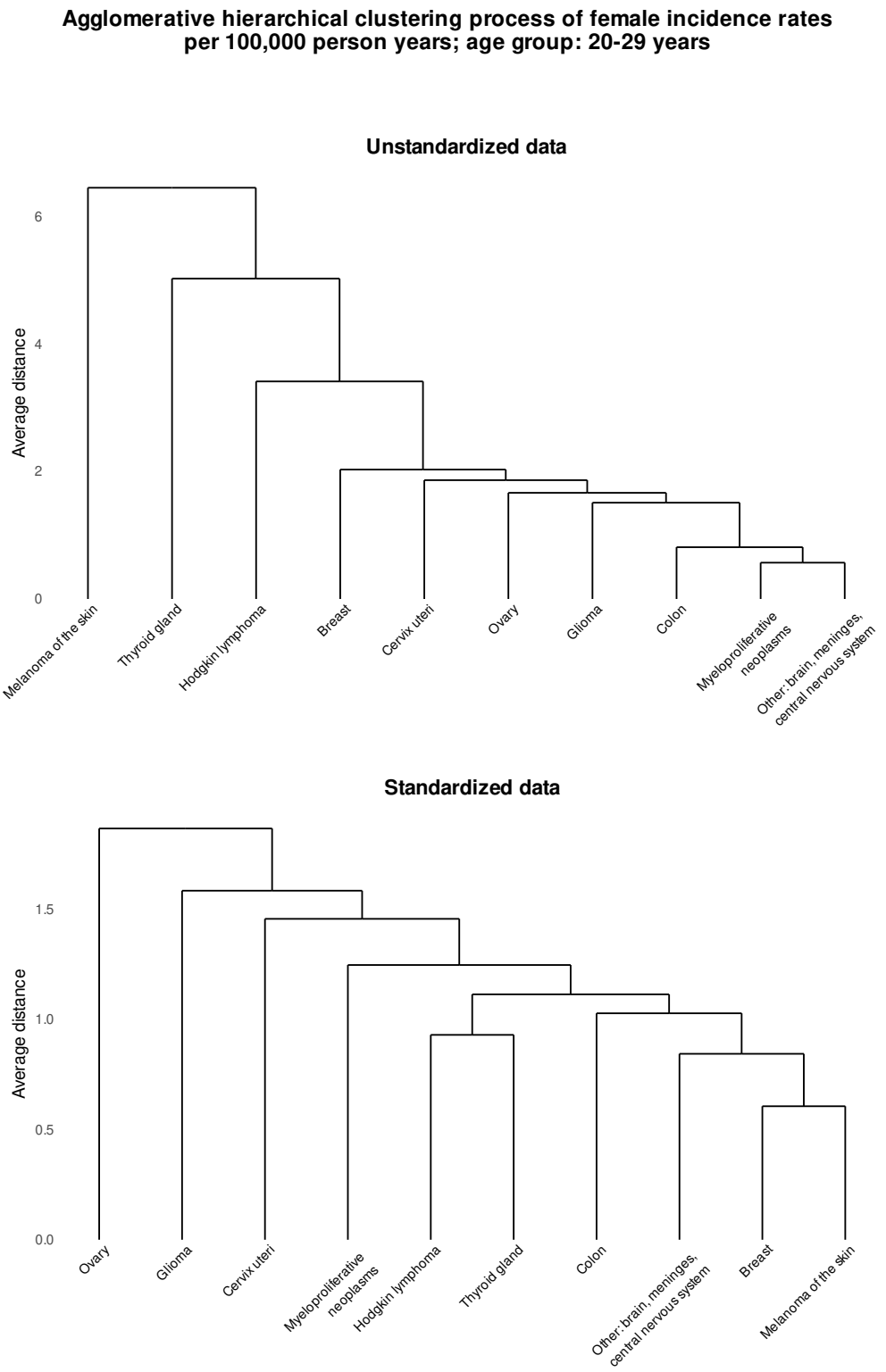


Figure 13: Dendrograms of clustered unstandardized and standardized cancer incidence data of the most common cancers among females aged 20-29 years in Finland from 1962 to 2022.

In addition to the clustering results shown by Figures 11 and 12, Figure 13 presents the corresponding dendrograms, or sequential representations of the clustering processes [56]. These dendrograms clarify the clustering process at each stage. For example, in the case of unstandardized data, the first merge occurs between myeloproliferative neoplasms and the group of the other cancers of the brain, meninges, and central nervous system. Subsequently, other cancers are merged one by one into the same growing cluster, while there are no parallel clusters emerging. This kind of clustering process is aligned with the earlier observation that in all cases of either 2, 3, or 4 resulting clusters, there is one large cluster containing most of the cancers, while the remaining cancers each form a cluster of their own. When using standardized data, the clustering process differs slightly. Namely, at one point, there are two separate clusters containing more than one cancer type. However, the smaller of the two clusters, formed by Hodgkin lymphoma and thyroid gland cancer, is soon merged into the larger cluster, after which the remaining cancers are again merged one by one into the large cluster.

The rest of this Section is organized as follows. Tables 5 - 9 gather and summarize the clustering results obtained by applying the agglomerative hierarchical clustering algorithm to the incidence rates per 100,000 person years of the most common cancers among females aged 20-29 to 70-79 years, and among males across the same age groups. The algorithm has been applied to both original and standardized data. Visualizations of the resulting clusters are presented by Figures B1 - B22 in Appendix B. Furthermore, the corresponding dendrograms are shown by Figures C1 - C11 in Appendix C.

Table 5 shows the clustering results obtained by applying the agglomerative hierarchical clustering algorithm to the incidence rates per 100,000 person years of the most common cancers among females in Finland across the age groups from 20-29 to 70-79 years. The corresponding visualizations of the results and the respective dendrograms are displayed by Figure 11 and Figures B1 - B5 in Appendix B, and by Figure 13 and Figures C1 - C5 in Appendix C, respectively. As a first observation from Table 5, it can be noted that in almost all cases in the age groups from 30-39 to 70-79 years breast cancer separates as its own cluster. The only exception is the age group of 30-39 years, and when the algorithm has been used to identify 2 resulting clustering, as breast cancer along with melanoma of the skin and thyroid gland cancer form one cluster, while the rest of the most common cancers form the other cluster. This outcome appears reasonable, as across all of these age groups the breast cancer incidence rate per 100,000 person years is clearly higher than that of any other cancer. On the other hand, among females aged 30-39 years, the difference between the incidence rate per 100,000 person years of breast cancer and the other most common cancers is not as big as across the other age groups. Additionally, as seen in Figure B1 in Appendix B, among females aged 30-39 years the incidence rates per 100,000 person years of breast cancer, melanoma of the skin, and thyroid gland cancer share a similar kind of increasing trend from 1962 to 2022, making it reasonable to include them into the same cluster in the case of 2 resulting clusters.

Age group	Cluster 1	Cluster 2	Cluster 3	Cluster 4
20-29 (4)	Hodgkin lymphoma	Melanoma of the skin	Thyroid gland	Rest
20-29 (3)	Melanoma of the skin	Thyroid gland	Rest	-
20-29 (2)	Melanoma of the skin	Rest	-	-
30-39 (4)	Breast	Cervix uteri	Melanoma of the skin, thyroid gland	Rest
30-39 (3)	Breast	Melanoma of the skin, thyroid gland	Rest	-
30-39 (2)	Breast, melanoma of the skin, thyroid gland	Rest	-	-
40-49 (4)	Breast	Cervix uteri	Melanoma of the skin, thyroid gland	Rest
40-49 (3)	Breast	Cervix uteri	Rest	-
40-49 (2)	Breast	Rest	-	-
50-59 (4)	Breast	Corpus uteri, ovary	Melanoma of the skin	Rest
50-59 (3)	Breast	Melanoma of the skin	Rest	-
50-59 (2)	Breast	Rest	-	-
60-69 (4)	Breast	Corpus uteri	Lung, trachea	Rest
60-69 (3)	Breast	Lung, trachea	Rest	-
60-69 (2)	Breast	Rest	-	-
70-79 (4)	Breast	Colon, corpus uteri, melanoma of the skin, skin squamous cell carcinoma	Lung, trachea	Other: brain, meninges, CNS; ovary, pancreas, rectum, rectosigmoid
70-79 (3)	Breast	Lung, trachea	Rest	-
70-79 (2)	Breast	Rest	-	-

Table 5: Summary of the clustering results after applying the agglomerative hierarchical clustering algorithm to the incidence rates per 100,000 person years of the most common cancers among females in Finland across age groups from 20-29 to 70-79 years.

Other observations from Table 5 include the following. First, in the age groups of 30-39 and 40-49 years, cervical cancer separates as its own cluster, when 4 resulting clusters are identified among females aged 30-39 years, and 3 or 4 resulting clusters are identified among females aged 40-49 years. Based on Figures B1 and B2 in Appendix B, this is explained by the decrease in the incidence rate per 100,000 person years of cervical cancer during the 1960s and early 1970s, while a similar decrease does not exist for the other most common cancers. Second, lung and tracheal cancer also tends to separate as its own cluster in the age groups of 60-69 and 70-79 years. The reason seems to be that in these age groups, the incidence rate per 100,000 person years of lung and tracheal cancer has increased in relative terms more than the incidence rates of the other most common cancers. Lastly, in the age groups from 20-29 to 40-49 years, melanoma of the skin and thyroid gland cancer appear to form either a cluster of their own or together, in particular, when 3 or 4 resulting clusters are identified. This seems reasonable, as the incidence rates per 100,000 person years of these cancer types is slightly higher than those of the other of the most common cancers during the 2010s and excluding breast cancer, and in relative terms these incidence rates seem to also have increased more from 1962 to 2022.

Table 6 presents the clustering results obtained by applying the agglomerative hierarchical clustering algorithm to the standardized incidence rates per 100,000 person years of the most common cancers among females in Finland across age groups from 20-29 to 70-79 years. The corresponding visualizations of the results and the respective dendrograms are shown by Figure 12 and Figures B6 - B10 in Appendix B, and by Figure 13 and Figures C1 - C5 in Appendix C, respectively. Based on Table 6, it can be observed that different cancers separate as their own smaller clusters compared to the resulting clusters when using unstandardized data, as Table 5 shows. For example, breast cancer, melanoma of the skin, and lung and tracheal cancer do not emerge as their own clusters, but it is rather glioma and ovarian cancer in the age groups of 20-29 and 30-39 years, cervical and other cancers of the brain, meninges, and central nervous system in the age groups of 40-49 and 50-59 years, and rectal and other cancers of the brain, meninges, and central nervous system in the age groups of 60-69 and 70-79 years. However, it should be noted that in the age groups of 20-29 and 30-39 years, no big differences can be observed in the underlying trends of the standardized incidence rates per 100,000 person years, but rather random fluctuations seem to have an influence on the resulting clusters, as illustrated by Figure 12 and Figure B6 in Appendix B. In contrast, more distinct differences can be observed in the underlying trends of the standardized incidence rates in the age groups from 40-49 to 70-79 years. In the age groups of 40-49 and 50-59 years, the standardized incidence rate per 100,000 person years of ovarian and corpus uteri cancer, as well as cervical cancer in the age group of 40-49 years, seem to exhibit a slightly decreasing trend from 1962 to 2022, while the standardized incidence rates per 100,000 person years of the other most common cancers in these age groups seem to increase over time. Also in the age group of 60-69 years, the standardized incidence rate per 100,000 person years of ovarian and corpus uteri cancer shows a slightly decreasing trend, but only from the 1990s onward, as seen in Figure B9 in Appendix B. Consequently, in the case of 3 or 4 resulting clusters, these two cancers form the cluster of their own. Lastly, in

Age group	Cluster 1	Cluster 2	Cluster 3	Cluster 4
20-29 (4)	Cervix uteri	Glioma	Ovary	Rest
20-29 (3)	Glioma	Ovary	Rest	-
20-29 (2)	Ovary	Rest	-	-
30-39 (4)	Cervix uteri, corpus uteri, other: brain, meninges, CNS	Glioma	Ovary	Rest
30-39 (3)	Cervix uteri, corpus uteri, other: brain, meninges, CNS	Ovary	Rest	-
30-39 (2)	Ovary	Rest	-	-
40-49 (4)	Cervix uteri	Corpus uteri, ovary	Other: brain, meninges, CNS	Rest
40-49 (3)	Cervix uteri	Corpus uteri, ovary	Rest	-
40-49 (2)	Cervix uteri, corpus uteri, ovary	Rest	-	-
50-59 (4)	Corpus uteri	Other: brain, meninges, CNS	Ovary	Rest
50-59 (3)	Corpus uteri, ovary	Other: brain, meninges, CNS	Rest	-
50-59 (2)	Corpus uteri, ovary	Rest	-	-
60-69 (4)	Corpus uteri, ovary	Other: brain, meninges, CNS	Rectum, rectosigmoid	Rest
60-69 (3)	Corpus uteri, ovary	Other: brain, meninges, CNS	Rest	-
60-69 (2)	Other: brain, meninges, CNS	Rest	-	-
70-79 (4)	Other: brain, meninges, CNS	Pancreas	Rectum, rectosigmoid	Rest
70-79 (3)	Other: brain, meninges, CNS	Rectum, rectosigmoid	Rest	-
70-79 (2)	Rectum, rectosigmoid	Rest	-	-

Table 6: Summary of the clustering results after applying the agglomerative hierarchical clustering algorithm to the standardized incidence rates per 100,000 person years of the most common cancers among females in Finland across age groups from 20-29 to 70-79 years.

the age group of 70-79 years, it is not ovarian and corpus uteri cancer, but rather rectal cancer that exhibits a slightly decreasing trend from 1962 to 2022 in terms of the standardized incidence rate per 100,000 person years, while the other most common cancers display an increasing trend over time. Therefore, in the case of the age group of 70-79 years, it appears reasonable that rectal cancer separates as its own cluster in all three cases of either 2, 3, or 4 resulting clusters.

Age group	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**
20-29 (4)	Hodgkin lymphoma	Melanoma of the skin	Testis	Rest
20-29 (3)	Hodgkin lymphoma	Testis	Rest	-
20-29 (2)	Testis	Rest	-	-
30-39 (4)	Hodgkin lymphoma	Melanoma of the skin	Testis	Rest
30-39 (3)	Melanoma of the skin	Testis	Rest	-
30-39 (2)	Testis	Rest	-	-
40-49 (4)	Lung, trachea	Melanoma of the skin	Prostate	Rest
40-49 (3)	Lung, trachea	Melanoma of the skin, prostate	Rest	-
40-49 (2)	Lung, trachea	Rest	-	-
50-59 (4)	Lung, trachea	Prostate	Stomach	Rest
50-59 (3)	Lung, trachea	Prostate	Rest	-
50-59 (2)	Lung, trachea	Rest	-	-
60-69 (4)	Colon, melanoma of the skin	Lung, trachea	Prostate	Rest
60-69 (3)	Lung, trachea	Prostate	Rest	-
60-69 (2)	Prostate	Rest	-	-
70-79 (4)	Lung, trachea	Prostate	Stomach	Rest
70-79 (3)	Lung, trachea; stomach	Prostate	Rest	-
70-79 (2)	Prostate	Rest	-	-

Table 7: Summary of the clustering results after applying the agglomerative hierarchical clustering algorithm to the incidence rates per 100,000 person years of the most common cancers among males in Finland across age groups from 20-29 to 70-79 years.

Table 7 shows the clustering results obtained by applying the agglomerative hierarchical clustering algorithm to the incidence rates per 100,000 person years of the most common cancers among males in Finland across age groups from 20-29

to 70-79 years. The corresponding visualizations of the results and the respective dendrograms are displayed by Figures B11 - B16 in Appendix B, and by Figures C6 - C11 in Appendix C, respectively. It can be observed that in most of the cases of either 2, 3, or 4 resulting clusters, there are only individual cancers forming clusters of their own, whereas the majority of the most common cancers across all of the considered age groups are grouped together. In the age groups of 20-29 and 30-39 years, it is Hodgkin lymphoma, melanoma of the skin, and testicular cancer that appear to separate as their own clusters due to their higher incidence rates per 100,000 person years compared to the other most common cancer types, as illustrated by Figures B11 and B12 in Appendix B. Also in the age group of 40-49 years, melanoma of the skin, as well as prostate cancer, separate as their own clusters in the case of 4 resulting clusters, and form a cluster together in the case of 3 resulting clusters. This seems reasonable, given their higher incidence rates per 100,000 person years compared to the other most common cancers during the 2000s. However, given even higher decrease in the incidence rate per 100,000 person years of lung and tracheal cancer from 1962 to 2022 makes it separate as its own cluster in all three either cases of 2, 3, or 4 resulting clusters.

When it comes to the age groups from 50-59 to 70-79 years, lung and tracheal cancer together with prostate cancer stand out from Table 7. That is, across all of these age groups and in the case of either 3 or 4 resulting clusters, these two cancers separate as their own clusters. Moreover, in the case of 2 resulting clusters, lung and tracheal cancer forms its own cluster among males aged 50-59 years, in a similar way as prostate cancer does in the age groups of 60-69 and 70-79 years, while the other most common cancers are grouped together. This outcome seems reasonable for two reasons. First, the incidence rates per 100,000 person years are higher than those of the other most common cancers in these age groups. Second, in the age group of 50-59 years, the decrease in the incidence rate per 100,000 person years of lung and tracheal cancer from 1962 to 2022 seems to be slightly higher than the increase in the incidence rate per 100,000 person years of prostate cancer, as presented by Figure B14 in Appendix B. However, the opposite, that is, higher increase in the incidence rate per 100,000 person years of prostate cancer than a decrease in the incidence rate of lung and tracheal cancer, seem to hold in the case of the age groups of 60-69 and 70-79 years, as seen in Figures B15 and B16 in Appendix B.

Tables 8 and 9 present the clustering results obtained by applying the agglomerative hierarchical clustering algorithm to the standardized incidence rates per 100,000 person years of the most common cancers among males in Finland across age groups from 20-29 to 70-79 years. The corresponding visualizations of the results and the respective dendrograms are shown by Figures B17 - B22 in Appendix B, and by Figures C6 - C11 in Appendix C, respectively. Similar to females, no big differences can be observed in the underlying trends of the standardized incidence rates per 100,000 person years of the most common cancers in the age groups 20-29 and 30-39 years, as seen in Figures B17 and B18 in Appendix B. Consequently, random fluctuations seem to guide the clustering process and the resulting cluster structures. In the other age groups from 40-49 to 70-79 years, one of the most apparent observations from Tables 8 and 9 is that lung and tracheal cancer tends in most cases separate as it own cluster. Based on

Age group	Cluster 1	Cluster 2	Cluster 3	Cluster 4
20-29 (4)	Acute lymphoblastic leukemia / lymphoma	Acute myeloid leukemia	Other: brain, meninges, CNS	Rest
20-29 (3)	Acute lymphoblastic leukemia / lymphoma	Acute myeloid leukemia	Rest	-
20-29 (2)	Acute myeloid leukemia	Rest	-	-
30-39 (4)	Bladder and urinary tract, Hodgkin lymphoma	Glioma	Other: brain, meninges, CNS	Rest
30-39 (3)	Bladder and urinary tract, Hodgkin lymphoma	Glioma	Rest	-
30-39 (2)	Bladder and urinary tract, Hodgkin lymphoma	Rest	-	-
40-49 (4)	Bladder and urinary tract	Lung, trachea	Other: brain, meninges, CNS	Rest
40-49 (3)	Bladder and urinary tract	Lung, trachea; other: brain, meninges, CNS	Rest	-
40-49 (2)	Lung, trachea; other: brain, meninges, CNS	Rest	-	-
50-59 (4)	Bladder and urinary tract	Kidney	Lung trachea; stomach	Rest
50-59 (3)	Bladder and urinary tract, kidney	Lung, trachea; stomach	Rest	-
50-59 (2)	Lung, trachea; stomach	Rest	-	-

Table 8: Summary of the clustering results after applying the agglomerative hierarchical clustering algorithm to the standardized incidence rates per 100,000 person years of the most common cancers among males in Finland across age groups from 20-29 to 50-59 years.

Age group	Cluster 1	Cluster 2	Cluster 3	Cluster 4
60-69 (4)	Bladder and urinary tract, kidney	Lung, trachea	Pancreas	Rest
60-69 (3)	Lung, trachea	Pancreas	Rest	-
60-69 (2)	Lung, trachea	Rest	-	-
70-79 (4)	Lung, trachea	Pancreas	Stomach	Rest
70-79 (3)	Lung, trachea; stomach	Pancreas	Rest	-
70-79 (2)	Lung, trachea; stomach	Rest	-	-

Table 9: Summary of the clustering results after applying the agglomerative hierarchical clustering algorithm to the standardized incidence rates per 100,000 person years of the most common cancers among males in Finland across age groups from 60-69 to 70-79 years.

Figures B19 - B10 in Appendix B this outcome appears reasonable, as in terms of the standardized incidence rates per 100,000 person years it exhibits a decreasing trend from 1962 to 2022, while most of the other most common cancers rather show an increasing trend during the same time interval. The few exceptions are other cancers of the brain, meninges, and central nervous system, whose standardized incidence rate per 100,000 person years decreased from 1962 to 2010 among males aged 40-49 years, and stomach cancer, whose standardized incidence rate per 100,000 person years also shows a decreasing trend in the age groups of 50-59 and 70-79 years, thus making these two cancers appear in the same cluster as lung and tracheal cancer in the respective age groups.

5.2 Cluster structures of cancer mortality data over time in Finland

Table 10 shows the clustering results obtained by applying the agglomerative hierarchical clustering algorithm to the mortalities per 100,000 person years of the most common cancers among females in Finland across age groups from 20-29 to 70-79 years. The corresponding visualizations of the results and the respective dendrograms are displayed by Figures B35 - B40 in Appendix B, and by Figures C12 - C17 in Appendix C, respectively. The first observation from Table 10 is that also in terms of mortalities per 100,000 person years, breast cancer separates as its own cluster in most cases, excluding the age group of 20-29 years, the age group of 60-69 years, when 2 resulting clusters are identified, and the age group of 70-79, when either 2 or 3 resulting clusters are identified. Given that the mortality per 100,000 person years of breast cancer is the highest among the most common cancers across age groups from 30-39 to 50-59 years, the outcomes appears again reasonable. On the other hand, as the mortality per 100,000 person years of lung and tracheal cancer exceeds that of breast cancer during the 2010s in the age groups of 60-69 and 70-79 years, as seen in

Age group	Cluster 1	Cluster 2	Cluster 3	Cluster 4
20-29 (4)	Glioma	Hodgkin lymphoma	Melanoma of the skin	Rest
20-29 (3)	Glioma	Hodgkin lymphoma	Rest	-
20-29 (2)	Glioma	Rest	-	-
30-39 (4)	Breast	Cervix uteri	Glioma	Rest
30-39 (3)	Breast	Cervix uteri	Rest	-
30-39 (2)	Breast	Rest	-	-
40-49 (4)	Breast	Cervix uteri	Ovary	Rest
40-49 (3)	Breast	Cervix uteri	Rest	-
40-49 (2)	Breast	Rest	-	-
50-59 (4)	Breast	Lung, trachea	Ovary	Rest
50-59 (3)	Breast	Lung, trachea	Rest	-
50-59 (2)	Breast	Rest	-	-
60-69 (4)	Breast	Lung, trachea	Pancreas	Rest
60-69 (3)	Breast	Lung, trachea	Rest	-
60-69 (2)	Lung, trachea	Rest	-	-
70-79 (4)	Breast	Colon, rectum, rectosigmoid	Lung, trachea	Rest
70-79 (3)	Colon, rectum, rectosigmoid	Lung, trachea	Rest	-
70-79 (2)	Lung, trachea	Rest	-	-

Table 10: Summary of the clustering results after applying the agglomerative hierarchical clustering algorithm to the mortalities per 100,000 person years of the most common cancers among females in Finland across age groups from 20-29 to 70-79 years.

Figures B27 and B28 in Appendix B, and also has increased in relative terms more than the mortalities per 100,000 person years of the other most common cancers, it seems sensible that it eventually separates as its own cluster in the aforementioned cases. Another interesting observation from Figure B28 in Appendix B is that for the age group of 70-79 years and in the case of either 3 or 4 resulting clusters, cancers of colon and rectum form the cluster of their own, even though the scale and trend of the corresponding mortalities per 100,000 person years appear to coincide rather well with those of the other most common cancers and excluding breast and lung and tracheal cancer.

Tables 11 and 12 present the clustering results obtained by applying the agglomerative hierarchical clustering algorithm to the standardized mortalities rates per 100,000 person years of the most common cancers among females in Finland across age groups from 20-29 to 70-79 years. The corresponding visualizations of the results and the respective dendrograms are shown by Figures B29 - B34 in Appendix B, and by Figures C12 - C17 in Appendix C, respectively. One observation arising from the visualizations presented by Figures B29 - B31 in Appendix B is that in the age groups

Age group	Cluster 1	Cluster 2	Cluster 3	Cluster 4
20-29 (4)	Breast, colon, ovary	Cervix uteri, glioma	Thyroid gland	Rest
20-29 (3)	Breast, colon, ovary	Glioma, thyroid gland	Rest	-
20-29 (2)	Glioma, thyroid gland	Rest	-	-
30-39 (4)	Corpus uteri	Meningeoma	Thyroid gland	Rest
30-39 (3)	Corpus uteri, meningeoma	Thyroid gland	Rest	-
30-39 (2)	Thyroid gland	Rest	-	-
40-49 (4)	Colon	Melanoma of the skin	Meningeoma	Rest
40-49 (3)	Melanoma of the skin	Meningeoma	Rest	-
40-49 (2)	Melanoma of the skin, meningeoma	Rest	-	-
50-59 (4)	Breast, melanoma of the skin	Colon	Lung, trachea	Rest
50-59 (3)	Colon	Lung, trachea	Rest	-
50-59 (2)	Colon, lung, trachea	Rest	-	-
60-69 (4)	Breast, other: brain, meninges, CNS; ovary	Lung, trachea; pancreas	Melanoma of the skin	Rest
60-69 (3)	Breast, other: brain, meninges, CNS; ovary	Lung, trachea; melanoma of the skin, pancreas	Rest	-
60-69 (2)	Lung, trachea; melanoma of the skin, pancreas	Rest	-	-

Table 11: Summary of the clustering results after applying the agglomerative hierarchical clustering algorithm to the standardized mortalities per 100,000 person years of the most common cancers among females in Finland across age groups from 20-29 to 60-69 years.

Age group	Cluster 1	Cluster 2	Cluster 3	Cluster 4
70-79 (4)	Breast, other: brain, meninges, CNS	Colon, rectum, rectosigmoid; skin squamous cell carcinoma	Corpus uteri	Rest
70-79 (3)	Colon, rectum, rectosigmoid; skin squamous cell carcinoma	Corpus uteri	Rest	-
70-79 (2)	Colon, rectum, rectosigmoid; skin squamous cell carcinoma	Rest	-	-

Table 12: Summary of the clustering results after applying the agglomerative hierarchical clustering algorithm to the standardized mortalities per 100,000 person years of the most common cancers among females in Finland aged 70-79 years.

from 20-29 to 40-49 years, no distinct trend can be observed for the standardized mortality per 100,000 person years for any individual cancer, but rather the standardized mortalities per 100,000 person years appear more or less constant from 1962 to 2022 in the age groups of 20-29 and 30-39 years, and slightly decreasing in the case of the age group of 40-49 years. Consequently, relatively few clusters result that consist of only one cancer, compared to the other cluster structures analyzed above. Also in the age groups of 50-59 and 60-69 years, the standardized mortalities per 100,000 person years of most of the most common cancer exhibit a slightly decreasing trend over time, while colon and lung and tracheal cancer in the age group of 50-59 years and pancreatic cancer, melanoma of the skin, and lung and tracheal cancer in the age group of 60-69 years show a slightly increasing trend, as seen in Figures B32 and B33 in Appendix B. Consequently, in the case of 2 resulting clusters, these cancers eventually end up in the same clusters in the respective age groups. For the age group of 70-79 years, the situation is the opposite, that is, in terms of the standardized mortalities per 100,000 person years, most of the most common cancers show a slightly increasing trend, while the standardized mortality per 100,000 person years of the cancers of colon and rectum as well as squamous cell carcinoma of the skin has decreased from 1962 to 2022, as displayed by Figure B34 in Appendix B. Again, this leads these cancers to form the cluster of their own in the case of 2 resulting clusters, while the rest of the most common cancers end up in the other cluster.

Table 13 shows the clustering results obtained by applying the agglomerative hierarchical clustering algorithm to the mortalities per 100,000 person years of the most common cancers among males in Finland across age groups from 20-29 to 70-79 years. The corresponding visualizations of the results and the respective dendrograms are displayed by Figures B35 - B40 in Appendix B, and by Figures C18 - C23 in Appendix C, respectively. In the age groups from 20-29 to 40-49 years, glioma exhibits one of the highest mortalities per 100,000 person years, as seen in Figures B35 - B37

Age group	Cluster 1	Cluster 2	Cluster 3	Cluster 4
20-29 (4)	Acute myeloid leukemia	Glioma	Hodgkin lymphoma	Rest
20-29 (3)	Acute myeloid leukemia	Hodgkin lymphoma	Rest	-
20-29 (2)	Acute myeloid leukemia, Hodgkin lymphoma	Rest	-	-
30-39 (4)	Glioma	Hodgkin lymphoma	Melanoma of the skin	Rest
30-39 (3)	Glioma	Hodgkin lymphoma	Rest	-
30-39 (2)	Hodgkin lymphoma	Rest	-	-
40-49 (4)	Glioma	Lung, trachea	Melanoma of the skin	Rest
40-49 (3)	Glioma	Lung, trachea	Rest	-
40-49 (2)	Lung, trachea	Rest	-	-
50-59 (4)	Kidney	Lung, trachea	Stomach	Rest
50-59 (3)	Lung, trachea	Stomach	Rest	-
50-59 (2)	Lung, trachea	Rest	-	-
60-69 (4)	Bladder and urinary tract	Lung, trachea	Prostate	Rest
60-69 (3)	Lung, trachea	Prostate	Rest	-
60-69 (2)	Lung, trachea	Rest	-	-
70-79 (4)	Lung, trachea	Prostate	Stomach	Rest
70-79 (3)	Lung, trachea	Stomach	Rest	-
70-79 (2)	Lung, trachea; stomach	Rest	-	-

Table 13: Summary of the clustering results after applying the agglomerative hierarchical clustering algorithm to the mortalities per 100,000 person years of the most common cancers among males in Finland across age groups from 20-29 to 70-79 years.

in Appendix B. However, its mortality per 100,000 person years is only slightly higher than that of the other most common cancers in these age groups. As a result, in the case of 2 resulting clusters, glioma does not form a cluster of its own, although it does so in most cases of 3 or 4 resulting clusters. In contrast, the slightly decreasing mortality per 100,000 person years of Hodgkin lymphoma over time among males aged 30-39 years and the similarly decreasing trend of lung and tracheal cancer combined with also higher mortality per 100,000 person years than glioma among males aged 40-49 years appear more distinct in the clustering process. Consequently, these cancers form their own clusters in the case of 2 resulting clusters in the respective age groups.

Also in the age groups from 50-59 to 70-79 years, the decreasing trend of lung and tracheal cancer, combined with its higher mortality per 100,000 person years compared to the other most common cancers, as seen in Figures B38 - B40 in Appendix B, seems to explain why it forms its own cluster in most cases. The only exception is the age group of 70-79 years in the case of 2 resulting clusters, when stomach cancer is clustered together with lung and tracheal cancer, while the rest of the other most common cancers form the other cluster. This result seems reasonable, as both cancers share a similar decreasing trend in terms of mortality per 100,000 person years over time. However, interestingly, among males aged 50-59 years, stomach and lung and tracheal cancer do not form a shared cluster, neither in the case of 2 nor 3 or 4 resulting clusters, despite showing a similar decreasing trend in that age group as well.

Tables 14 and 15 present the clustering results obtained by applying the agglomerative hierarchical clustering algorithm to the standardized mortalities rates per 100,000 person years of the most common cancers among males in Finland across age groups from 20-29 to 70-79 years. The corresponding visualizations of the results and the respective dendrograms are shown by Figures B41 - B46 in Appendix B, and by Figures C18 - C23 in Appendix C, respectively. In the age groups from 20-29 to 40-49 years, no clear trends are observable in the standardized mortalities per 100,000 person years, visualized by Figures B41 - B43 in Appendix B. Instead, random fluctuations in the standardized mortalities per 100,000 person years seem to guide the formation of the resulting clusters. For the age group of 50-59 years, the 3 clusters described by Table 15 appear reasonable. Namely, as seen in Figure B44 in Appendix B, the cancers belonging to cluster 1 seem to exhibit a slightly decreasing trend from 1962 to 2022 in terms of the standardized mortality per 100,000 person years, while the standardized mortality of the cancers in cluster 2 appears to have first increased slightly until the late 1990s and then decreased. The standardized mortality per 100,000 person years of diffuse B lymphoma forming cluster 3 is a bit different from the other two trends in the sense that there appears to be no reported mortality until the mid-1990s, which would explain the constant line seen in Figure B44. This also seems to affect the resulting cluster structure in the case of 2 clusters, as diffuse B lymphoma forms its own cluster, while the other most common cancers are clustered together. When it comes to the age groups of 60-69 and 70-79 years, they seem to share similar characteristics. That is, based on Figures B45 and B46 in Appendix B, there appears to be some cancers, including melanoma of the skin, colon, and pancreatic cancer for both age groups and liver cancer among males aged 60-69 years as well kidney cancer among males aged 70-79 years, whose standardized mortalities per 100,000 person years seem to exhibit slightly increasing trend over time, while the standardized mortalities per 100,000 person years of the other most common cancers show a slightly decreasing trend, especially from the 1990s onward. These two different trends are best captured in the resulting cluster structure, when the agglomerative hierarchical clustering algorithm is applied to identify 2 resulting clusters.

Age group	Cluster 1	Cluster 2	Cluster 3	Cluster 4
20-29 (4)	Acute lymphoblastic leukemia / lymphoma, diffuse B lymphoma, glioma	CNS, nerve sheath tumor	Thyroid gland	Rest
20-29 (3)	Acute lymphoblastic leukemia / lymphoma, diffuse B lymphoma, glioma	CNS, nerve sheath tumor	Rest	-
20-29 (2)	Acute lymphoblastic leukemia / lymphoma, diffuse B lymphoma, glioma	Rest	-	-
30-39 (4)	Bladder and urinary tract, CNS, nerve steath tumor; kidney, melanoma of the skin	Colon, Hodgkin lymphoma, other: brain, meninges, CNS; testis	Glioma	Thyroid gland
30-39 (3)	Bladder and urinary tract, CNS, nerve steath tumor; kidney, melanoma of the skin	Glioma	Rest	-
30-39 (2)	Glioma	Rest	-	-
40-49 (4)	Colon, rectum, rectosigmoid	Glioma	Prostate	Rest
40-49 (3)	Glioma	Prostate	Rest	-
40-49 (2)	Glioma	Rest	-	-

Table 14: Summary of the clustering results after applying the agglomerative hierarchical clustering algorithm to the standardized mortalities per 100,000 person years of the most common cancers among males in Finland across age groups from 20-29 to 40-49 years.

Age group	Cluster 1	Cluster 2	Cluster 3	Cluster 4
50-59 (4)	Bladder and urinary tract, lung, trachea; rectum, rectosigmoid; stomach	Colon, kidney, melanoma of the skin, prostate	Diffuse B lymphoma	Pharynx
50-59 (3)	Bladder and urinary tract, lung, trachea; rectum, rectosigmoid; stomach	Diffuse B lymphoma	Rest	-
50-59 (2)	Diffuse B lymphoma	Rest	-	-
60-69 (4)	Bladder and urinary tract, lung, trachea; skin squamous cell carcinoma	Colon, liver, melanoma of the skin	Prostate, rectum, rectosigmoid	Pancreas
60-69 (3)	Colon, liver, melanoma of the skin	Pancreas	Rest	-
60-69 (2)	Colon, liver, melanoma of the skin, pancreas	Rest	-	-
70-79 (4)	Colon, kidney, melanoma of the skin	Pancreas	Skin squamous cell carcinoma	Rest
70-79 (3)	Colon, kidney, melanoma of the skin	Pancreas	Rest	-
70-79 (2)	Colon, kidney, melanoma of the skin, pancreas	Rest	-	-

Table 15: Summary of the clustering results after applying the agglomerative hierarchical clustering algorithm to the standardized mortalities per 100,000 person years of the most common cancers among males in Finland across age groups from 50-59 to 70-79 years.

6 Interpretation of the cluster structures

When it comes to the threefold research question presented in Section 1.3, all the components needed to answer the question have been provided in Section 5 above. Moreover, in Section 5, the different cluster structures identified from the cancer incidence and mortality data over time in Finland are presented. In this Section, these cluster structures are further discussed, differences in the cluster structures across different subgroups determined by age and gender are evaluated, and whether hormone-related cancers emerged in the same cluster is considered.

One common feature observable among the cluster structures of the incidence rates per 100,000 person years across females and males is that for some of the cancers, which tend to separate as their own clusters, there exists either a national screening program or individualized, opportunistic screening tool. These cancers include breast and cervical cancer, for which there are national screening programs in Finland [52], [53], and prostate cancer that can be detected using the so-called PSA, or prostate-specific antigen, test [62], [63]. This is reasonable, as screening and testing can affect the incidence rate per 100,000 person years and mortality as well. Given that breast cancer screening may result in earlier diagnoses and possibly overdiagnosing some cases, as discussed in Section 3.3, the screening program is assumed to increase the incidence rate per 100,000 person years, while the corresponding mortality per 100,000 person years is assumed to decrease, compared to a situation without the screening program. The observed increase in incidence and decrease in mortality of prostate cancer since the launch of PSA testing in Finland [63] suggest that PSA testing shares similar characteristics. However, somewhat surprisingly, after excluding the effect of magnitude, that is, standardizing the data, breast cancer no longer tends to separate as its own cluster, but rather is clustered together with the majority of the other most common cancers. The same happens with prostate cancer, although there is a sharp increase in the incidence rate per 100,000 person years of the cancer during the late 1990s and early 2000s. In the case of cervical cancer the situation is a bit different. That is, cervical cancer screening allows for detecting cancer precursors, which, if cured, will not develop as actual cancer [64]. Consequently, the incidence rate per 100,000 person years of cervical cancer has decreased after the launch of the screening program [53], even so that also in the standardized case cervical cancer is clustered together with other cancers that exhibit a decreasing standardized incidence rate per 100,000 person years over time in the relevant age groups.

Another cancer type that was observed to separate as its own cluster in many cases, including both incidence and mortality data, and females and males, is lung and tracheal cancer. Among females, and in particular in the age groups of 60-69 and 70-79 years, it was noted that the relatively higher increase of incidence rate per 100,000 person years of lung and tracheal cancer over time than that of the other most common cancers resulted in clusters formed only by lung and tracheal cancer. A similar increase in mortality per 100,000 person years of lung and tracheal cancer led to the formation of the same kind of clusters also in the case of the mortality data, that is, those that contain only lung and tracheal cancer. There are few possible interpretations for these observations. First, there is a reason to believe that, over

time, gender differences between females and males have become smaller in Finland, making smoking today more approachable and common among females than it used to be in the past. Second, as indicated by Table 3, smokers are on average 15-30 times more likely to develop lung cancer than those who do not smoke. Third, lung cancer is also among the cancers causing the most deaths worldwide [8]. A combination of these aspects could help explain the cluster structure related to lung and tracheal cancer observed among female incidence rates and mortalities per 100,000 person years of the most common cancers in Finland. On the other hand, among males, both the incidence rate and mortality per 100,000 person years of lung and tracheal cancer has decreased over time. This could suggest that the awareness of the dangers of smoking have become better known and also internalized by people. In addition, the increased awareness of the association of asbestos and lung cancer as well as the regulation that has followed [65] could also partly explain the decrease of both incidence and mortality of lung and tracheal cancer among males in Finland.

It can be also observed how the interesting behavior of the standardized incidence rate per 100,000 person years of other cancers of the brain, meninges, and central nervous system caused this cancer type to form its own cluster in many cases both among females and males and across different age groups, as displayed by Tables 6 and 8. The behavior of this standardized incidence rate per 100,000 person years is interesting in the sense that in absolute terms, there appears to be no large increase in the incidence rate per 100,000 person years, but after standardization, the relative increase during the 2020s is greater than that of the other most common cancers in applicable age groups among females and males. Given that a similar increase in the standardized incidence rate per 100,000 person years of the other cancers of the brain, meninges, and central nervous system can be observed in multiple age groups and both among females and males, this could hint that there is some underlying common factor causing the change. For example, the diagnostics of this cancer type has perhaps improved, or there has been a change in exposure of some carcinogen causing this cancer in the environment. However, based on the available results, nothing can be concluded with certainty, and better understanding of the true reason behind this phenomenon would require further analyses.

When it comes to the differences, and also similarities, in the cluster structures across different subgroups determined by age and gender, a couple observations arise. First, for both females and males, random fluctuations appeared to guide the formation of the resulting cluster structures in the younger age groups of 20-29 and 30-39 years. This is reasonable, as aging remains as one of the major risk factors causing cancer [12] - [14]. On the other hand, in the older age groups from 40-49 to 70-79 years, it was possible to give also other interpretations to the cluster structures, including the possible effects of changes in smoking behavior, as discussed above. It can be also observed how lung and tracheal cancer tend to separate as its own cluster among females aged 60-69 and 70-79 years, while among males the observation holds across all age groups from 40-49 to 70-79 years. Another difference in the cluster structures between females and males is that in terms of female incidence rates per 100,000 person years, melanoma of the skin had a bit higher tendency to separate as its own cluster than in the case of male incidence rates per 100,000 person years.

In terms of hormone-related cancers, it can be observed that these cancers hardly formed common clusters. For example, breast cancer rather separated as its own cluster, when unstandardized incidence rates and mortalities per 100,000 person years were considered. A similar observation holds for males in terms of testicular and prostate cancer that also formed the cluster of their own instead of forming one together or with other hormone-related cancers. Therefore, in the context of the thesis, no connection between testicular and prostate cancer can be concluded, although the incidence rate per 100,000 person years of both cancers has seemingly increased in a similar way during the late 1990s and early 2000s. However, in the case of prostate cancer, this increase can be at least partly explained by the launch of PSA testing, as discussed above, while for the increase in testicular cancer incidence there is no unambiguous explanation yet [35].

What remains to be seen in the future is the effect of certain environmental and lifestyle factors on the resulting cluster structures. For example, among the younger people, the amount of vaping has been observed to increase, while there is also a concern that vaping might increase the risk of lung cancer [66]. Although in this thesis no separate clusters formed by lung and tracheal cancer among the younger age groups were observed, the current development might still have an effect on the cluster structures to be identified in the future, because as the authors also point out, there is a lag period between carcinogenic exposure and malignant transformation. Another interesting aspect is the impact of microplastics and xenoestrogens on the resulting cluster structures. However, given that the role of these two in cancer development is not yet fully understood, as discussed in Section 2, it is difficult to conclude anything definitive within the scope of the thesis. Lastly, there is a possible link between mobile phone use, which has increased in recent decades, and an increased risk of cancer [67]. However, in their meta-analysis, the authors of [67] identify some inconsistencies in the association between the cancer risk and mobile phone use, and thereby suggest further research on the topic. Hence, a better understanding of the effect of mobile phone use on the cluster structures found among the cancer incidence and mortality data would require further analyses.

7 Summary

In this thesis, cluster structures of cancer incidence and mortality data over time in Finland were studied. In particular, the interest was to determine what kind of cluster structures can be identified from the data, how the cluster structures differ across different subgroups determined by age and gender, and whether hormone-related cancers emerge in the same cluster due to the Western lifestyle. The cluster structures were identified using the agglomerative hierarchical clustering algorithm, and presented in Section 5. In total, the clustering was performed for 12 different subgroups, from the age group of 20-29 to 70-79 years, both females and males, and for both cancer incidence and mortality data. The main difference in cluster structures between subgroups was that random fluctuations appeared to have a greater impact on the results for the age groups of 20–29 and 30–39 years, whereas more interpretable cluster structures emerged among the age groups from 40-49 to 70-79 years.

Further contributions to the literature include the following. In Section 2, description of the Western lifestyle and its connection to cancer was composed. In Section 3.1, the most common cancers in terms of incidence in Finland in 2022 were identified. In Sections 3.2 and 3.3, a descriptive analysis of the development of incidence and mortality per 100,000 person years of the most common cancers among females and males and across different age groups from 20-29 to 70-79 years was provided. In Section 4, existing shape sensitive clustering methods for functional data were briefly reviewed, and an appropriate clustering method, that is, agglomerative hierarchical clustering, was chosen for the purposes of the thesis. Furthermore, in Section 4, few essential properties of the proximity measure (3), used as the basis of the agglomerative hierarchical clustering algorithm, were formally shown.

The thesis induces some possible future research directions. First, one could try to assess more carefully the effect of each individual factor of the Western lifestyle on the cluster structures. Second, in addition to the time series representing either cancer incidence or mortality, time series of different substances that cause cancer, such as certain food items, alcohol, microplastics, and xenoestrogens, could be included in the cluster analysis to find a possible association between these substances and cancer. Third, continuous approximations of the cancer incidence and mortality time series could be constructed and results compared with those obtained using the discrete observations as in the thesis. Fourth, a similar analysis as done here in the Finnish context could be replicated in other countries or worldwide. Fifth, different clustering methods, such as k-means, as well as proximity measures, could be experimented with. Lastly, from the theoretical point of view, one could study the asymptotic behavior and convergence of the chosen proximity measure. These future research directions coincide with the limitations of the thesis. An additional limitation is related to the data at hand. That is, it should be noted that the incidence rate per 100,000 person years does not convey the true number of cancer cases, only diagnosed cases. Nevertheless, the identified cluster structures based on the available data contribute to a better understanding of cancer epidemiology over time in Finland.

The thesis forms a foundation for a continued analysis examining cluster structures of Finnish cancer incidence data, which is being prepared for publication [68].

References

[1] Encyclopædia Britannica. “Cell”. Accessed: Jan. 2, 2025. [Online]. Available: https://www.britannica.com/science/cell-biology

[2] R. A. Weinberg, “How Cancer Arises,” *Scientific American*, vol. 275, no. 3, pp. 62–70, Sep. 1996, doi: 10.1038/scientificamerican0996-62.

[3] Cancer Society of Finland. “What is cancer?”. Accessed: Jan. 2, 2025. [Online]. Available: https://allaboutcancer.fi/facts-about-cancer/what-is-cancer/

[4] American Cancer Society. “Gene changes and cancer”. Accessed: Jan. 2, 2025. [Online]. Available: https://www.cancer.org/cancer/understanding-cancer/genes-and-cancer/gene-changes.html

[5] Cancer Research UK. “How does cancer start?”. Accessed: Jan. 2, 2025. [Online]. Available: https://www.cancerresearchuk.org/about-cancer/what-is-cancer/how-cancer-starts

[6] World Health Organization. “Health topics, Cancer”. Accessed: Jan. 2, 2025. [Online]. Available: https://www.who.int/health-topics/cancer

[7] Our World in Data. “Causes of death, World, 2021”. Accessed: Jan. 3, 2025. [Online]. Available: https://ourworldindata.org/grapher/annual-number-of-deaths-by-cause

[8] F. Bray, M. Laversanne, H. Sung, J. Ferlay, R. L. Siegel, I. Soerjomataram, and A. Jemal, “Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” *CA: A Cancer Journal for Clinicians*, vol. 74, no. 3, pp. 229–263, Apr. 2024, doi: 10.3322/caac.21834.

[9] H. Ritchie and E. Mathieu, “How many people die and how many are born each year?” *Our World in Data*, 2023. [Online]. Available: https://ourworldindata.org/births-and-deaths

[10] United Nations. “Global issues, Population”. Accessed: Jan. 3, 2025. [Online]. Available: https://www.un.org/en/global-issues/population

[11] United Nations. “Global issues, Aging”. Accessed: Jan. 3, 2025. [Online]. Available: https://www.un.org/en/global-issues/ageing

[12] K. Smetana, L. Lacina, P. Szabo, B. Dvořánková, P. Brož, and A. Šedo, “Ageing as an important risk factor for cancer,” *Anticancer Research*, vol. 36, no. 10, pp. 5009–5017, Sep. 2016, doi: 10.21873/anticanres.11069.

[13] L. Berben, G. Floris, H. Wildiers, and S. Hatse, “Cancer and Aging: Two Tightly Interconnected Biological Processes,” *Cancers*, vol. 13, no. 6, p. 1400, Mar. 2021, doi: 10.3390/cancers13061400.

[14] L. Montégut, C. López-Otín, and G. Kroemer, "Aging and cancer," *Molecular Cancer*, vol. 23, no. 1, p. 106, May 2024, doi: 10.1186/s12943-024-02020-z.

[15] World Health Organization. "Newsroom, Cancer". Accessed: Jan. 3, 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cancer

[16] American Cancer Society. "Determining if something is a carcinogen". Accessed: Jan. 3, 2025. [Online]. Available: https://www.cancer.org/cancer/risk-prevention/understanding-cancer-risk/determining-if-something-is-a-carcinogen.html

[17] K. I. Avgerinos, N. Spyrou, C. S. Mantzoros, and M. Dalamaga, "Obesity and cancer risk: Emerging biological mechanisms and perspectives," *Metabolism Clinical and Experimental*, vol. 92, pp. 121–135, Mar. 2019, doi: 10.1016/j.metabol.2018.11.001.

[18] A. Berrington de Gonzalez and S. Darby, "Risk of cancer from diagnostic X-rays: estimates for the UK and 14 other countries," *The Lancet*, vol. 363, no. 9406, pp. 345–351, Jan. 2004, doi: 10.1016/s0140-6736(04)15433-0.

[19] A. I. Osman, M. Hosny, A. S. Eltaweil, S. Omar, A. M. Elgarahy, M. Farghali, P.-S. Yap, Y.-S. Wu, S. Nagandran, K. Batumalaie, S. C. B. Gopinath, O. D. John, M. Sekar, T. Saikia, P. Karunanithi, M. H. M. Hatta, and K. A. Akinyede, "Microplastic sources, formation, toxicity and remediation: a review," *Environmental Chemistry Letters*, vol. 21, no. 4, pp. 2129–2169, Apr. 2023, doi: 10.1007/s10311-023-01593-3.

[20] Finnish Cancer Registry. "Cancer statistics". Accessed: Jan. 16, 2025. [Online]. Available: https://cancerregistry.fi/statistics/cancer-statistics/

[21] R. Eckersley, "Is modern Western culture a health hazard?" *International Journal of Epidemiology*, vol. 35, no. 2, pp. 252–258, Nov. 2006, doi: 10.1093/ije/dyi235.

[22] M. Arfan, H. M. Rasheed, M. Mahmood, T. Aqsa, U. Liaquat, and S. Ali, "The Elements Of Contemporary Western Culture-An Analytical Study," *Journal of Positive School Psychology*, pp. 1–17, 2023.

[23] L. L. Marchand, L. R. Wilkens, L. N. Kolonel, J. H. Hankin, and L.-C. Lyu, "Associations of Sedentary Lifestyle, Obesity, Smoking, Alcohol Use, and Diabetes with the Risk of Colorectal Cancer," *Cancer Research*, vol. 57, no. 21, pp. 4787–4794, Nov. 1997.

[24] L. Cordain, S. B. Eaton, A. Sebastian, N. Mann, S. Lindeberg, B. A. Watkins, J. H. O'Keefe, and J. Brand-Miller, "Origins and evolution of the Western diet: health implications for the 21st century," *The American Journal of Clinical Nutrition*, vol. 81, no. 2, pp. 341–354, Feb. 2005, doi: 10.1093/ajcn.81.2.341.

[25] A. Azzam, "Is the world converging to a 'Western diet'?" *Public Health Nutrition*, vol. 24, no. 2, pp. 309–317, Oct. 2020, doi: 10.1017/S136898002000350X.

[26] P. Carrera-Bastos, M. Fontes-Villalba, J. H. O'Keefe, S. Lindeberg, and L. Cordain, "The western diet and lifestyle and diseases of civilization," *Research Reports in Clinical Cardiology*, pp. 15–35, Mar. 2011, doi: 10.2147/RRCC.S16919.

[27] M. Á. Martínez-González, J. A. Martínez, F. B. Hu, M. J. Gibney, and J. Kearney, "Physical inactivity, sedentary lifestyle and obesity in the European Union," *International Journal of Obesity*, vol. 23, no. 11, pp. 1192–1201, Nov. 1999, doi: 10.1038/sj.ijo.0801049.

[28] K. H. Pietiläinen, J. Kaprio, P. Borg, G. Plasqui, H. Yki-Järvinen, U. M. Kujala, R. J. Rose, K. R. W. R, and A. Rissanen, "Physical Inactivity and Obesity: A Vicious Circle," *Obesity*, vol. 16, no. 2, pp. 409–414, Jan. 2008, doi: 10.1038/oby.2007.72.

[29] C. W. Warren, N. Jones, M. P. Eriksen, and S. Asma, "Patterns of global tobacco use in young people and implications for future chronic disease burden in adults," *The Lancet*, vol. 367, no. 9512, pp. 749–753, Mar. 2006, doi: 10.1016/S0140-6736(06)68192-0.

[30] M. Ng, M. K. Freeman, T. D. Fleming, M. Robinson, L. Dwyer-Lindgren, B. Thomson, A. Wollum, E. Sanman, S. Wulf, A. D. Lopez, C. J. L. Murray, and E. Gakidou, "Smoking Prevalence and Cigarette Consumption in 187 Countries, 1980-2012," *Jama*, vol. 311, no. 2, pp. 183–192, Jan. 2014, doi: 10.1001/jama.2013.284692.

[31] World Health Organization. "Global status report on alcohol and health and treatment of substance use disorders". Accessed: Jan. 24, 2025. [Online]. Available: https://www.who.int/publications/i/item/9789240096745

[32] K. D. Cox, G. A. Covernton, H. L. Davies, J. F. Dower, F. Juanes, and S. E. Dudas, "Human Consumption of Microplastics," *Environmental Science & Technology*, vol. 53, no. 12, pp. 7068–7074, Jan. 2019, doi: 10.1021/acs.est.9b01517.

[33] D. E. Bronowicka-Kłys, M. Lianeri, and P. P. Jagodziński, "The role and impact of estrogens and xenoestrogen on the development of cervical cancer," *Biomedicine & Pharmacotherapy*, vol. 84, pp. 1945–1953, Dec. 2016, doi: 10.1016/j.biopha.2016.11.007.

[34] D. W. Singleton and S. A. Khan, "Xenoestrogen exposure and mechanisms of endocrine disruption," *Frontiers in Bioscience*, vol. 8, pp. s110–s118, Jan. 2003, doi: 10.2741/1010.

[35] A. Fucic, M. Gamulin, Z. Ferencic, J. Katic, M. K. von Krauss, A. Bartonova, and D. F. Merlo, "Environmental exposure to xenoestrogens and oestrogen related cancers: reproductive system, breast, lung, kidney, pancreas, and brain," *Environmental Health*, vol. 11, pp. 1–9, Feb. 2012, doi: 10.1186/1476-069x-11-s1-s8.

[36] World Cancer Research Fund/American Institute for Cancer Research. "Diet, Nutrition, Physical Activity and Cancer: a Global Perspective". Accessed: Jan. 28, 2025. [Online]. Available: https://www.wcrf.org/wp-content/uploads/2024/11/Summary-of-Third-Expert-Report-2018.pdf

[37] S. Hanson, G. Thorpe, L. Winstanley, A. S. Abdelhamid, and L. Hooper, "Omega-3, omega-6 and total dietary polyunsaturated fat on cancer incidence: systematic review and meta-analysis of randomised trials," *British Journal of Cancer*, vol. 122, no. 8, pp. 1260–1270, Feb. 2020, doi 10.1038/s41416-020-0761-6.

[38] L. D'Elia, G. Rossi, R. Ippolito, F. P. Cappuccio, and P. Strazzullo, "Habitual salt intake and risk of gastric cancer: A meta-analysis of prospective studies," *Clinical Nutrition*, vol. 31, no. 4, pp. 489–498, Aug. 2012, doi: 10.1016/j.clnu.2012.01.003.

[39] A. Urbute, K. Frederiksen, and S. K. Kjaer, "Early adulthood overweight and obesity and risk of premenopausal ovarian cancer, and premenopausal breast cancer including receptor status: prospective cohort study of nearly 500,000 Danish women," *Annals of Epidemiology*, vol. 70, pp. 61–67, Apr. 2022, doi: 10.1016/j.annepidem.2022.03.013.

[40] J. A. Knight, "Physical inactivity: associated diseases and disorders," *Annals of Clinical & Laboratory Science*, vol. 42, no. 3, pp. 320–337, Jan. 2012.

[41] R. J. Shephard, "Physical Activity and Prostate Cancer: An Updated Review," *Sports Medicine*, vol. 47, pp. 1055–1073, June 2017, doi: 10.1007/s40279-016-0648-0.

[42] A. J. Sasco, M. Secretan, and K. Straif, "Tobacco smoking and cancer: a brief review of recent epidemiological evidence," *Lung Cancer*, vol. 45, pp. S3–S9, Aug. 2004, doi: 10.1016/j.lungcan.2004.07.998.

[43] P. Boffetta and M. Hashibe, "Alcohol and cancer," *The Lancet Oncology*, vol. 7, no. 2, pp. 149–156, Feb. 2006, doi: 10.1016/S1470-2045(06)70577-0.

[44] A. Baspakova, A. Zare, R. Suleimenova, A. B. Berdygaliev, B. Karimsakova, K. Tussupkaliyeva, N. M. Mussin, K. R. Zhilisbayeva, N. Tanideh, and A. Tamadon, "An updated systematic review about various effects of microplastics on cancer: A pharmacological and in-silico based analysis," *Molecular Aspects of Medicine*, vol. 101, p. 101336, Jan. 2025, doi: 10.1016/j.mam.2024.101336.

[45] Eurostat. "Demographic change in Europe - Country factsheets: Finland". Accessed: Jan. 13, 2025. [Online]. Available: https://ec.europa.eu/eurostat/documents/12743486/14207633/FI-EN.pdf

[46] American Cancer Society. "Risk factors and causes of childhood cancer". Accessed: Jan. 13, 2025. [Online]. Available: https://www.cancer.org/cancer/types/cancer-in-children/risk-factors-and-causes.html

[47] National Cancer Institute. “NCI Dictionary of cancer terms, Carcinoma in situ”. Accessed: Jan. 14, 2025. [Online]. Available: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/carcinoma-in-situ

[48] National Cancer Institute. “NCI Dictionary of cancer terms, Noninvasive”. Accessed: Jan. 14, 2025. [Online]. Available: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/noninvasive

[49] National Cancer Institute. “Ovarian borderline tumors treatment (PDQ®) - Health professional version”. Accessed: Jan. 14, 2025. [Online]. Available: https://www.cancer.gov/types/ovarian/hp/ovarian-borderline-tumors-treatment-pdq

[50] G. Gibson and I. Ahmed, “Perianal and genital basal cell carcinoma: A clinicopathologic review of 51 cases,” *Journal of the American Academy of Dermatology*, vol. 45, no. 1, pp. 68–71, July 2001, doi: 10.1067/mjd.2001.114588.

[51] American Cancer Society. “What are basal and squamous cell skin cancers”. Accessed: Jan. 14, 2025. [Online]. Available: https://www.cancer.org/cancer/types/basal-and-squamous-cell-skin-cancer/about/what-is-basal-and-squamous-cell.html

[52] Finnish Cancer Registry. “Breast cancer screening programme - Annual review 2023”. Accessed: Jan. 16, 2025. [Online]. Available: https://syoparekisteri.fi/assets/files/2023/12/The_breast_cancer_screening_programme_in_Finland_annual_review_2023.pdf

[53] A. Anttila, E. Pukkala, B. Söderman, M. Kallio, P. Nieminen, and M. Hakama, “Effect of organised screening on cervical cancer incidence and mortality in Finland, 1963–1995: Recent increase in cervical cancer incidence,” *International Journal of Cancer*, vol. 83, no. 1, pp. 59–65, Nov. 1999, doi: 10.1002/(SICI)1097-0215(19990924)83:1%3C59::AID-IJC12%3E3.0.CO;2-N.

[54] M. Milosevic, D. Jankovic, A. Milenkovic, and D. Stojanov, “Early diagnosis and detection of breast cancer,” *Technology and Health Care*, vol. 26, no. 4, pp. 729–759, Sep. 2018, doi: 10.3233/THC-181277.

[55] M. Ilic and I. Ilic, “Epidemiology of pancreatic cancer,” *World Journal of Gastroenterology*, vol. 22, no. 44, p. 9694, Nov. 2016, doi: 10.3748/wjg.v22.i44.9694.

[56] W. K. Härdle and L. Simar, *Applied Multivariate Statistical Analysis*, 4th ed. Springer Berlin Heidelberg, 2015.

[57] J.-L. Wang, J.-M. Chiou, and H.-G. Müller, “Functional Data Analysis,” *Annual Review of Statistics and Its Application*, vol. 3, no. 1, pp. 257–295, June 2016, doi: 10.1146/annurev-statistics-041715-033624.

[58] L. Ferreira and D. B. Hitchcock, "A Comparison of Hierarchical Methods for Clustering Functional Data," *Communications in Statistics-Simulation and Computation*, vol. 38, no. 9, pp. 1925–1949, Dec. 2009, doi: 10.1080/03610910903168603.

[59] M. Zhang and A. Parnell, "Review of Clustering Methods for Functional Data," *ACM Transactions on Knowledge Discovery from Data*, vol. 17, no. 7, pp. 1–34, Apr. 2023, doi: 10.1145/3581789.

[60] J. Stewart, D. Clegg, and S. Watson, *Calculus: Early Transcendentals*, 7th ed. Brooks/Cole Cengage Learning, 2012.

[61] P. J. Huber, *Robust Statistics*, 1st ed. John Wiley & Sons, Inc., 1981.

[62] C. Neppl-Huber, M. Zappa, J. W. Coebergh, E. Rapiti, J. Rachtan, B. Holleczek, S. Rosso, T. Aareleid, H. Brenner, and A. Gondos, "Changes in incidence, survival and mortality of prostate cancer in Europe and the United States in the PSA era: additional diagnoses and avoided deaths," *Annals of Oncology*, vol. 23, no. 5, pp. 1325–1334, Sep. 2011, doi: 10.1093/annonc/mdr414.

[63] H. A. Seikkula, A. J. Kaipia, M. E. Rantanen, J. M. Pitkäniemi, N. K. Malila, and P. J. Boström, "Stage-specific mortality and survival trends of prostate cancer patients in Finland before and after introduction of PSA," *Acta Oncologica*, vol. 56, no. 7, pp. 971–977, Feb. 2017, doi: 10.1080/0284186X.2017.1288298.

[64] Finnish Cancer Registry. "Cervical cancer screening". Accessed: Aug. 5, 2025. [Online]. Available: https://cancerregistry.fi/screening/cervical-cancer-screening

[65] J. Rantanen, S. Lehtinen, M. Huuskonen, P. Oksa, A. Tossavainen, T. Tuomi, and H. Vainio. "Prevention and Management of Asbestos-Related Diseases in Finland". Accessed: July 27, 2025. [Online]. Available: https://www.julkari.fi/bitstream/handle/10024/135520/PreventionandManagementofAsbestos-RelatedDiseasesinFinland.pdf

[66] D. Bracken-Clarke, D. Kapoor, A. M. Baird, P. J. Buchanan, K. Gately, S. Cuffe, and S. P. Finn, "Vaping and lung cancer–A review of current data and recommendations," *Lung Cancer*, vol. 153, pp. 11–20, Jan. 2021, doi: 10.1016/j.lungcan.2020.12.030.

[67] S.-K. Myung, W. Ju, D. D. McDonnell, Y. J. Lee, G. Kazinets, C.-T. Cheng, and J. M. Moskowitz, "Mobile Phone Use and Risk of Tumors: A Meta-Analysis," *Journal of Clinical Oncology*, vol. 27, no. 33, pp. 5565–5572, Oct. 2009, doi: 10.1200/JCO.2008.21.6366.

[68] T. Huhtinen and P. Ilmonen, "On Cluster Structures of Finnish Cancer Incidence Data," 2025, manuscript in preparation.

A Development of the incidence and mortality of the most common cancers in Finland as moving averages over time

This Appendix contains Figures A1 - A8 representing the incidence and mortality of the most common cancers among females and males in Finland as 10-year moving averages from 1962 to 2022. The most common cancers refer to the most common cancers identified in Section 3.1. In Sections 3.2 and 3.3, Figures 3 - 10 visualize the underlying time series used to produce Figures A1 - A8 below.

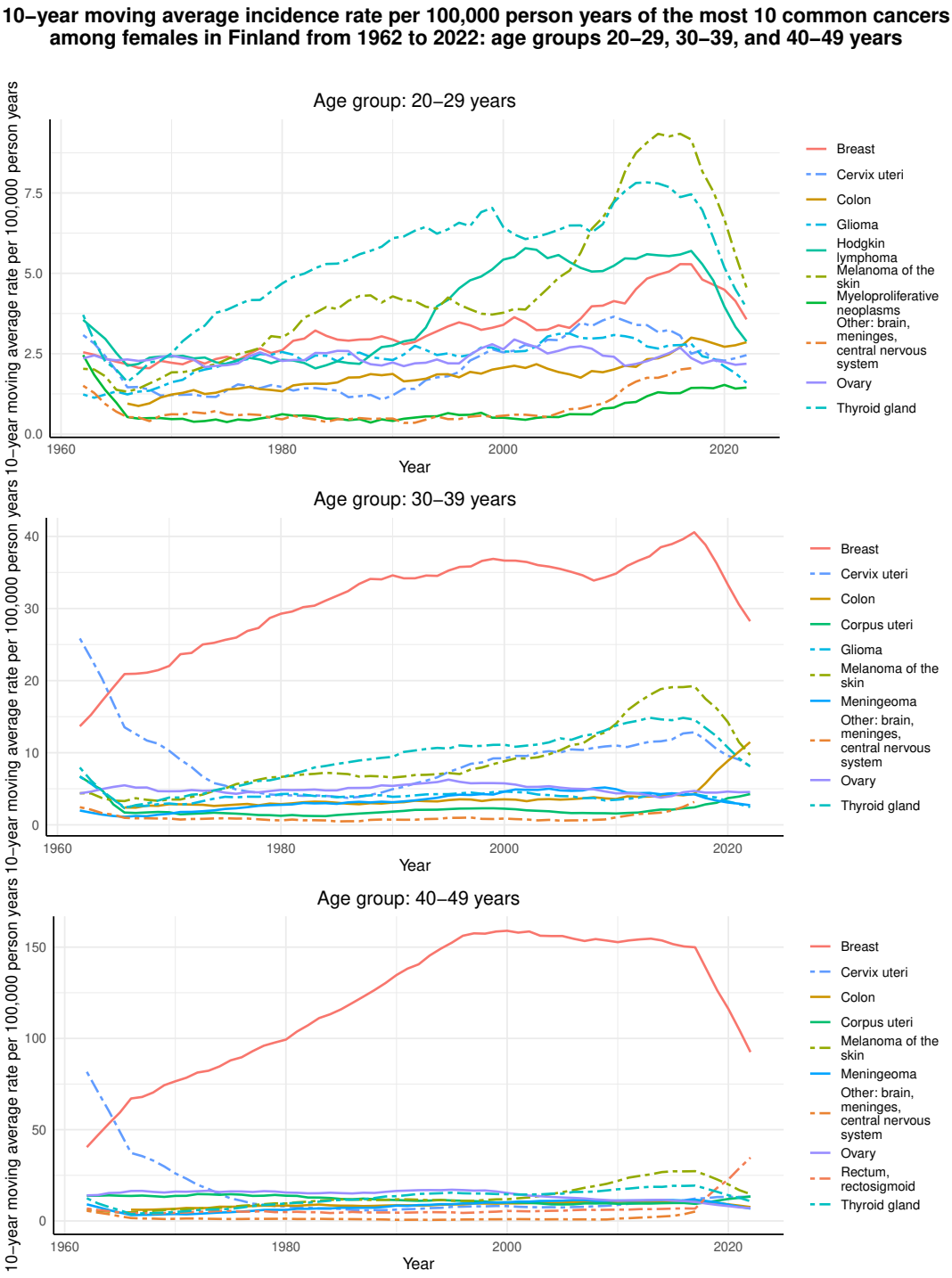


Figure A1: 10-year moving average incidence rate per 100,000 person years of the most common cancers among females in Finland from 1962 to 2022 for age the groups 20-29, 30-39, and 40-49 years.

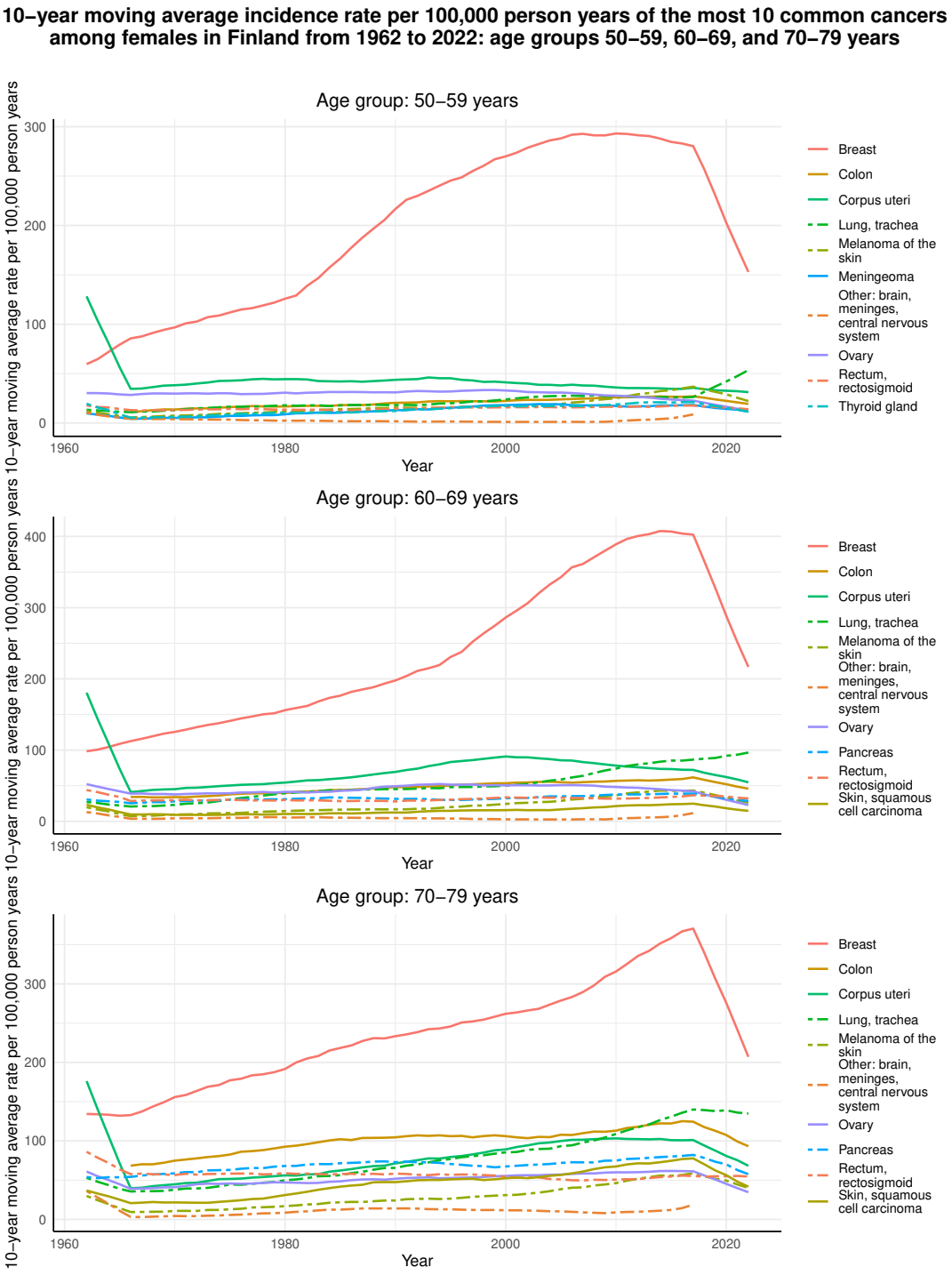


Figure A2: 10-year moving average incidence rate per 100,000 person years of the most common cancers among females in Finland from 1962 to 2022 for the age groups 50-59, 60-69, and 70-79 years.

10−year moving average incidence rate per 100,000 person years of the most 10 common cancers among males in Finland from 1962 to 2022: age groups 20−29, 30−39, and 40−49 years

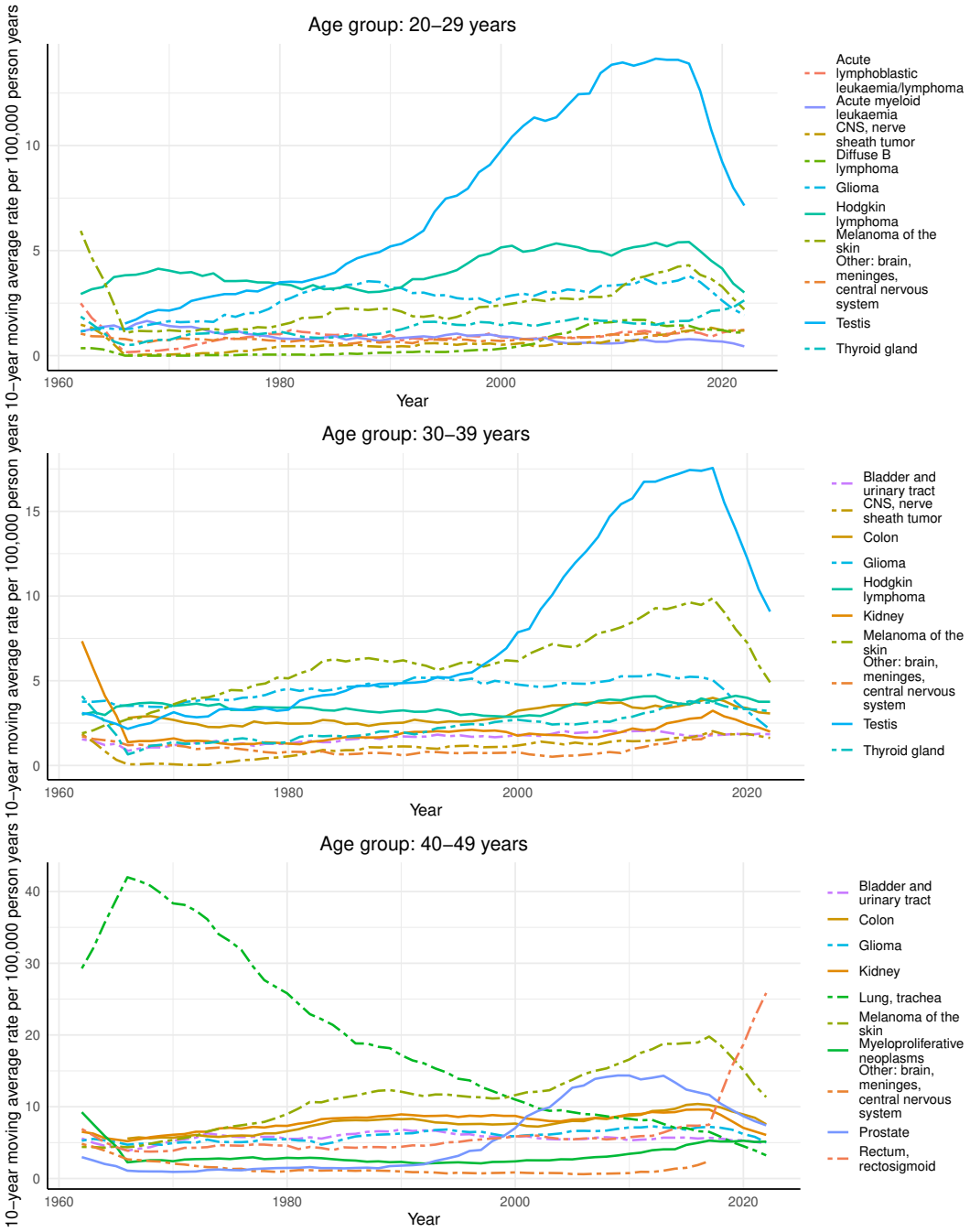


Figure A3: 10-year moving average incidence rate per 100,000 person years of the most common cancers among males in Finland from 1962 to 2022 for age the groups 20-29, 30-39, and 40-49 years.

Figure A4: 10-year moving average incidence rate per 100,000 person years of the most common cancers among males in Finland from 1962 to 2022 for the age groups 50-59, 60-69, and 70-79 years.

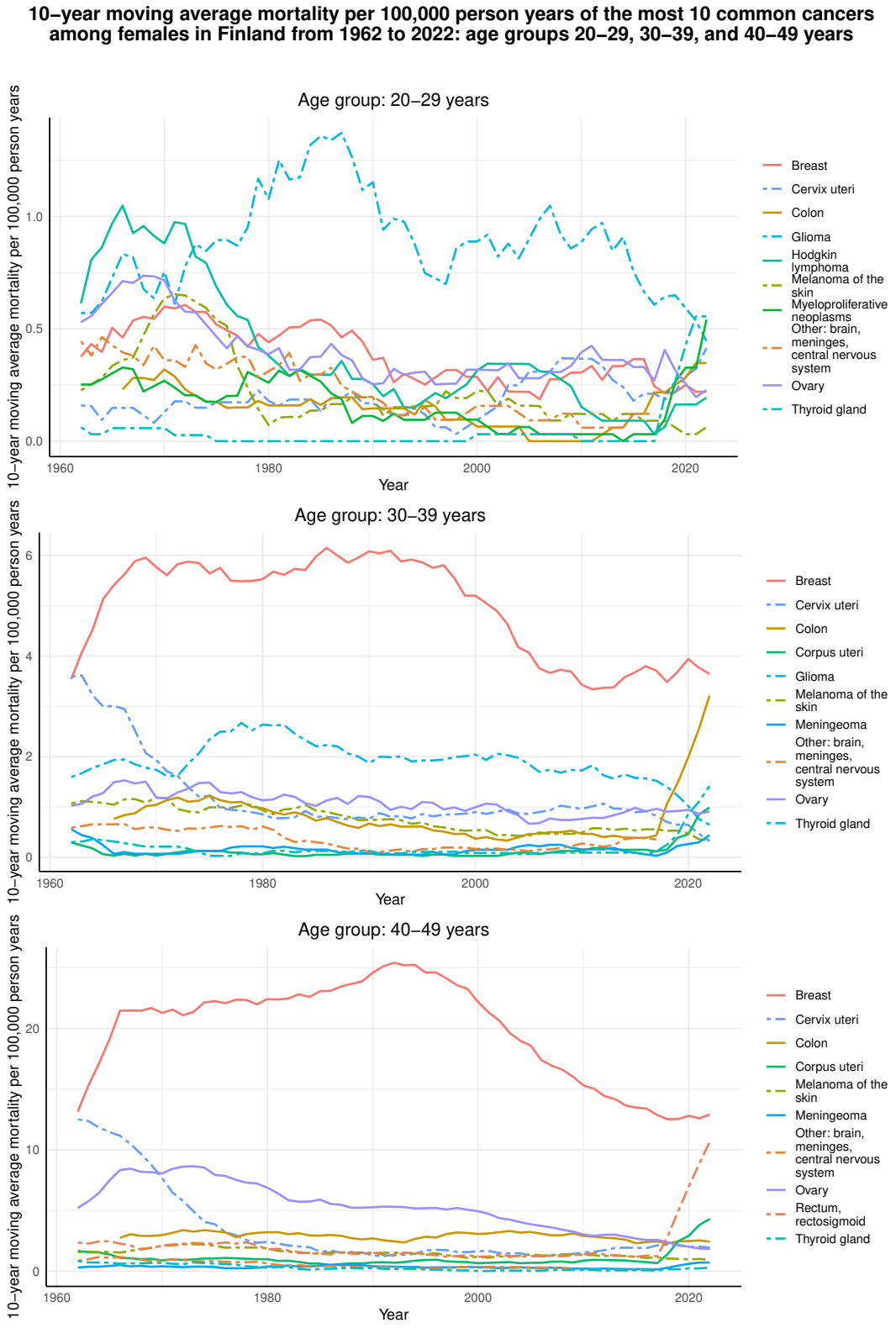


Figure A5: 10-year moving average mortality per 100,000 person years of the most common cancers among females in Finland from 1962 to 2022 for age the groups 20-29, 30-39, and 40-49 years.

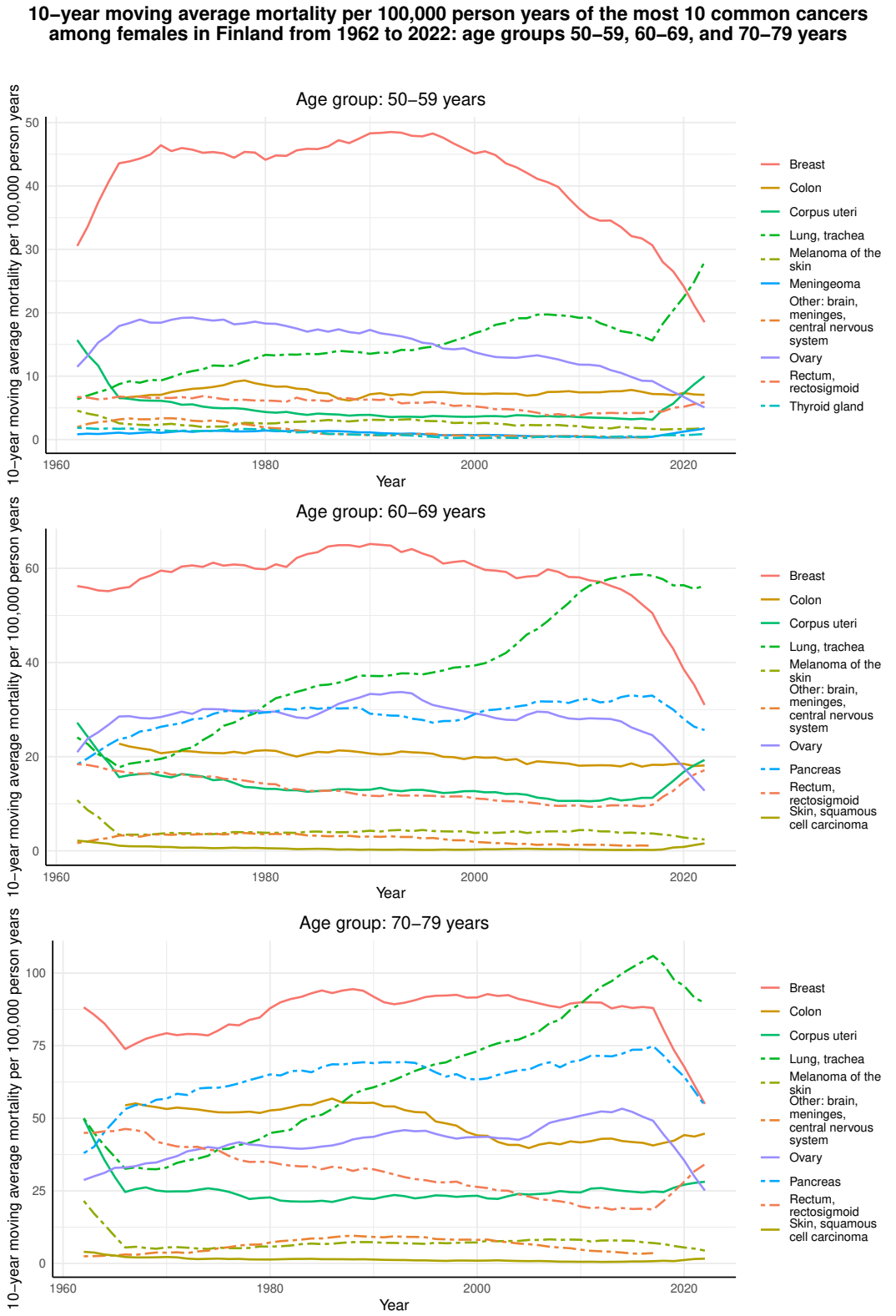


Figure A6: 10-year moving average mortality per 100,000 person years of the most common cancers among females in Finland from 1962 to 2022 for the age groups 50-59, 60-69, and 70-79 years.

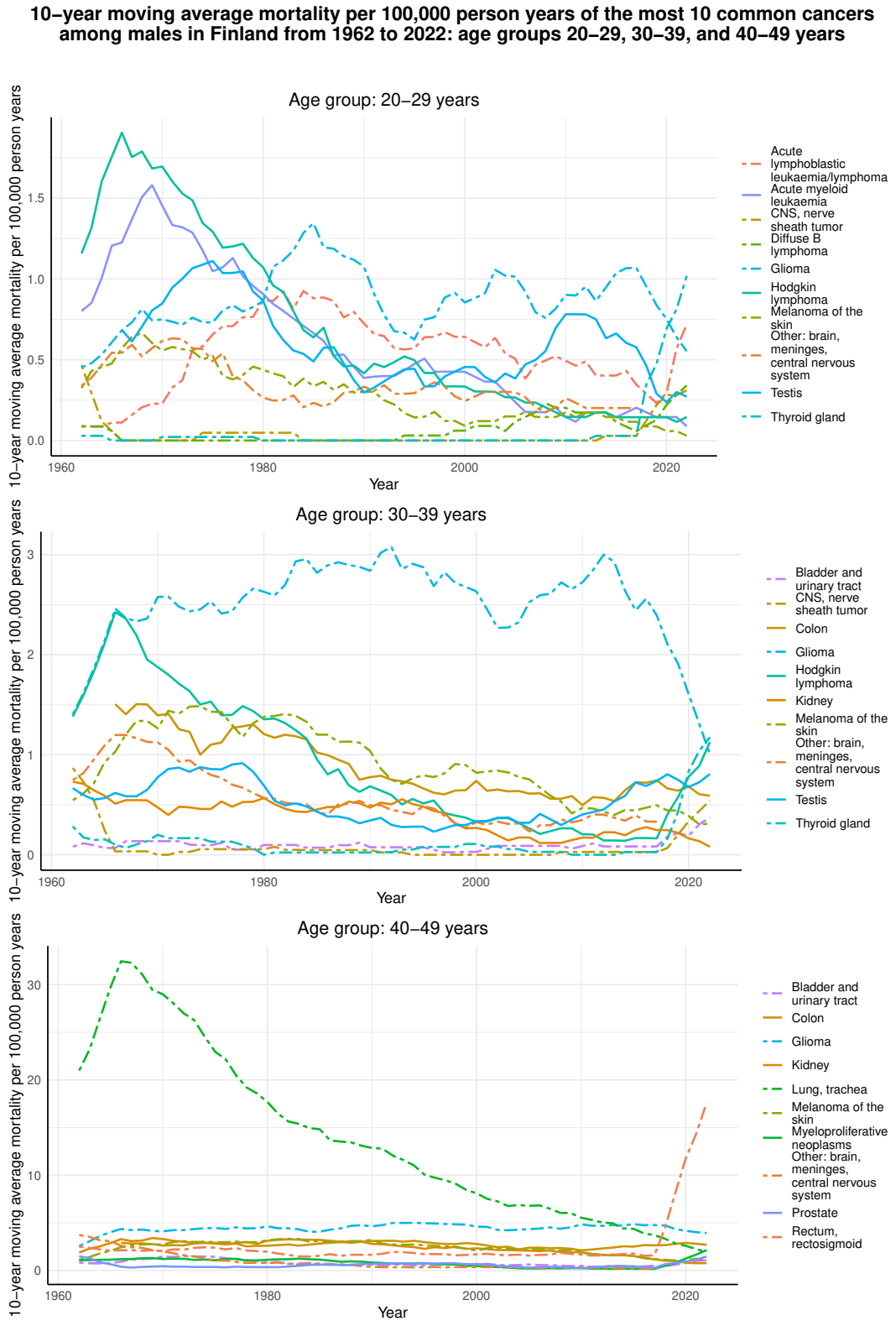


Figure A7: 10-year moving average mortality per 100,000 person years of the most common cancers among males in Finland from 1962 to 2022 for age the groups 20-29, 30-39, and 40-49 years.

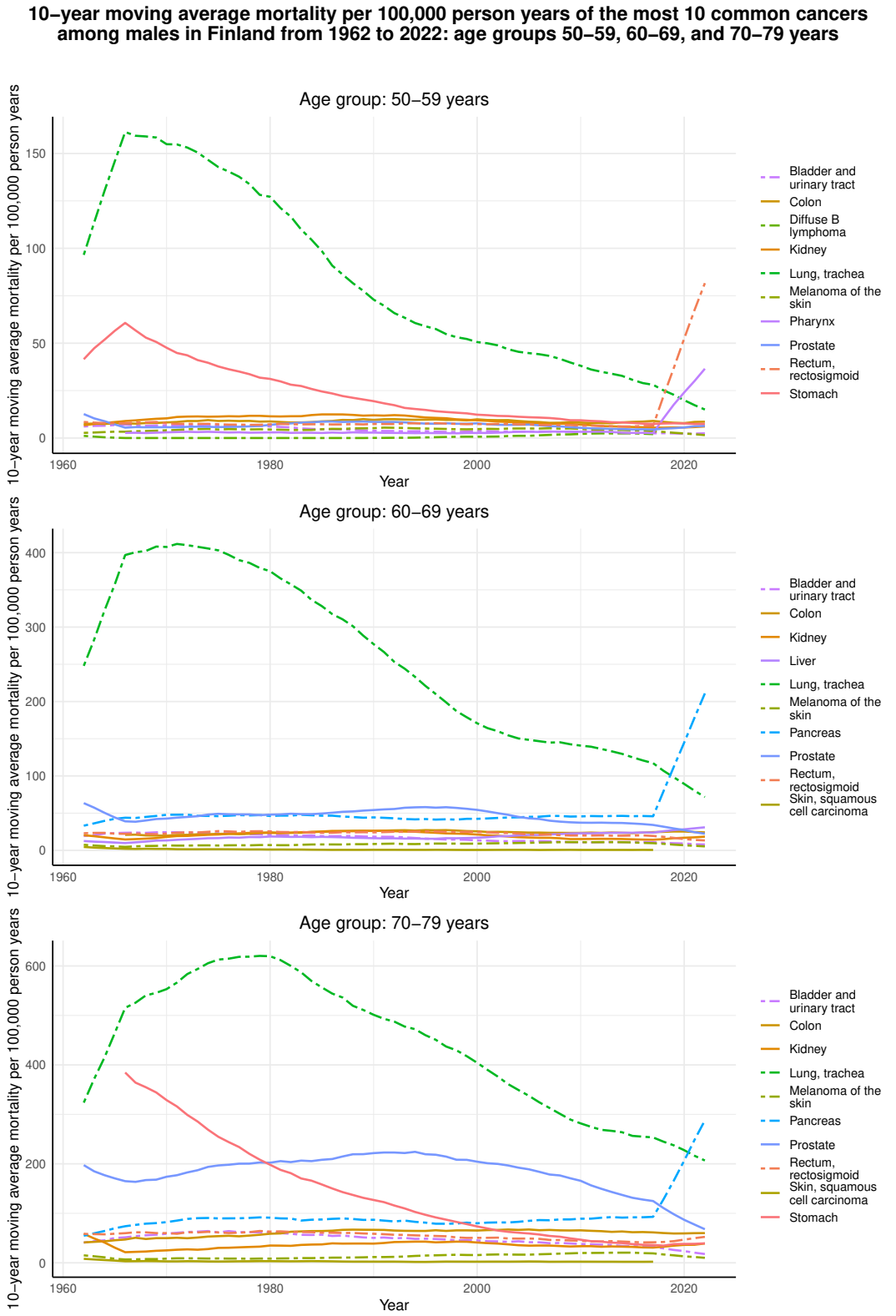


Figure A8: 10-year moving average mortality per 100,000 person years of the most common cancers among males in Finland from 1962 to 2022 for the age groups 50-59, 60-69, and 70-79 years.

B Agglomerative hierarchical clustering applied to the cancer incidence and mortality data over time in Finland

This Appendix contains Figures B1 - B22 showing the clustering results for both original and standardized cancer incidence rates per 100,000 person years of the most common cancers in Finland from 1962 to 2022. The age groups included are from 30-39 to 70-79 years for females and from 20-29 to 70-79 years for males, while the clustering results for the females aged 20-29 years are presented by Figures 11 and 12 in Section 5.1. In each case, the resulting cluster structures are illustrated for 2, 3, and 4 clusters. In addition, this Appendix contains Figures B23 - B46 that represent the clustering results for cancer mortalities per 100,000 person years of the most common cancers in Finland from 1962 to 2022. Again, the clustering has been performed using both original and standardized data. The results are shown for females and males for the age groups from 20-29 to 70-79 years. Also in the case of the mortality data, the resulting cluster structures are shown for 2, 3, and 4 clusters.

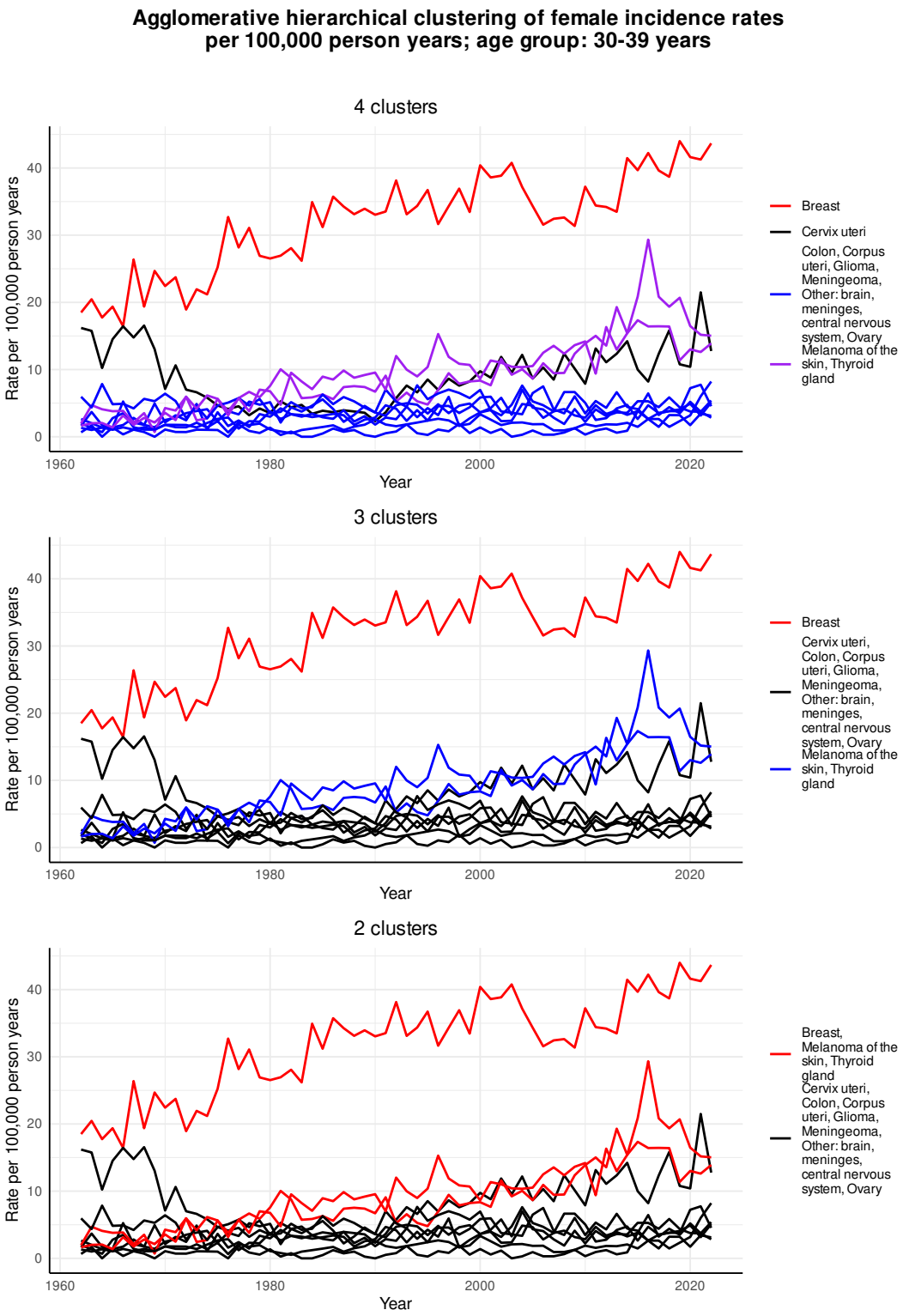


Figure B1: Agglomerative hierarchical clustering applied to the incidence rates per 100,000 person years of the most common cancers among females aged 30-39 years in Finland from 1962 to 2022.

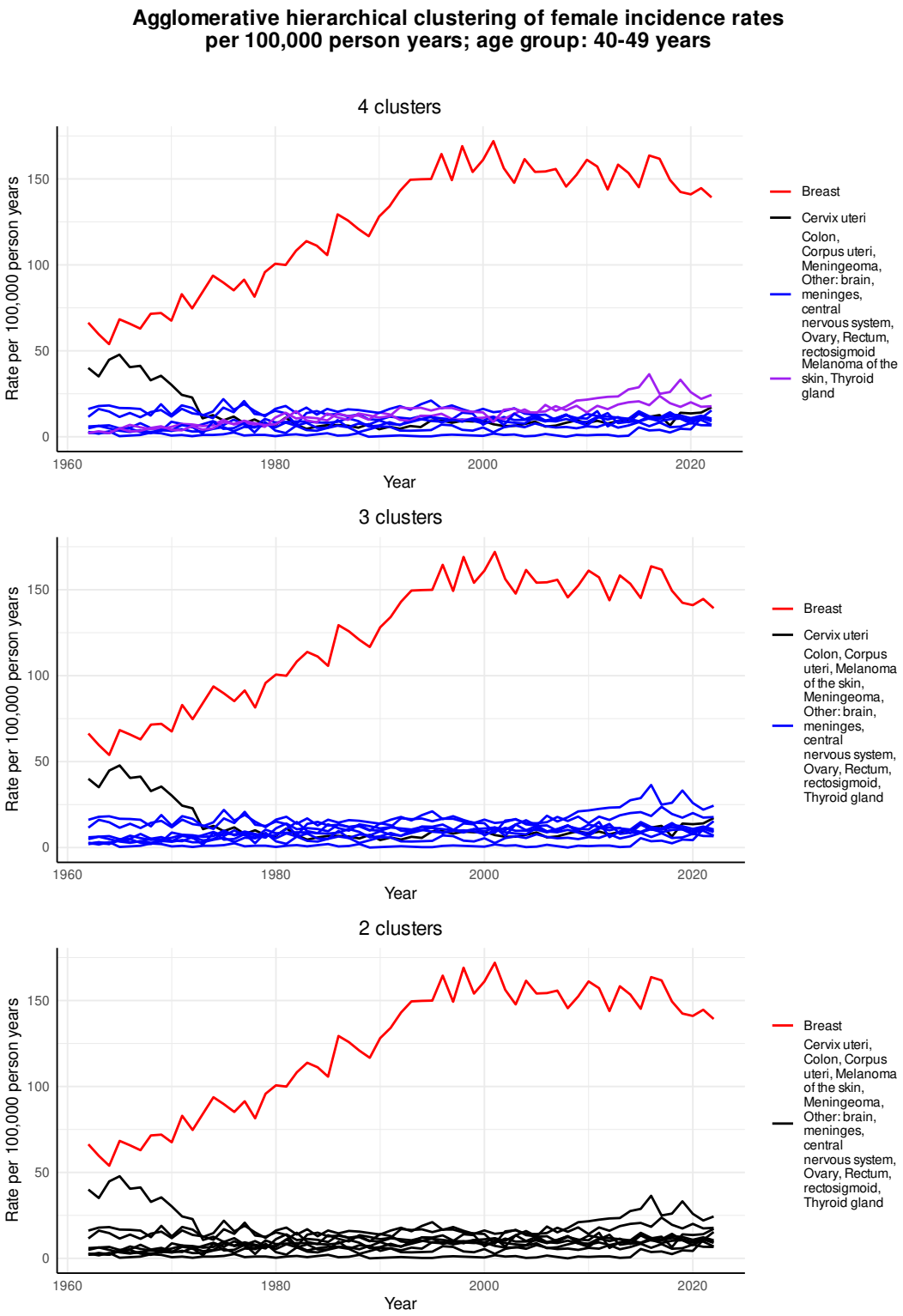


Figure B2: Agglomerative hierarchical clustering applied to the incidence rates per 100,000 person years of the most common cancers among females aged 40-49 years in Finland from 1962 to 2022.

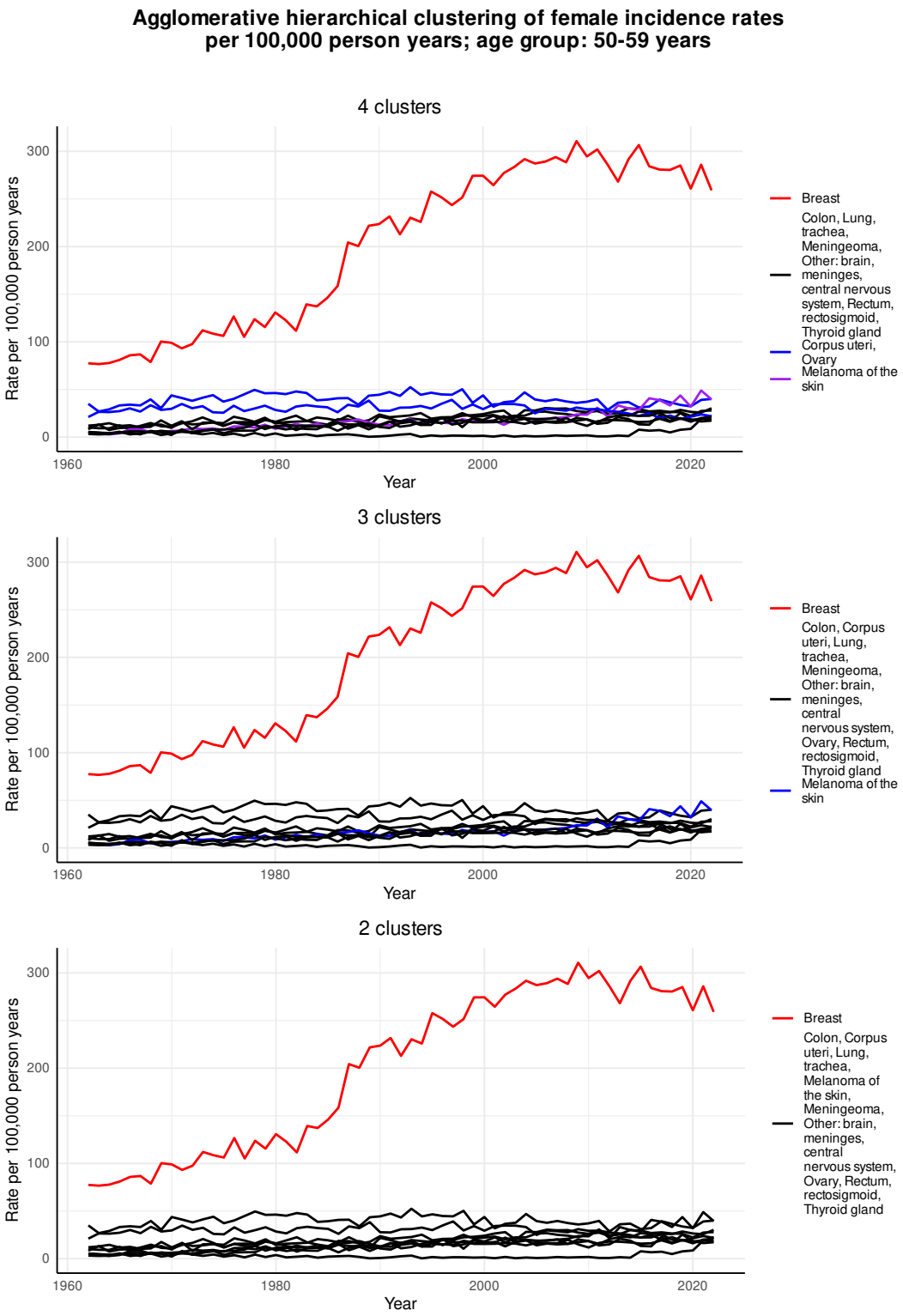


Figure B3: Agglomerative hierarchical clustering applied to the incidence rates per 100,000 person years of the most common cancers among females aged 50-59 years in Finland from 1962 to 2022.

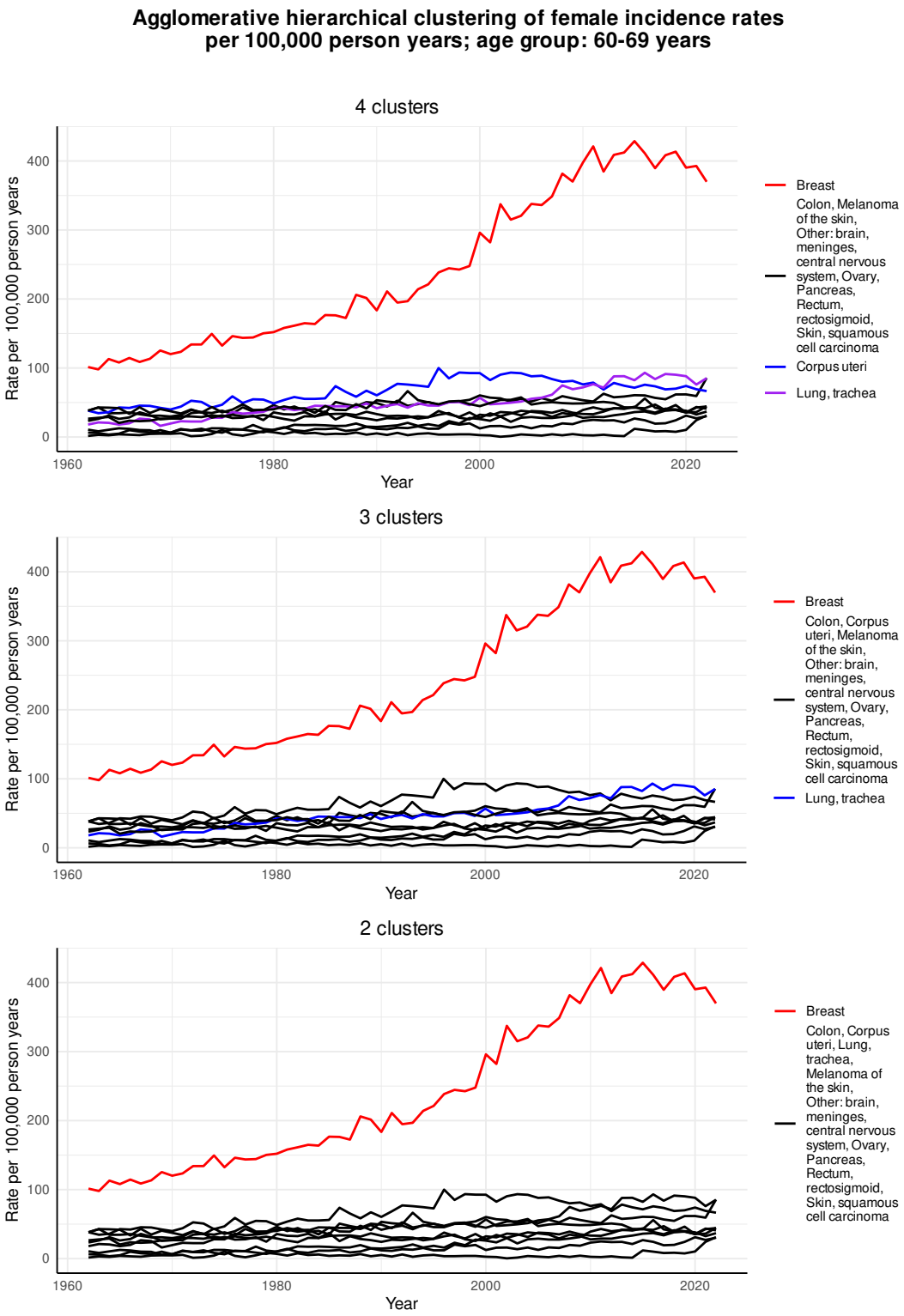


Figure B4: Agglomerative hierarchical clustering applied to the incidence rates per 100,000 person years of the most common cancers among females aged 60-69 years in Finland from 1962 to 2022.

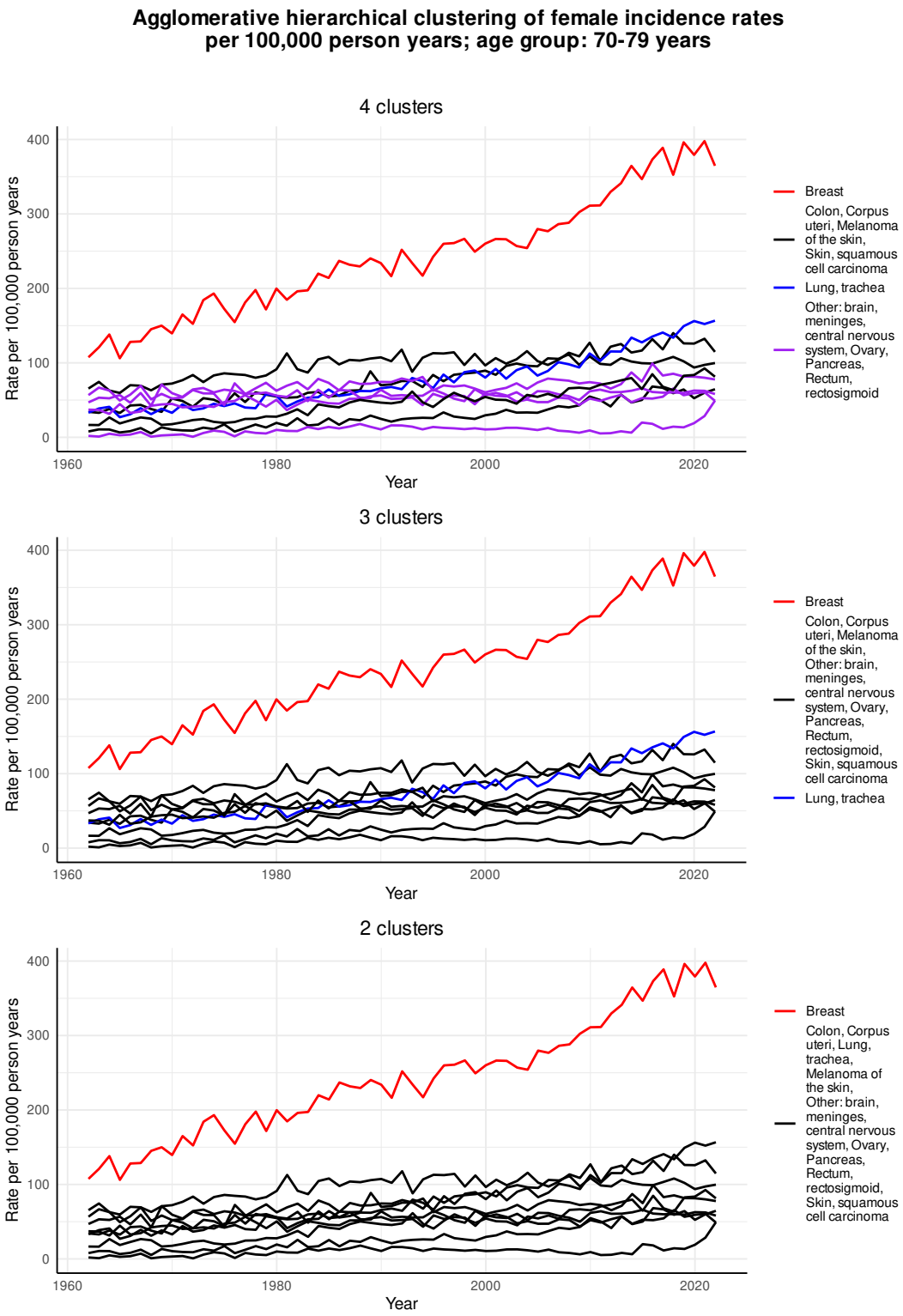


Figure B5: Agglomerative hierarchical clustering applied to the incidence rates per 100,000 person years of the most common cancers among females aged 70-79 years in Finland from 1962 to 2022.

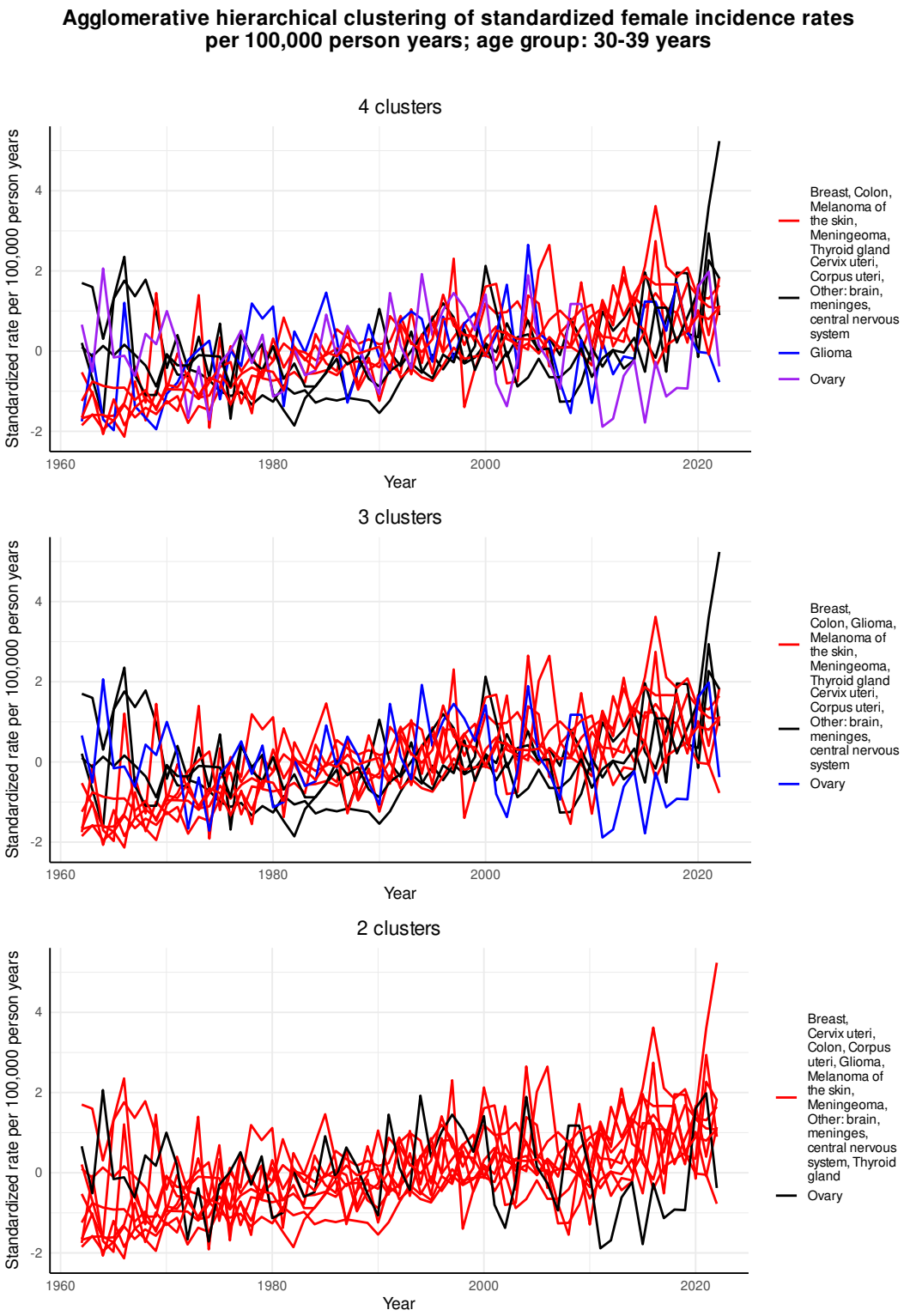


Figure B6: Agglomerative hierarchical clustering applied to the standardized incidence rates per 100,000 person years of the most common cancers among females aged 30-39 years in Finland from 1962 to 2022.

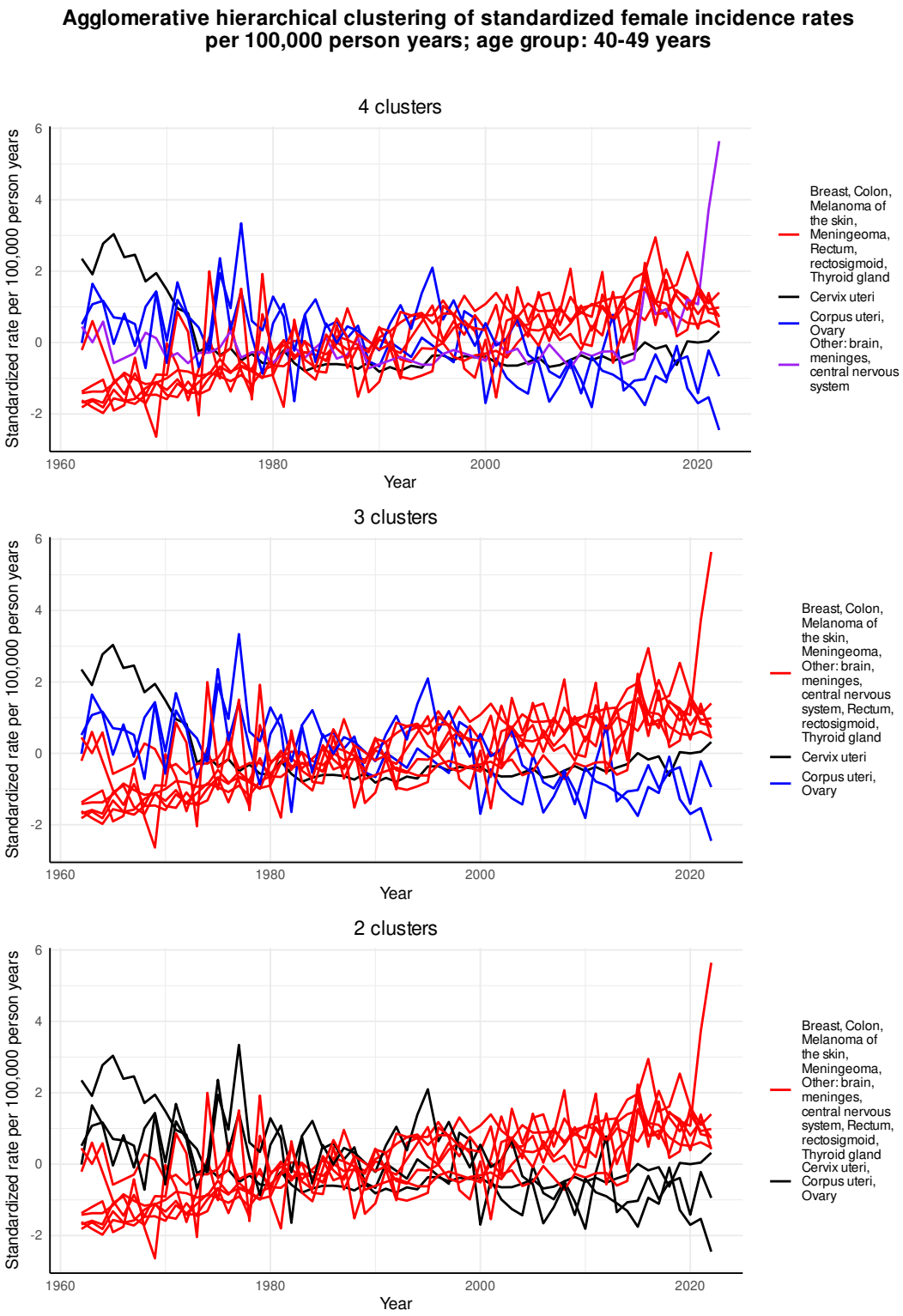


Figure B7: Agglomerative hierarchical clustering applied to the standardized incidence rates per 100,000 person years of the most common cancers among females aged 40-49 years in Finland from 1962 to 2022.

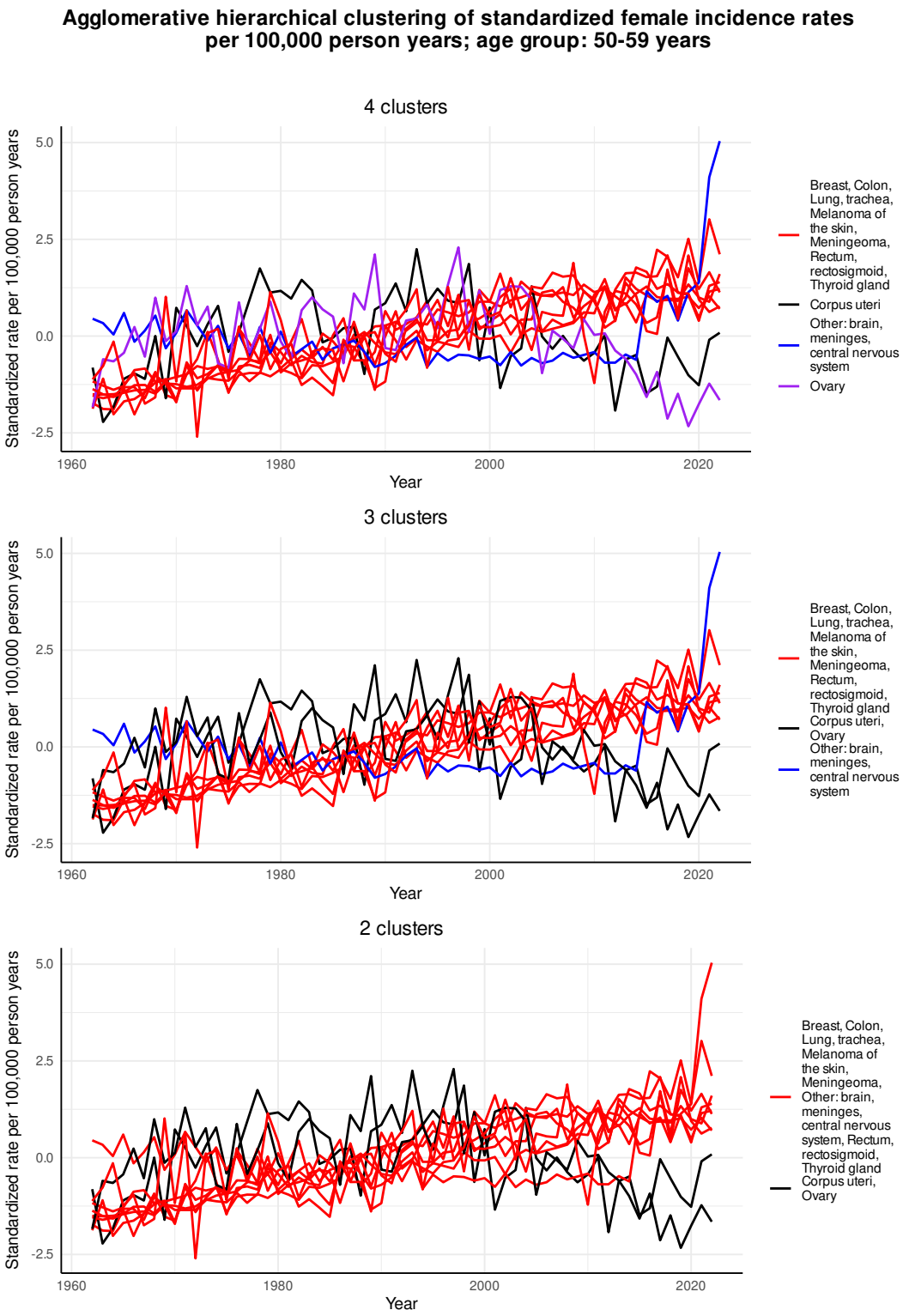


Figure B8: Agglomerative hierarchical clustering applied to the standardized incidence rates per 100,000 person years of the most common cancers among females aged 50-59 years in Finland from 1962 to 2022.

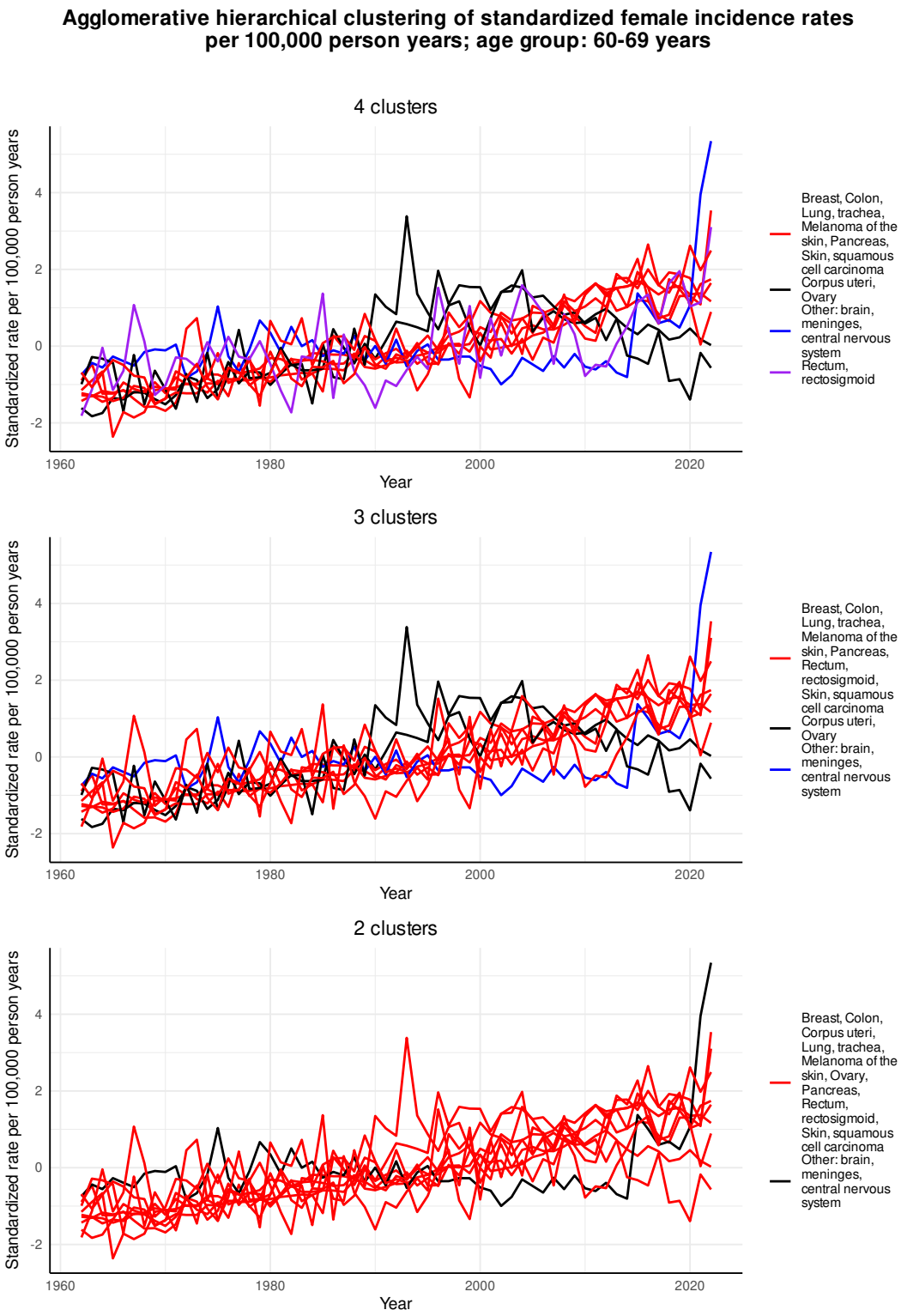


Figure B9: Agglomerative hierarchical clustering applied to the standardized incidence rates per 100,000 person years of the most common cancers among females aged 60-69 years in Finland from 1962 to 2022.

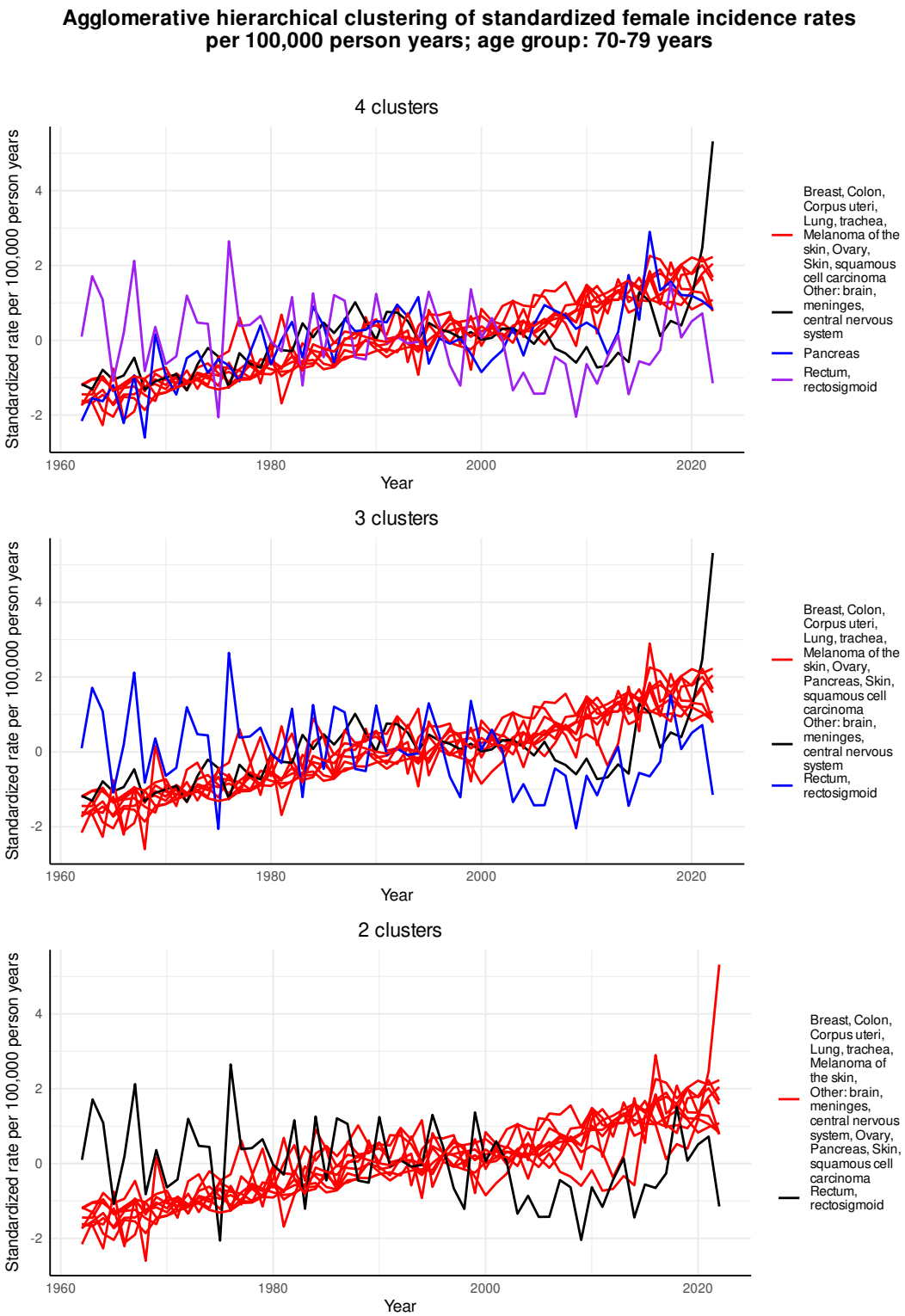


Figure B10: Agglomerative hierarchical clustering applied to the standardized incidence rates per 100,000 person years of the most common cancers among females aged 70-79 years in Finland from 1962 to 2022.

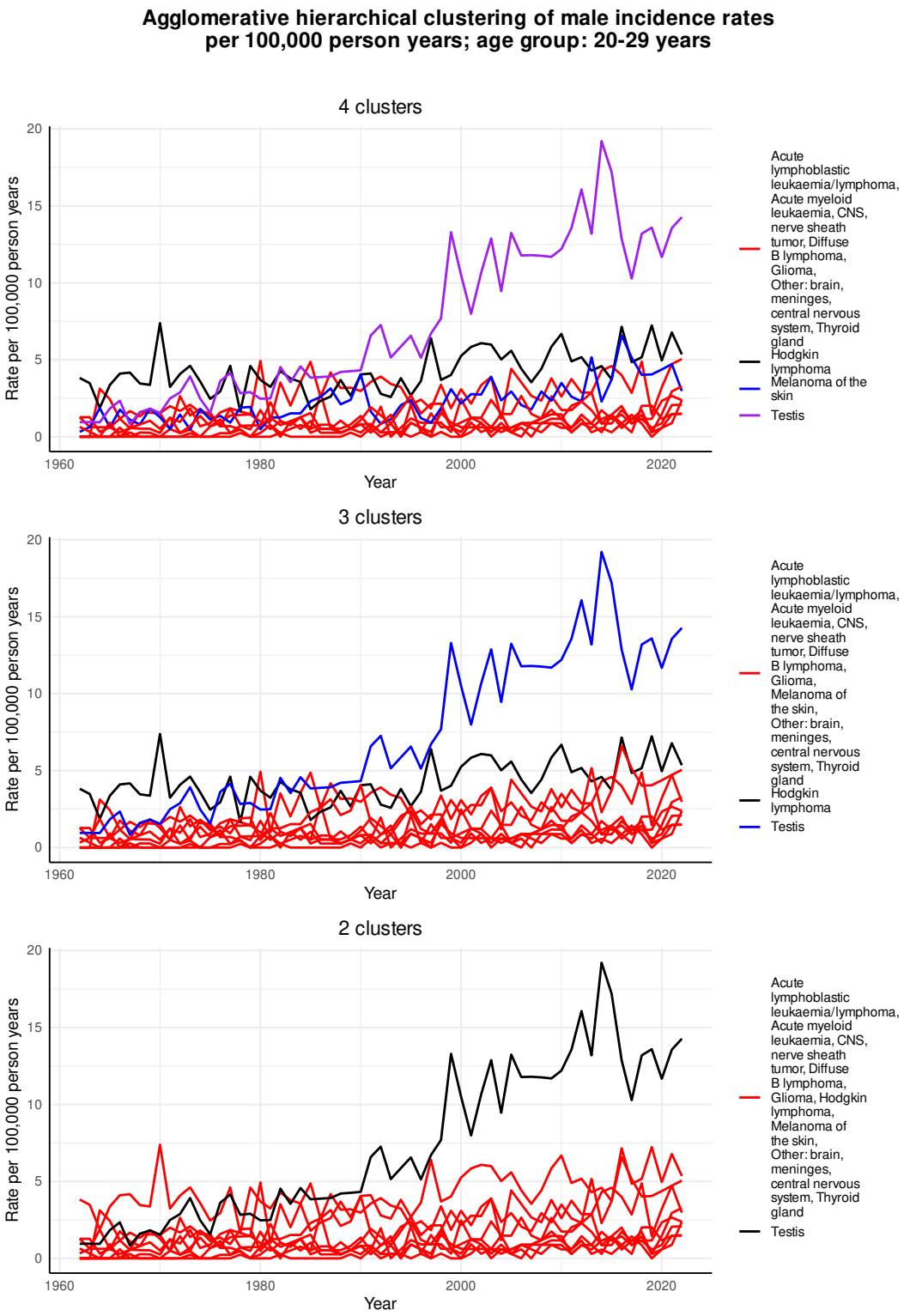


Figure B11: Agglomerative hierarchical clustering applied to the incidence rates per 100,000 person years of the most common cancers among males aged 20-29 years in Finland from 1962 to 2022.

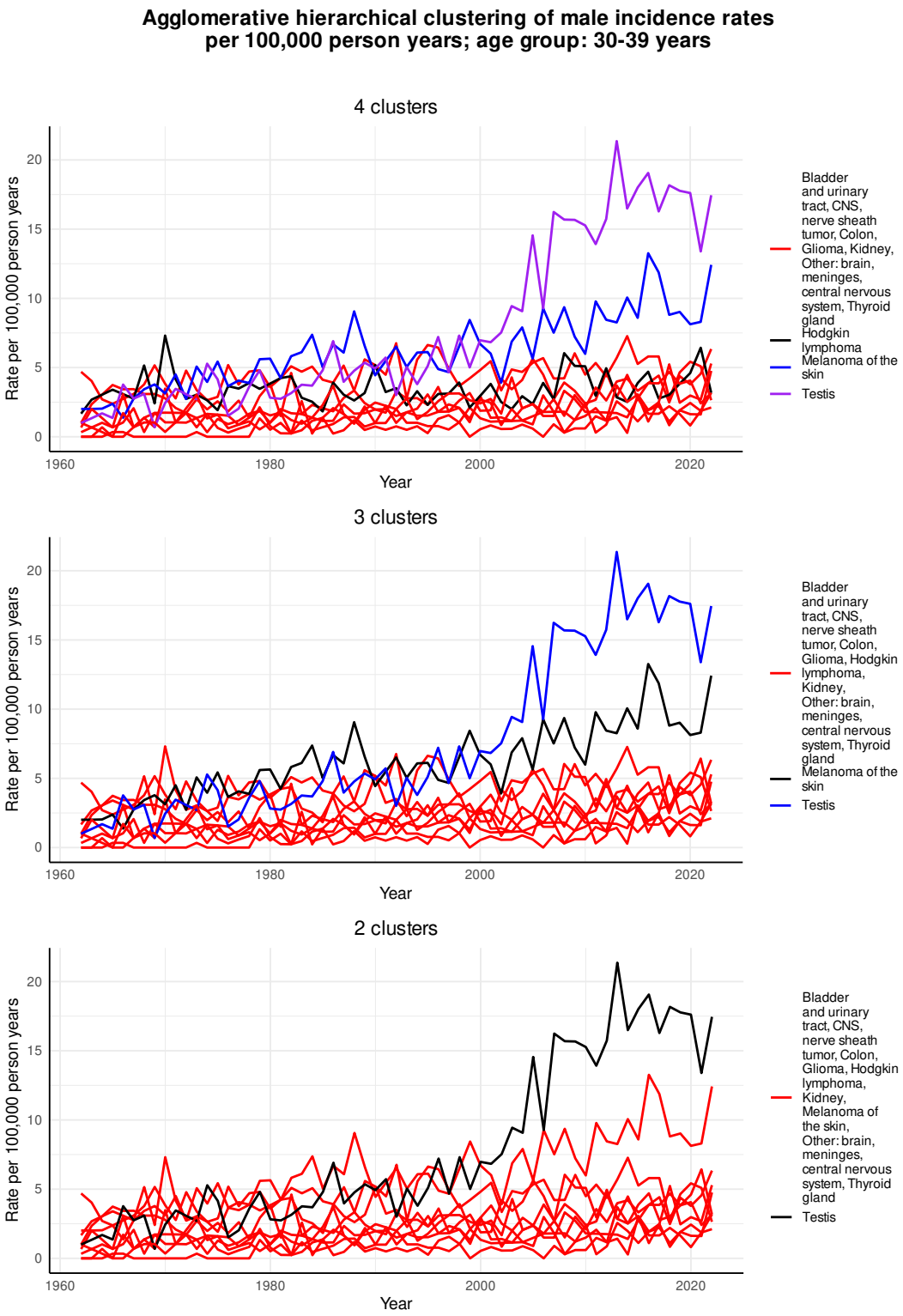


Figure B12: Agglomerative hierarchical clustering applied to the incidence rates per 100,000 person years of the most common cancers among males aged 30-39 years in Finland from 1962 to 2022.

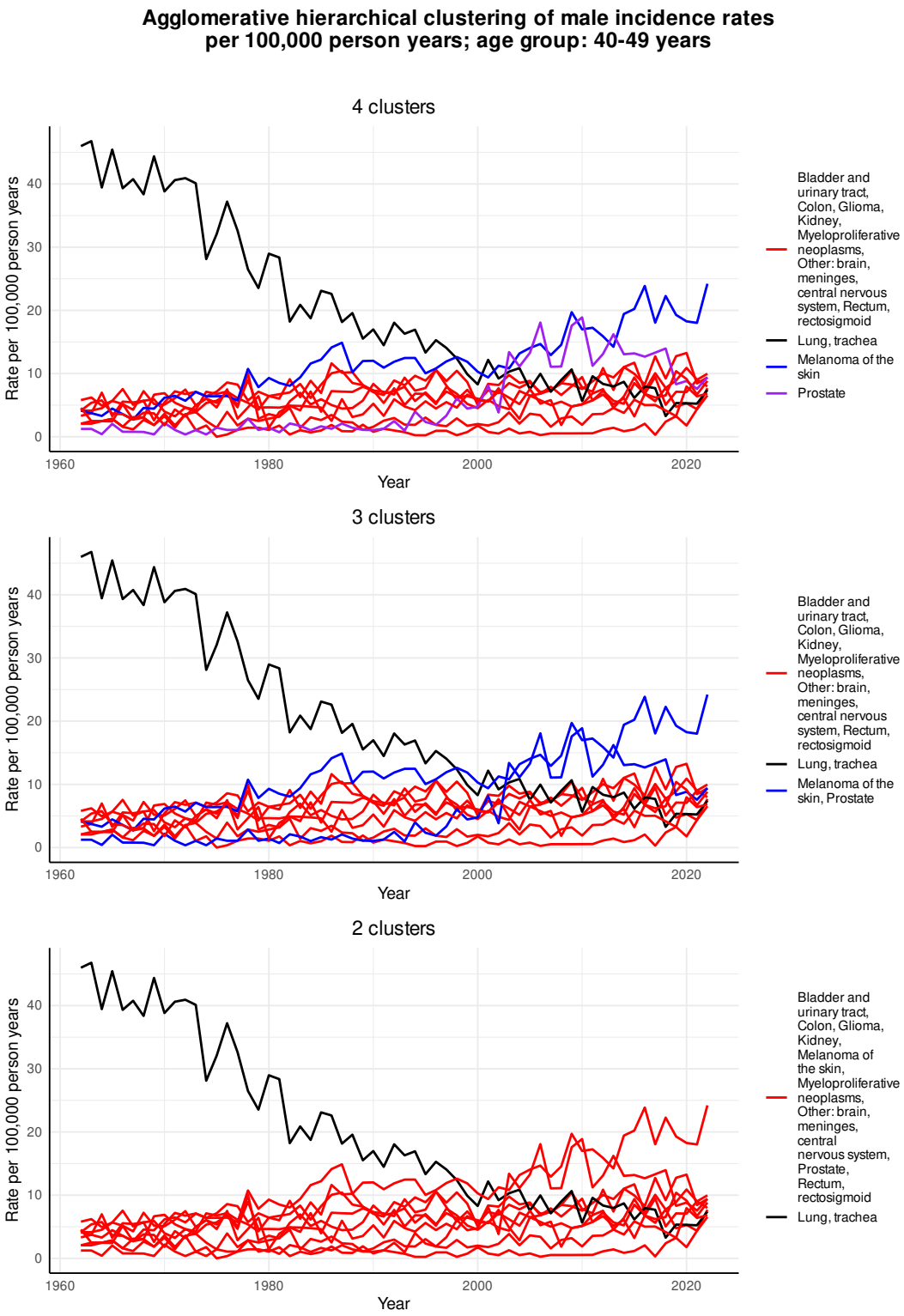


Figure B13: Agglomerative hierarchical clustering applied to the incidence rates per 100,000 person years of the most common cancers among males aged 40-49 years in Finland from 1962 to 2022.

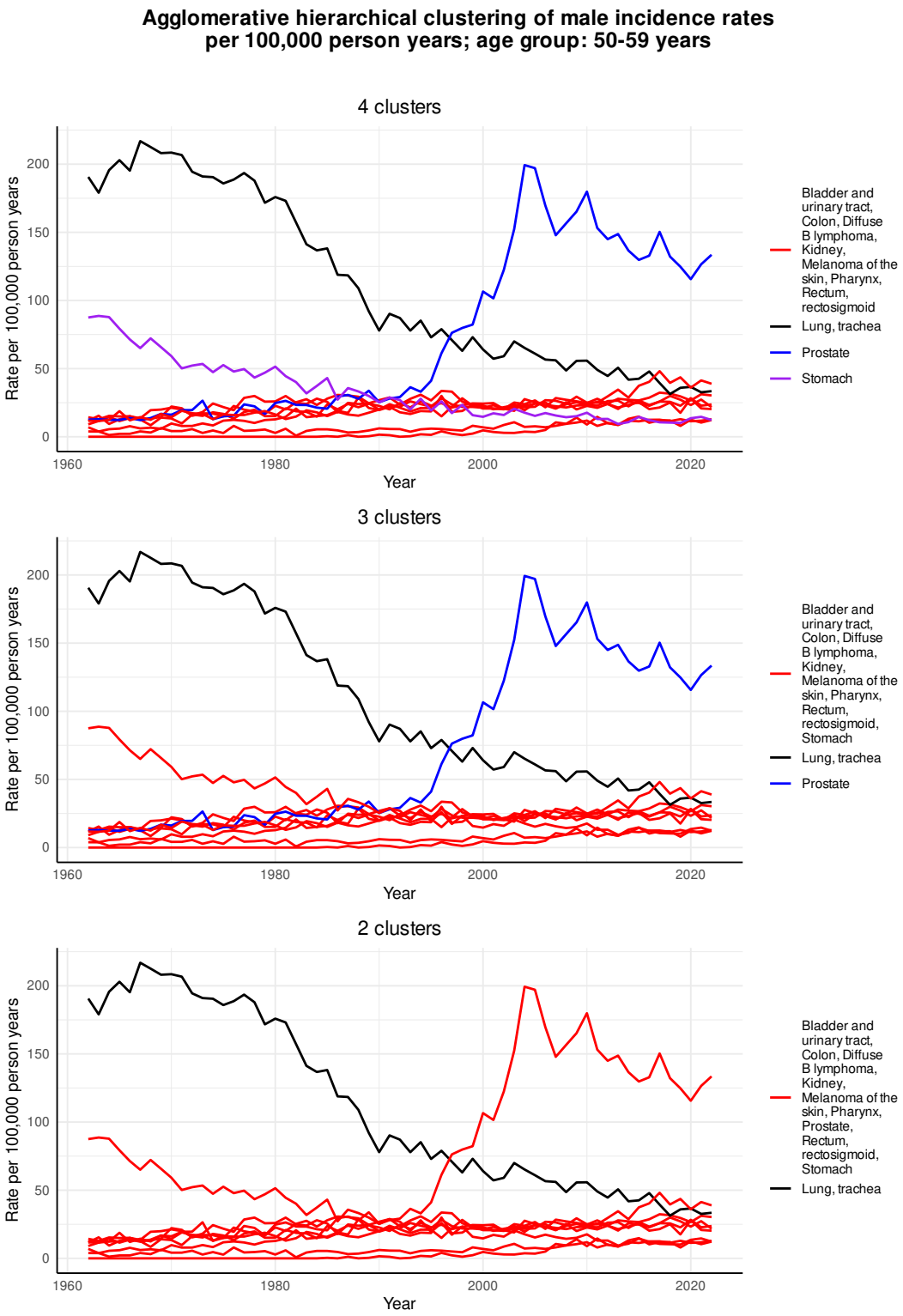


Figure B14: Agglomerative hierarchical clustering applied to the incidence rates per 100,000 person years of the most common cancers among males aged 50-59 years in Finland from 1962 to 2022.

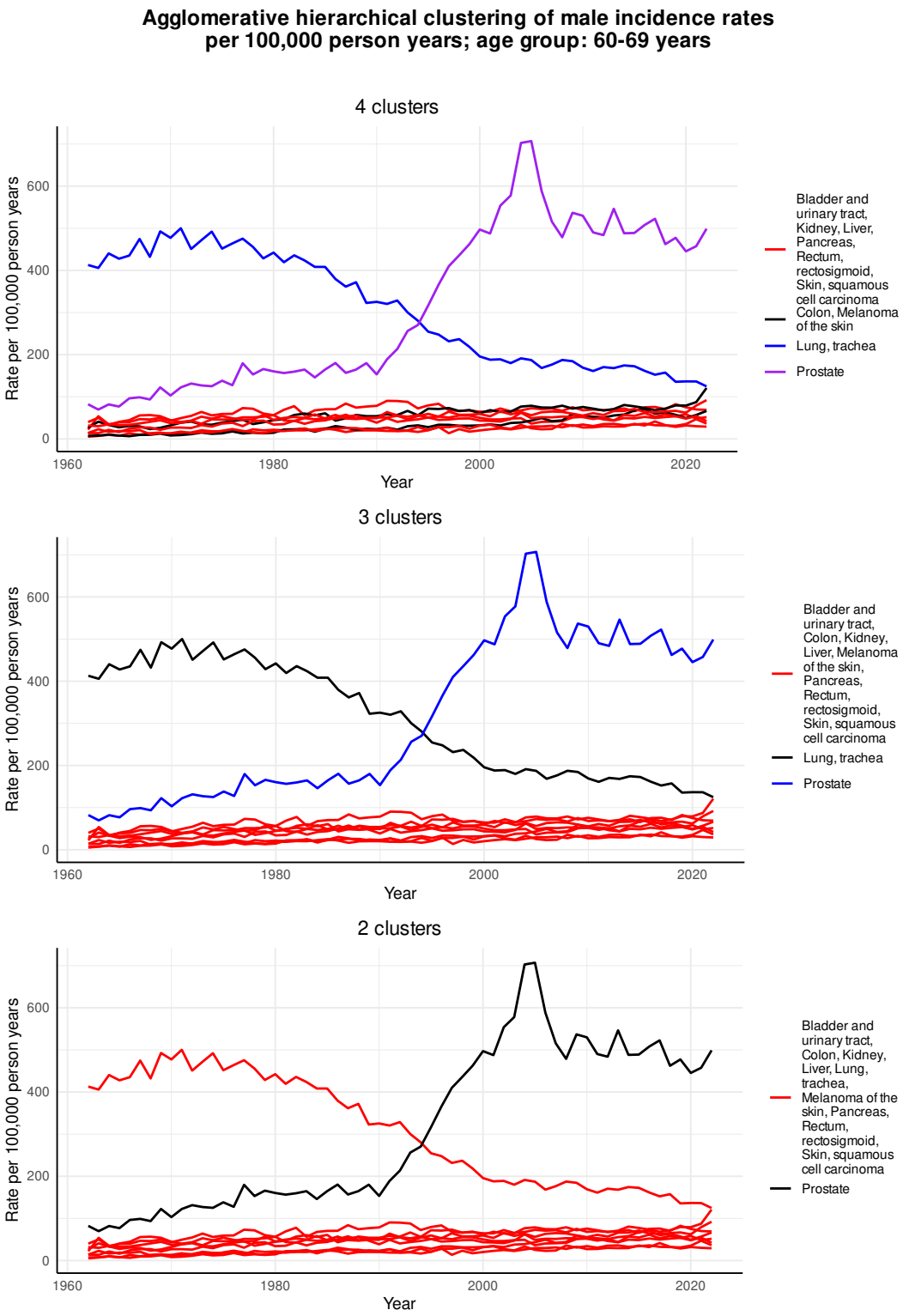


Figure B15: Agglomerative hierarchical clustering applied to the incidence rates per 100,000 person years of the most common cancers among males aged 60-69 years in Finland from 1962 to 2022.

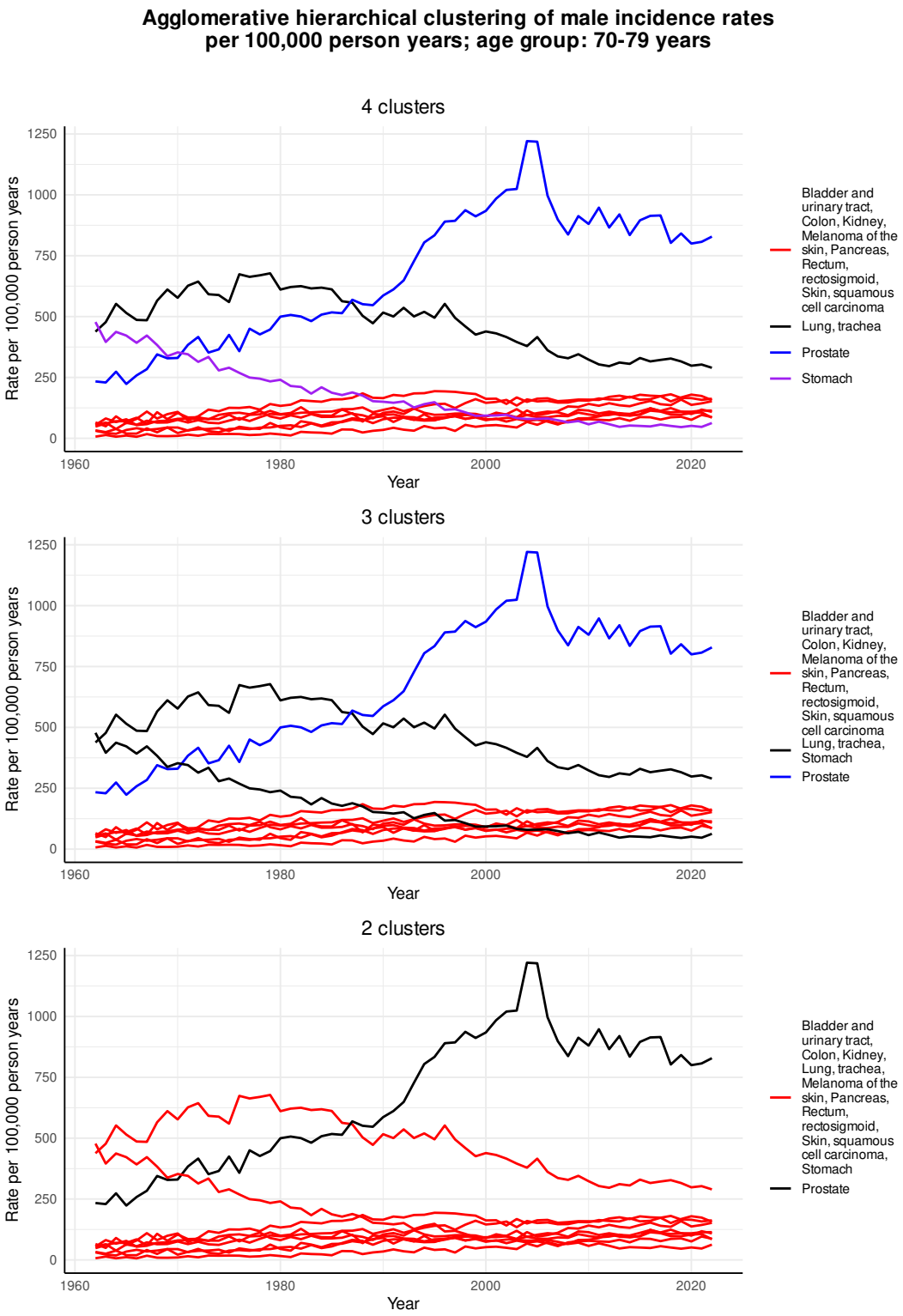


Figure B16: Agglomerative hierarchical clustering applied to the incidence rates per 100,000 person years of the most common cancers among males aged 70-79 years in Finland from 1962 to 2022.

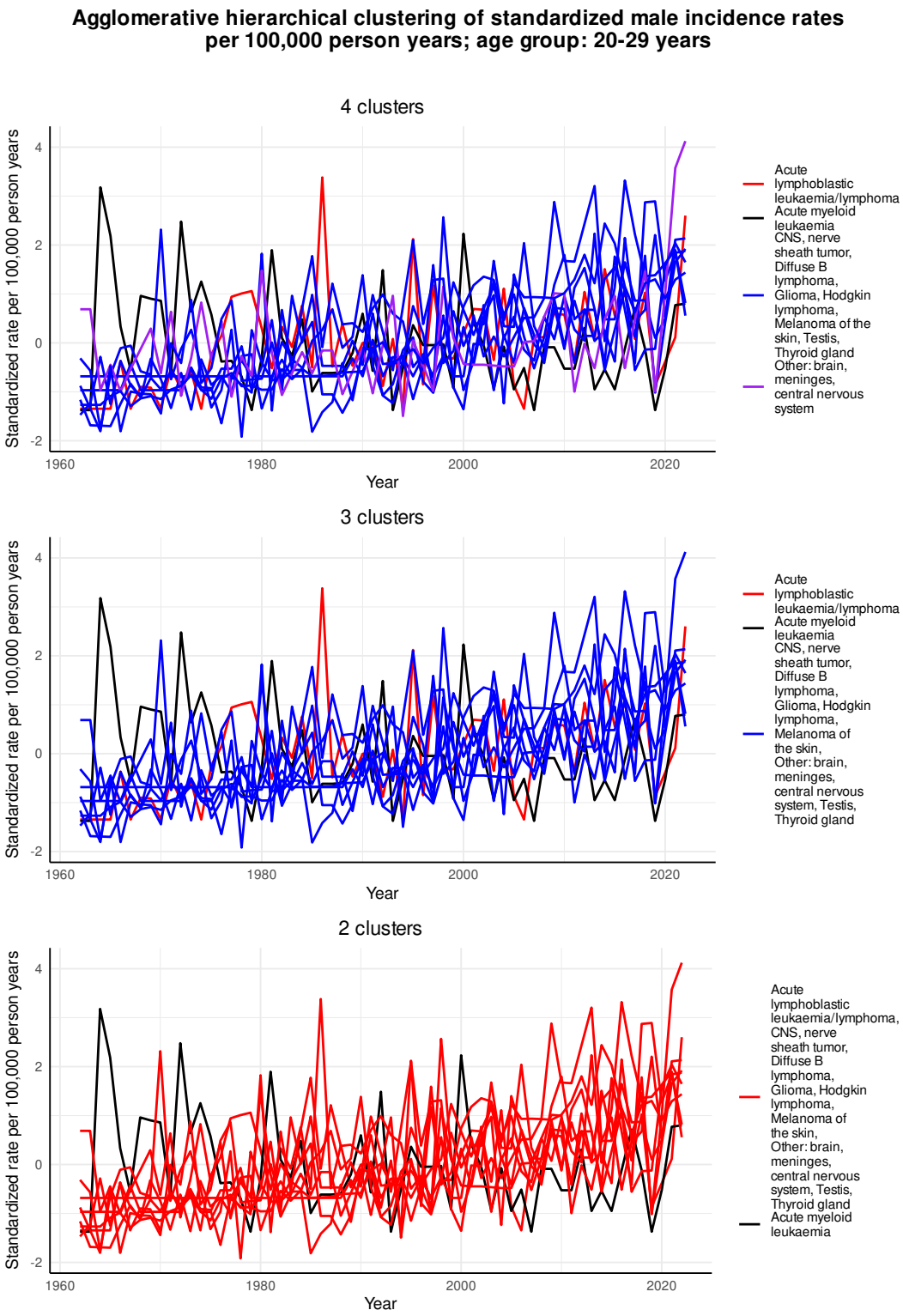


Figure B17: Agglomerative hierarchical clustering applied to the standardized incidence rates per 100,000 person years of the most common cancers among males aged 20-29 years in Finland from 1962 to 2022.

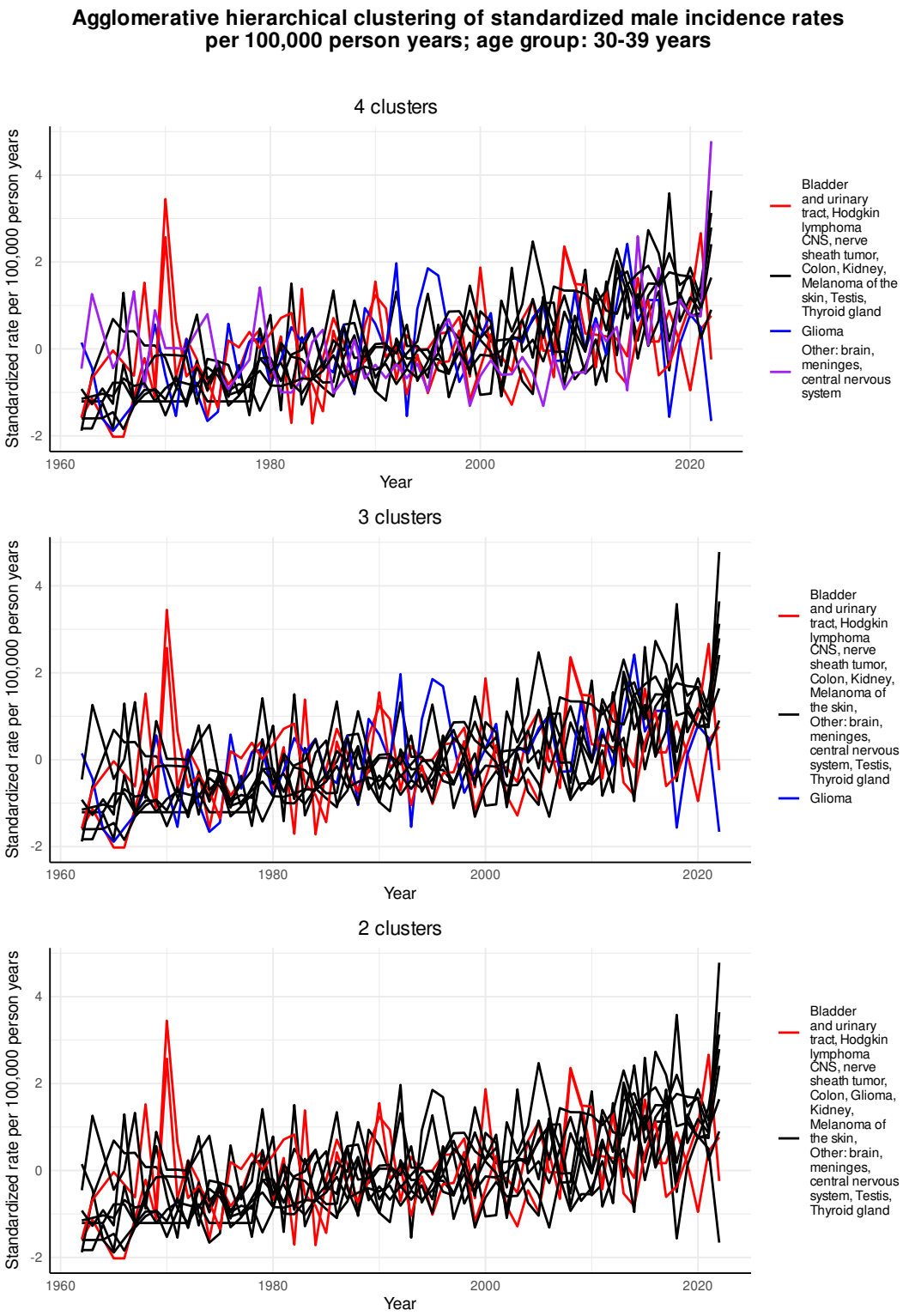


Figure B18: Agglomerative hierarchical clustering applied to the standardized incidence rates per 100,000 person years of the most common cancers among males aged 30-39 years in Finland from 1962 to 2022.

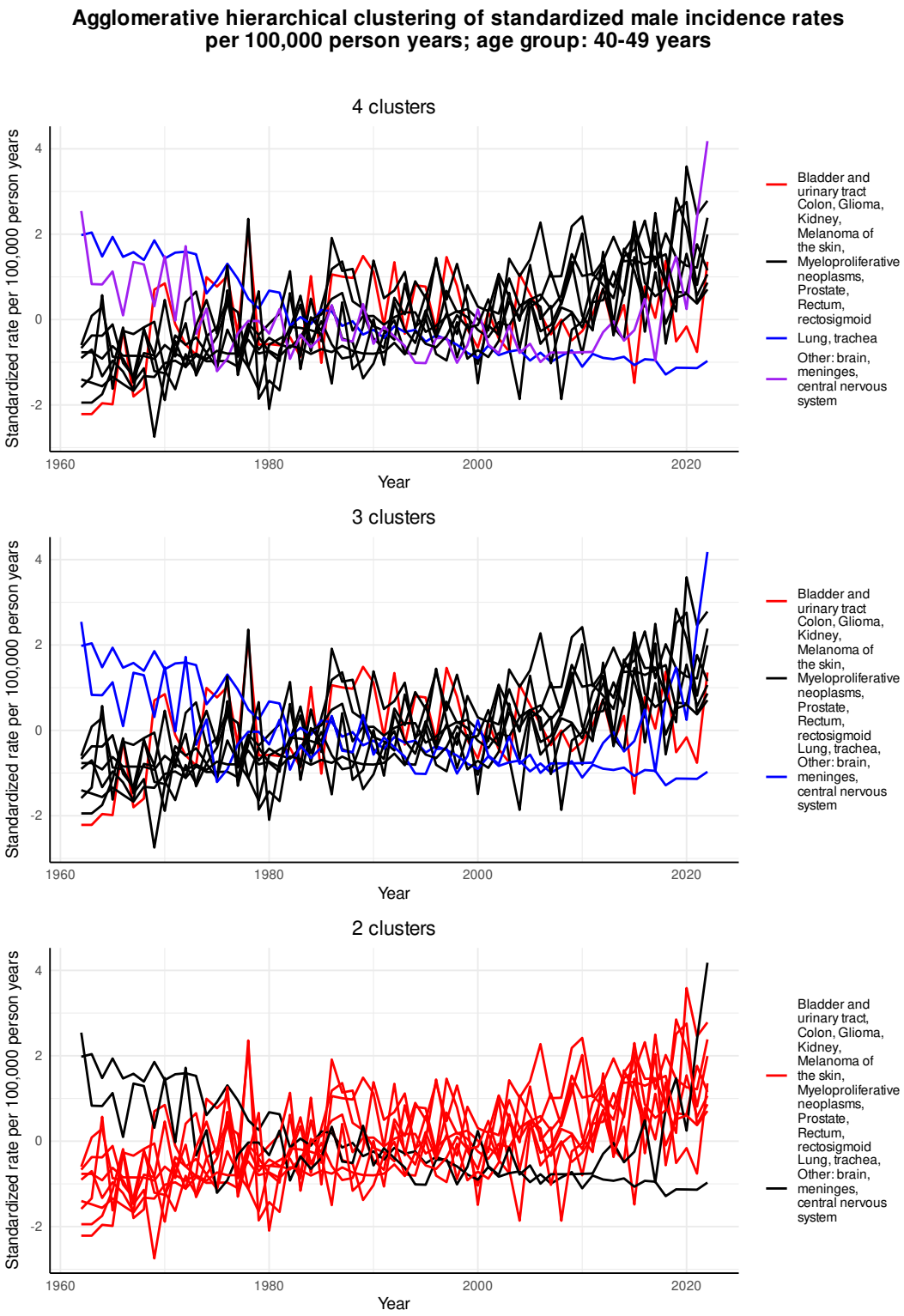


Figure B19: Agglomerative hierarchical clustering applied to the standardized incidence rates per 100,000 person years of the most common cancers among males aged 40-49 years in Finland from 1962 to 2022.

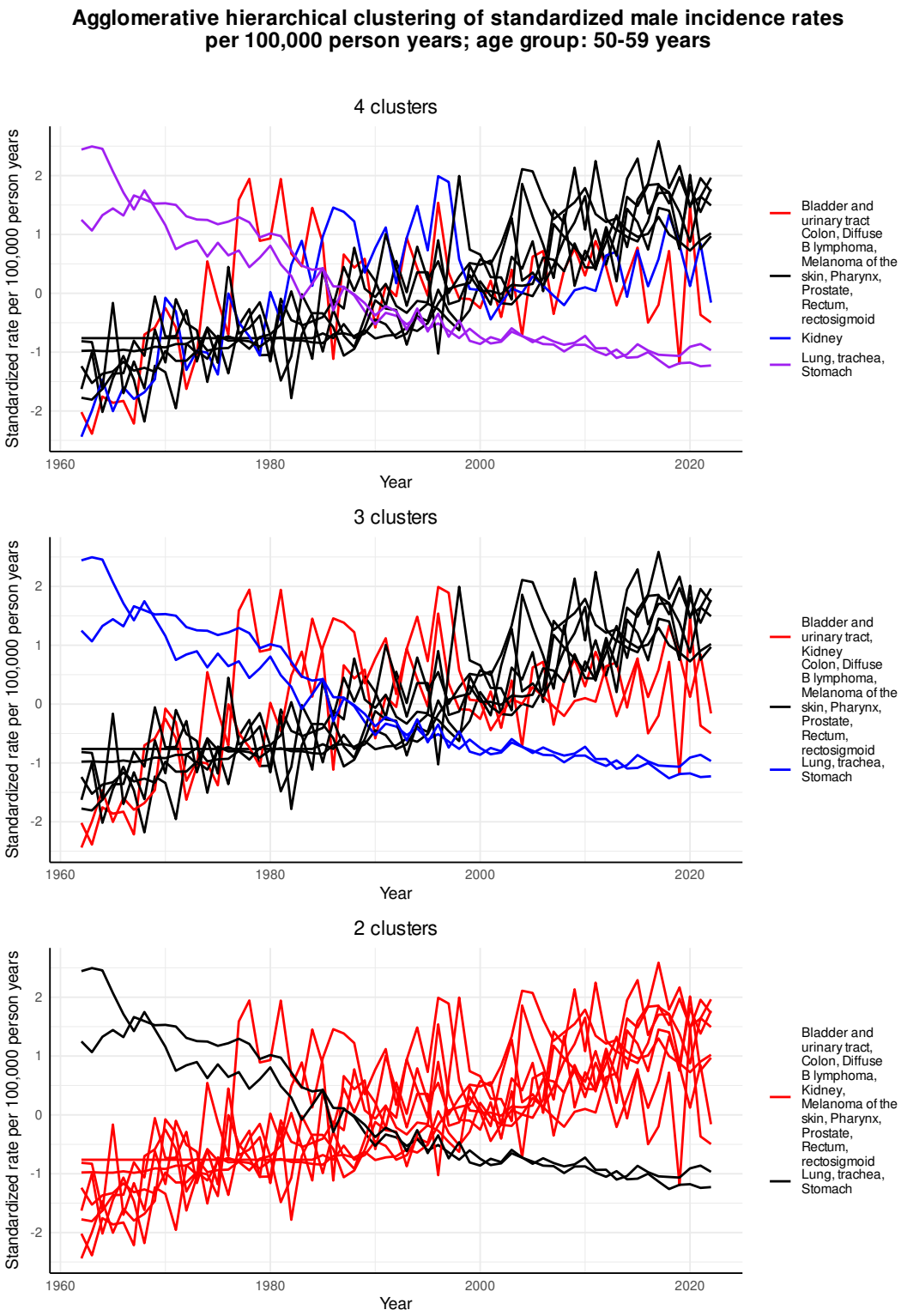


Figure B20: Agglomerative hierarchical clustering applied to the standardized incidence rates per 100,000 person years of the most common cancers among males aged 50-59 years in Finland from 1962 to 2022.

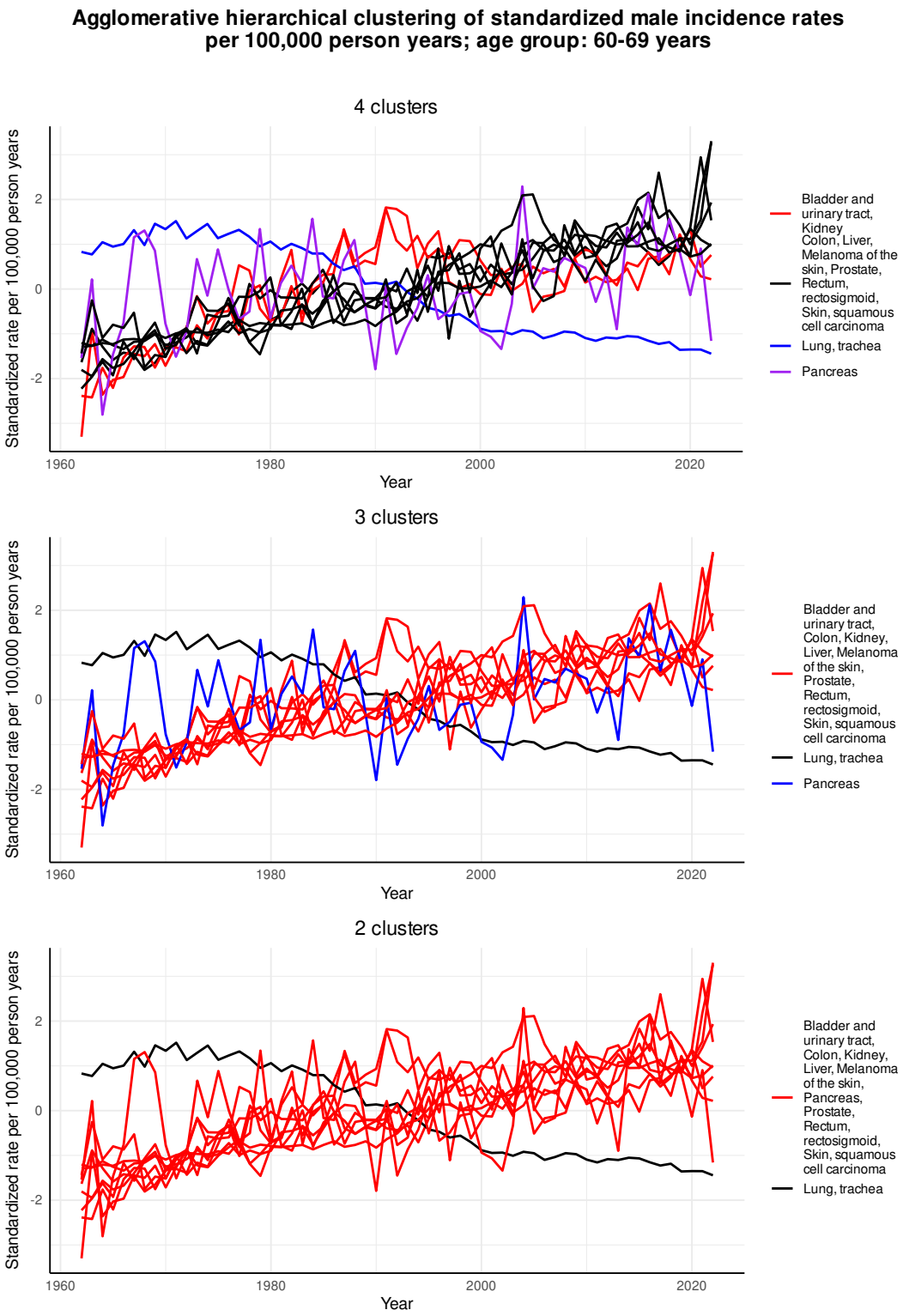


Figure B21: Agglomerative hierarchical clustering applied to the standardized incidence rates per 100,000 person years of the most common cancers among males aged 60-69 years in Finland from 1962 to 2022.

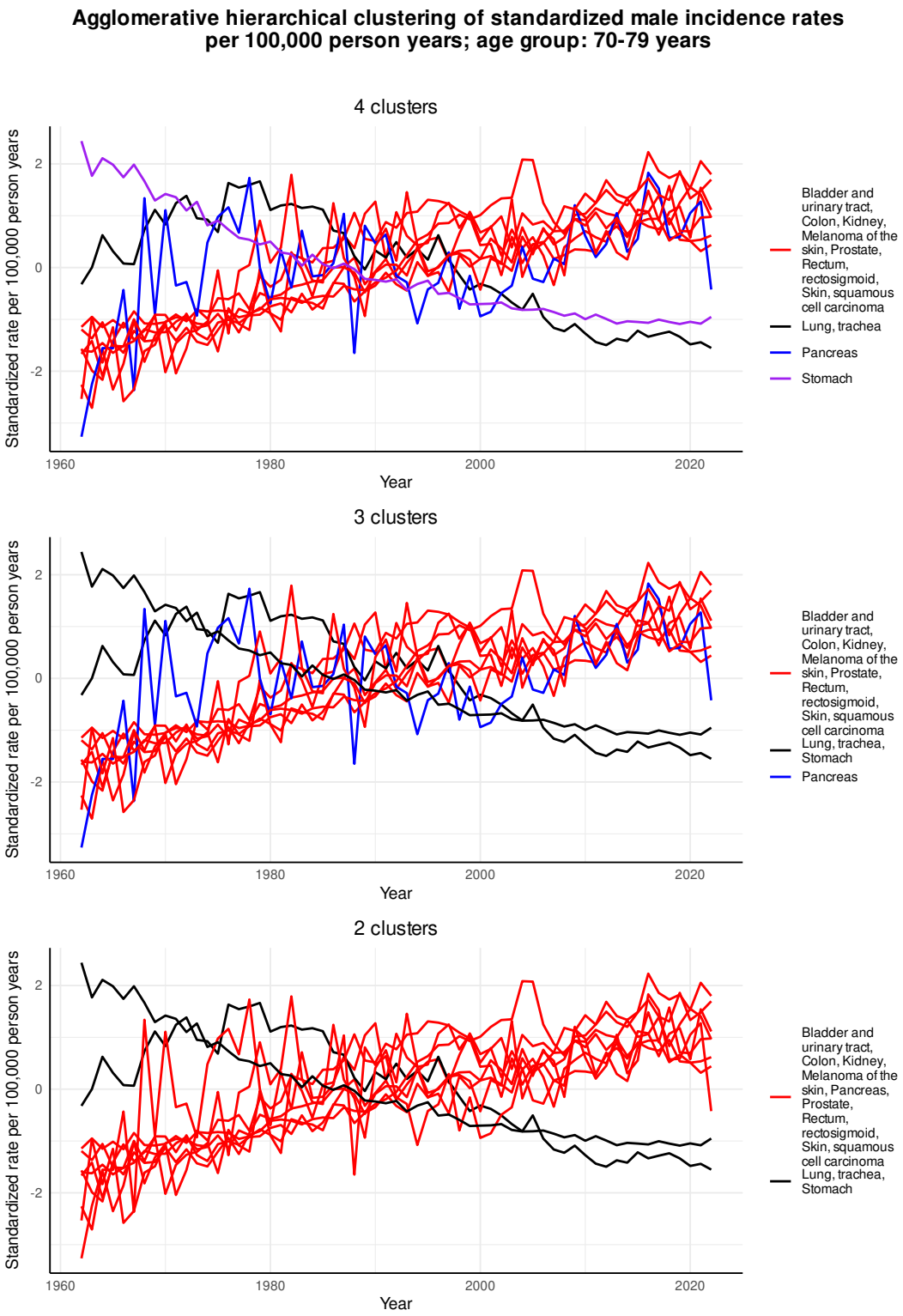


Figure B22: Agglomerative hierarchical clustering applied to the standardized incidence rates per 100,000 person years of the most common cancers among males aged 70-79 years in Finland from 1962 to 2022.

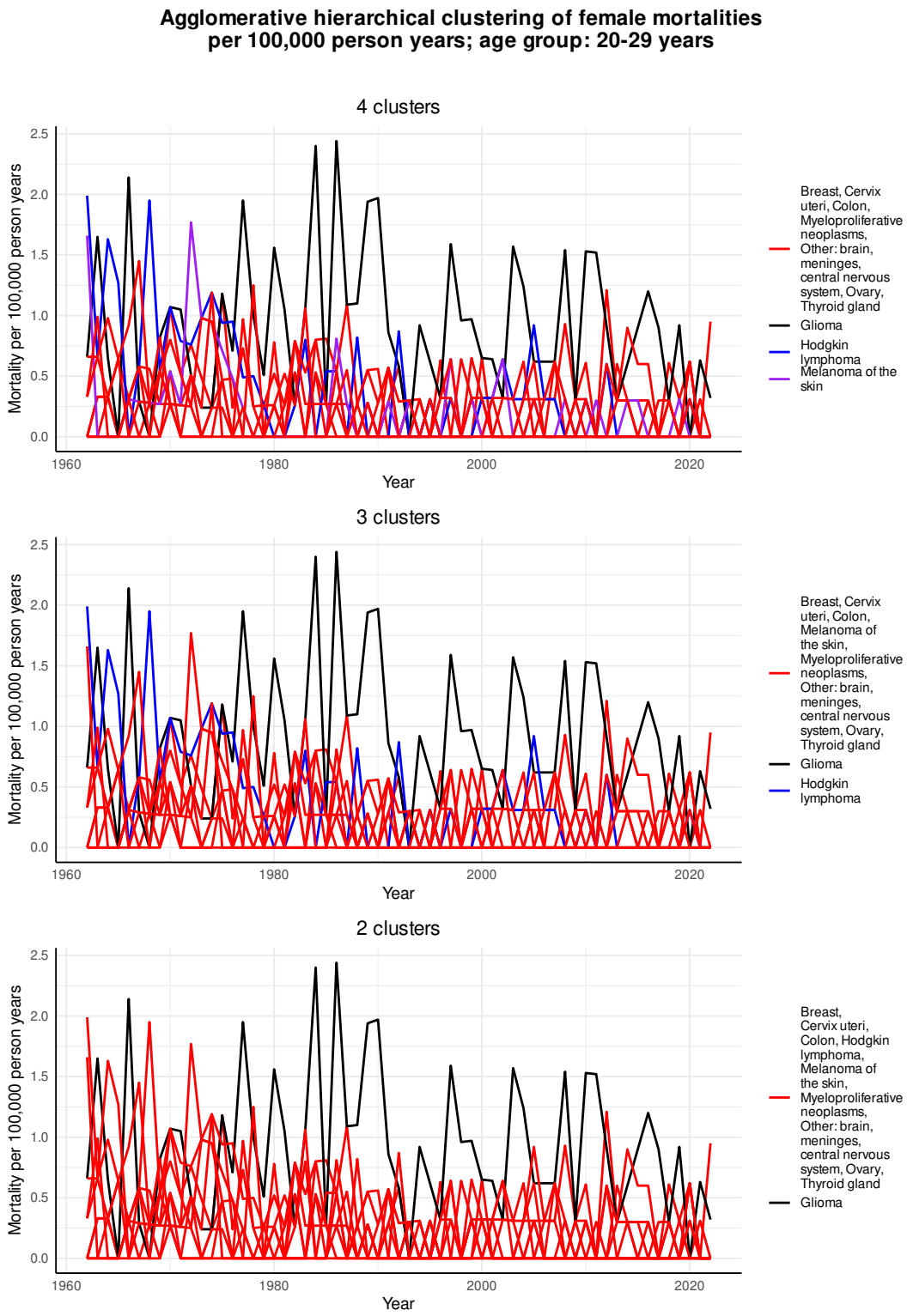


Figure B23: Agglomerative hierarchical clustering applied to the mortalities per 100,000 person years of the most common cancers among females aged 20-29 years in Finland from 1962 to 2022.

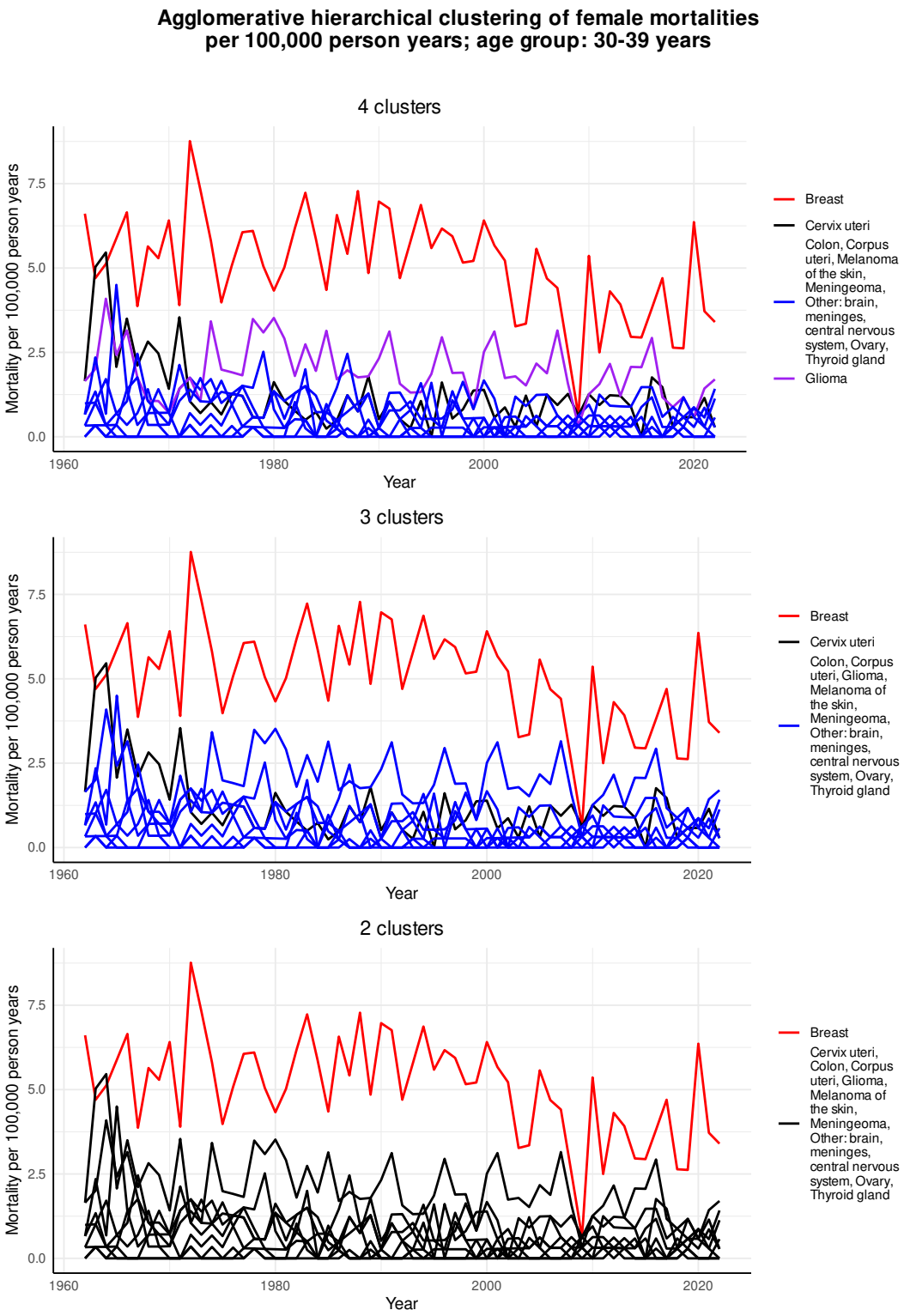


Figure B24: Agglomerative hierarchical clustering applied to the mortalities per 100,000 person years of the most common cancers among females aged 30-39 years in Finland from 1962 to 2022.

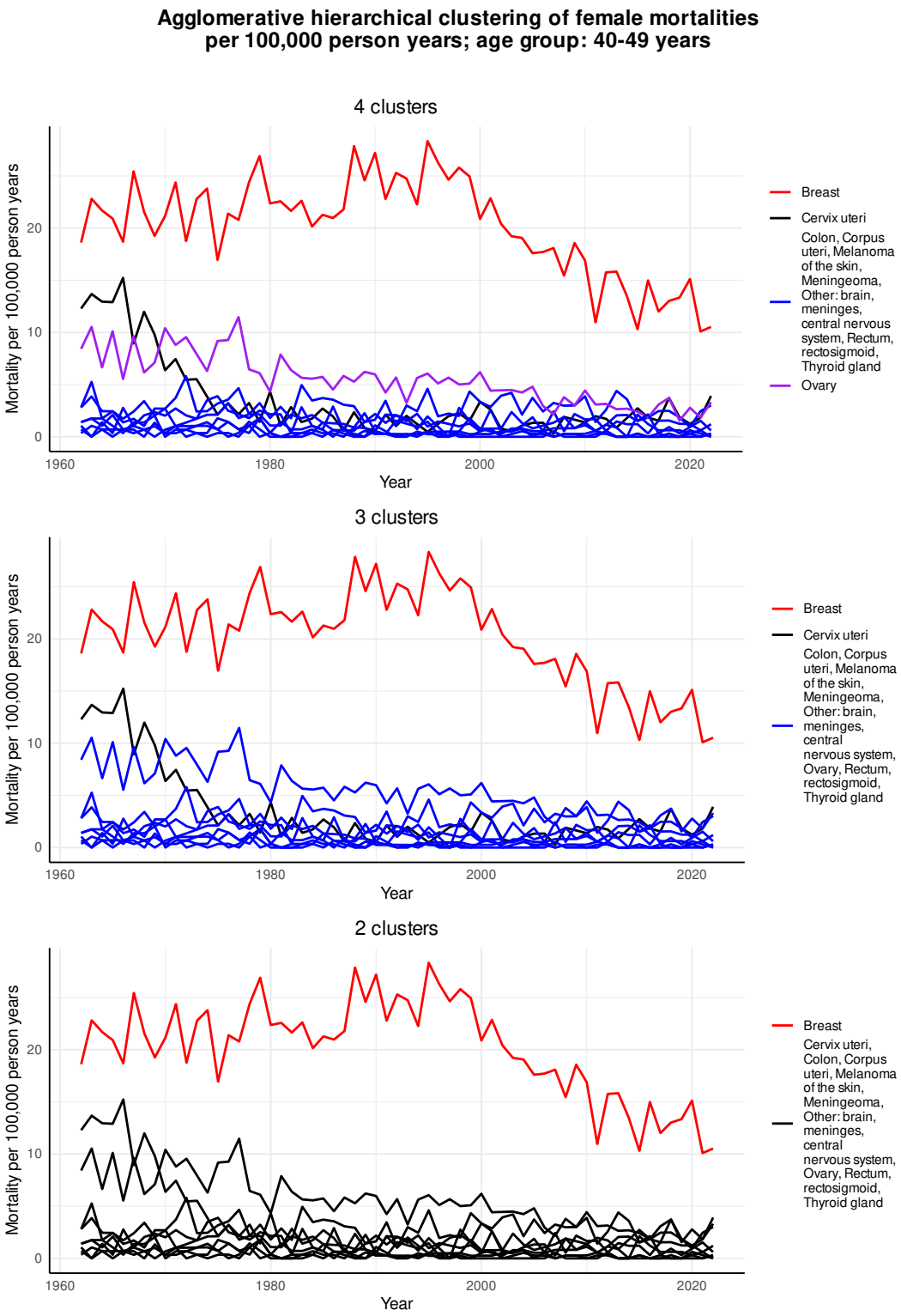


Figure B25: Agglomerative hierarchical clustering applied to the mortalities per 100,000 person years of the most common cancers among females aged 40-49 years in Finland from 1962 to 2022.

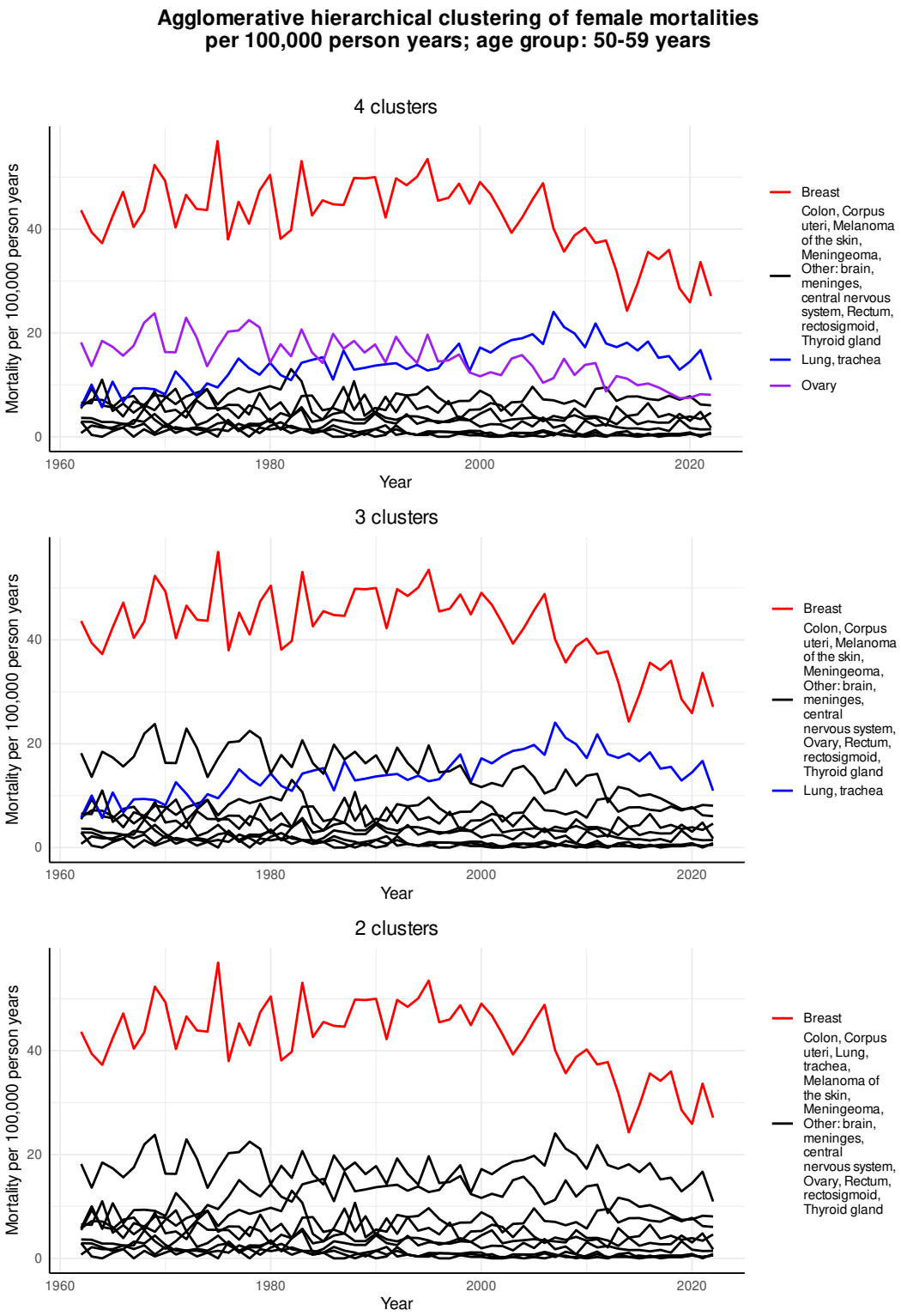


Figure B26: Agglomerative hierarchical clustering applied to the mortalities per 100,000 person years of the most common cancers among females aged 50-59 years in Finland from 1962 to 2022.

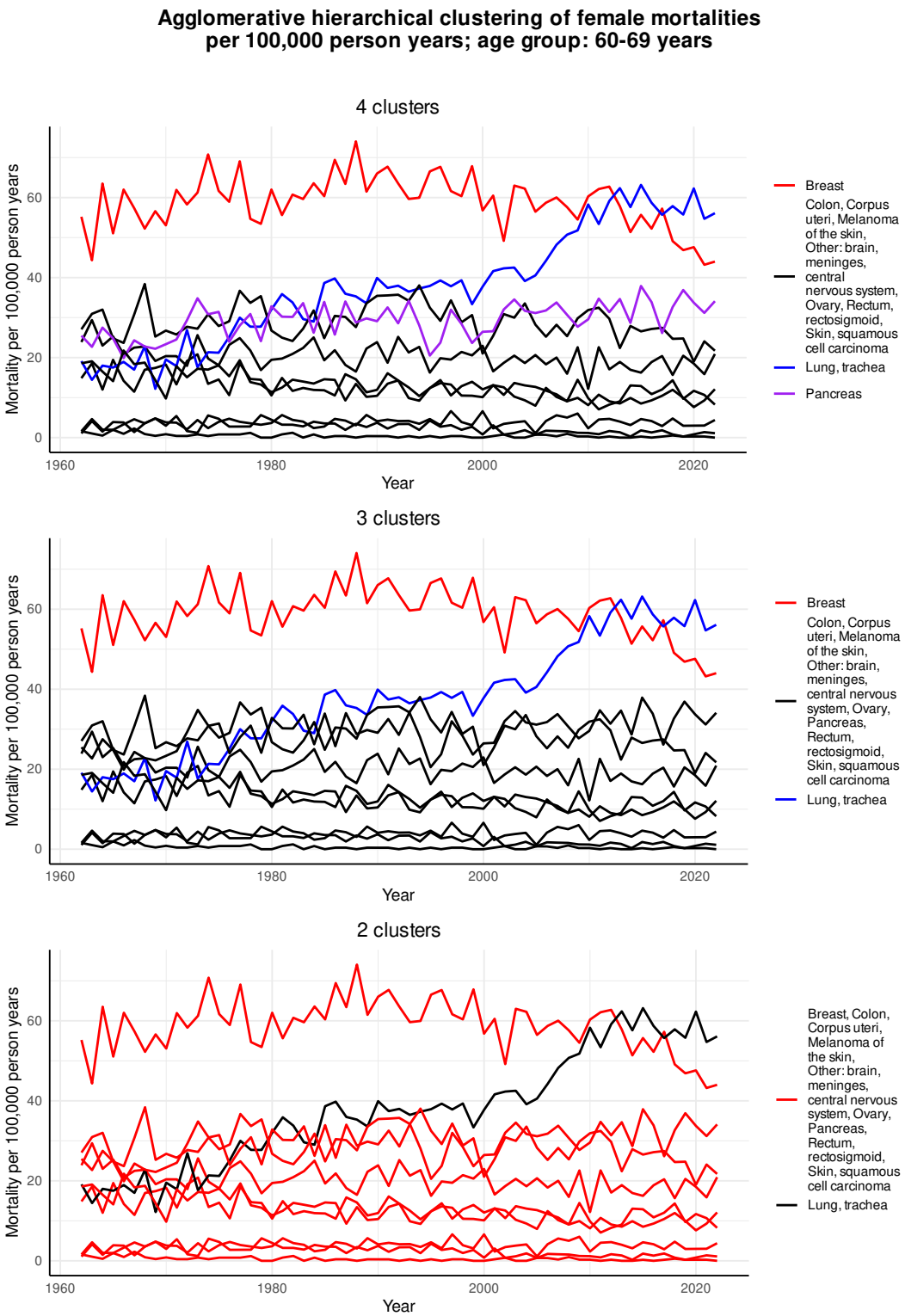


Figure B27: Agglomerative hierarchical clustering applied to the mortalities per 100,000 person years of the most common cancers among females aged 60-69 years in Finland from 1962 to 2022.

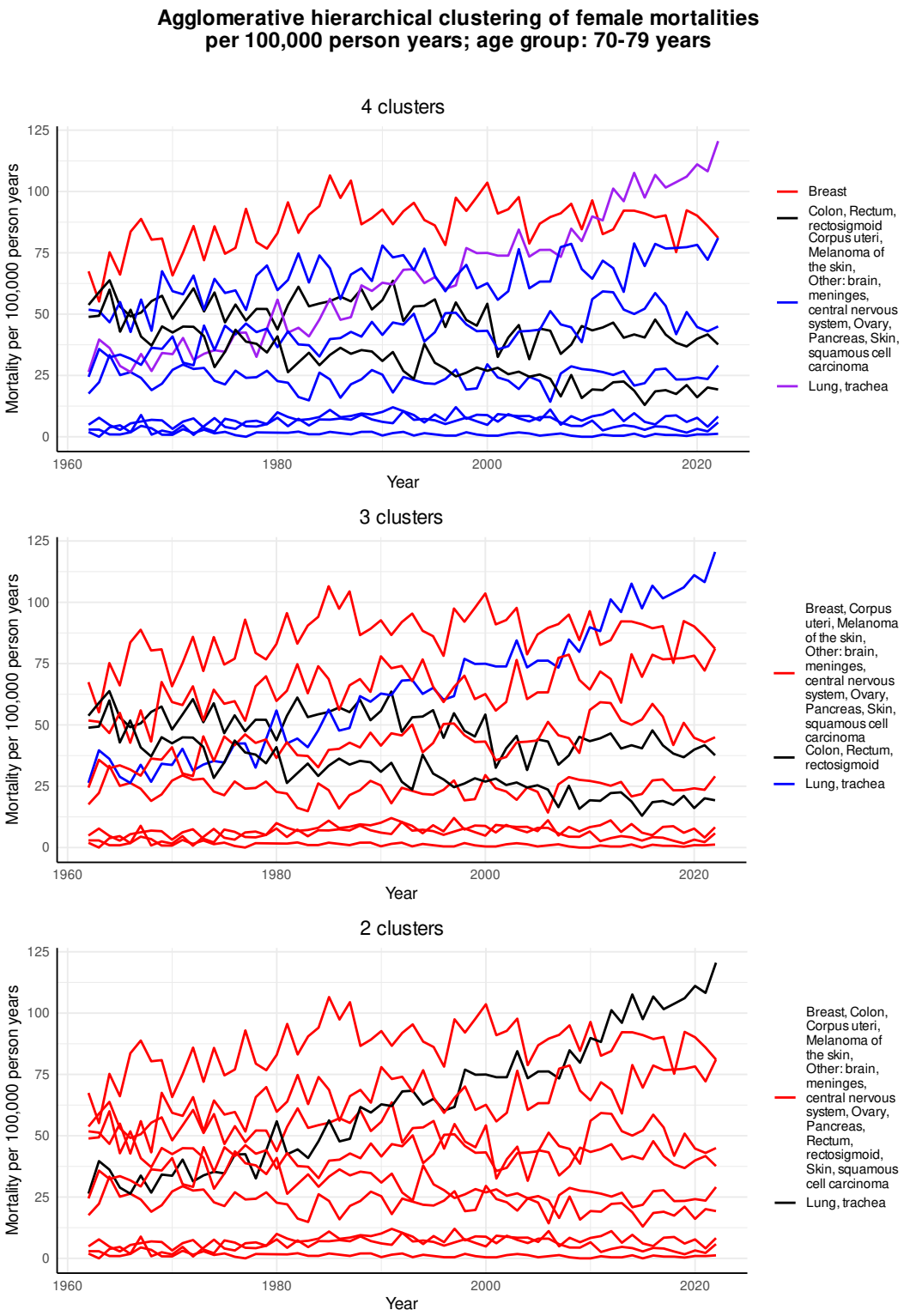


Figure B28: Agglomerative hierarchical clustering applied to the mortalities per 100,000 person years of the most common cancers among females aged 70-79 years in Finland from 1962 to 2022.

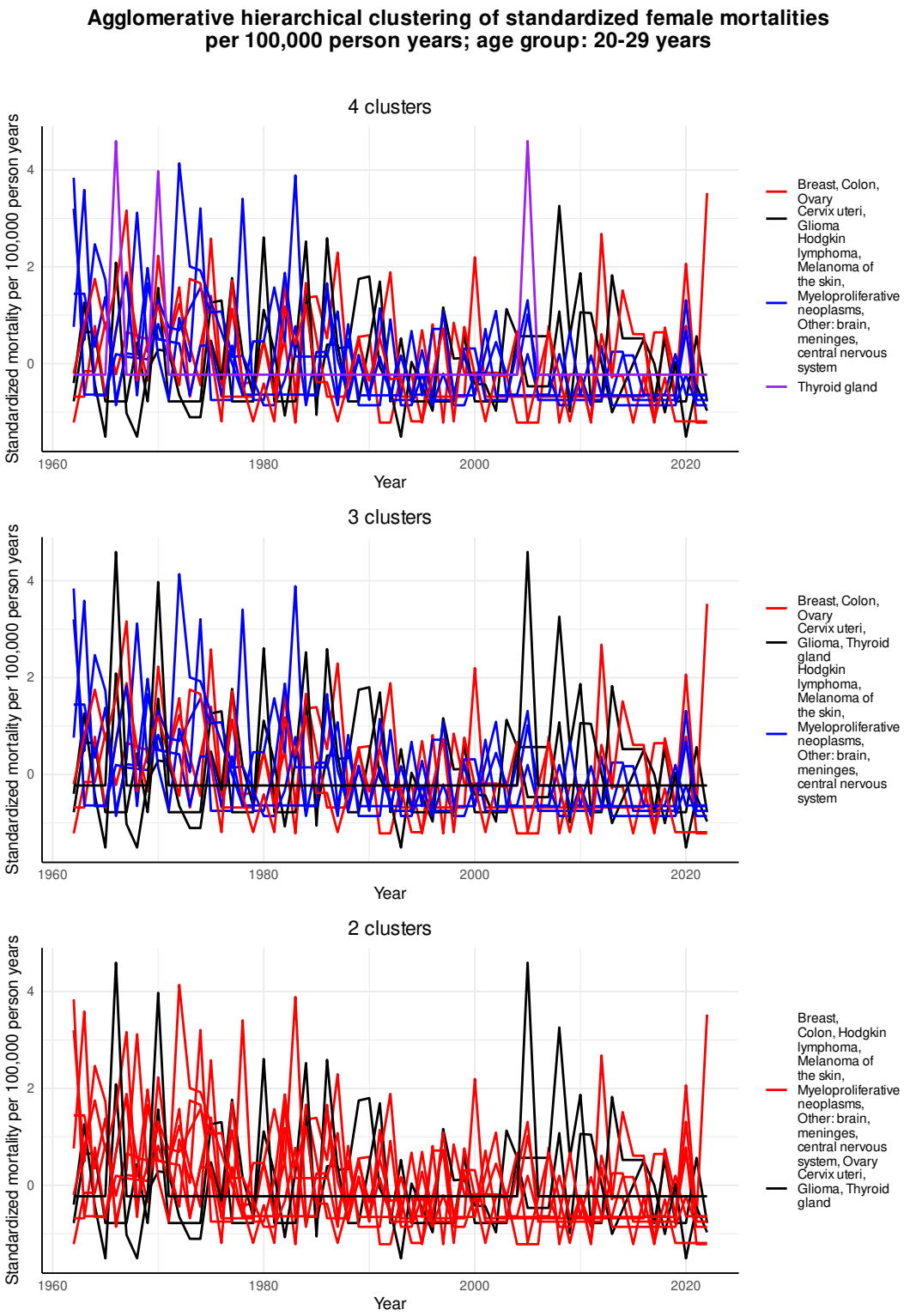


Figure B29: Agglomerative hierarchical clustering applied to the standardized mortalities per 100,000 person years of the most common cancers among females aged 20-29 years in Finland from 1962 to 2022.

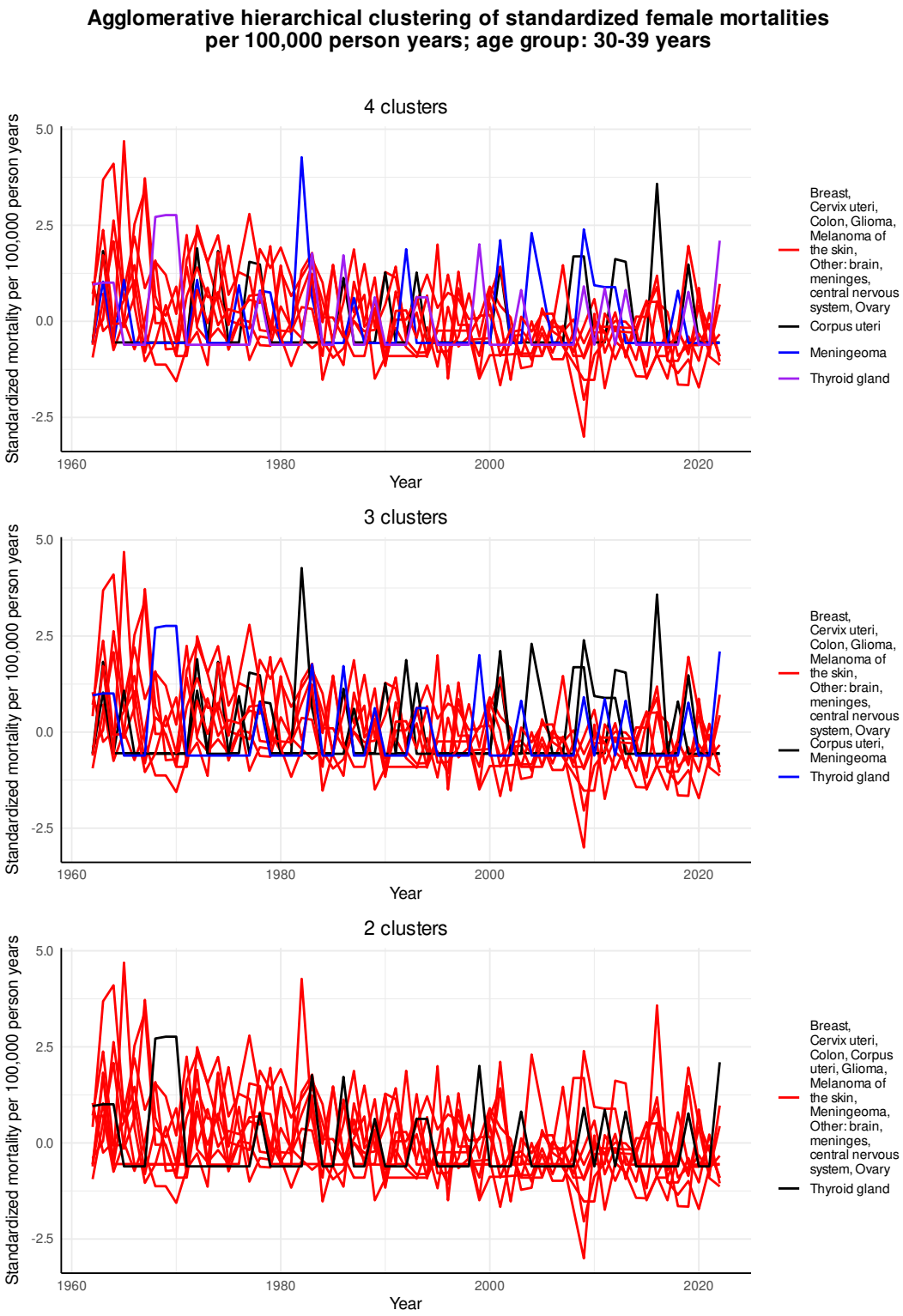


Figure B30: Agglomerative hierarchical clustering applied to the standardized mortalities per 100,000 person years of the most common cancers among females aged 30-39 years in Finland from 1962 to 2022.

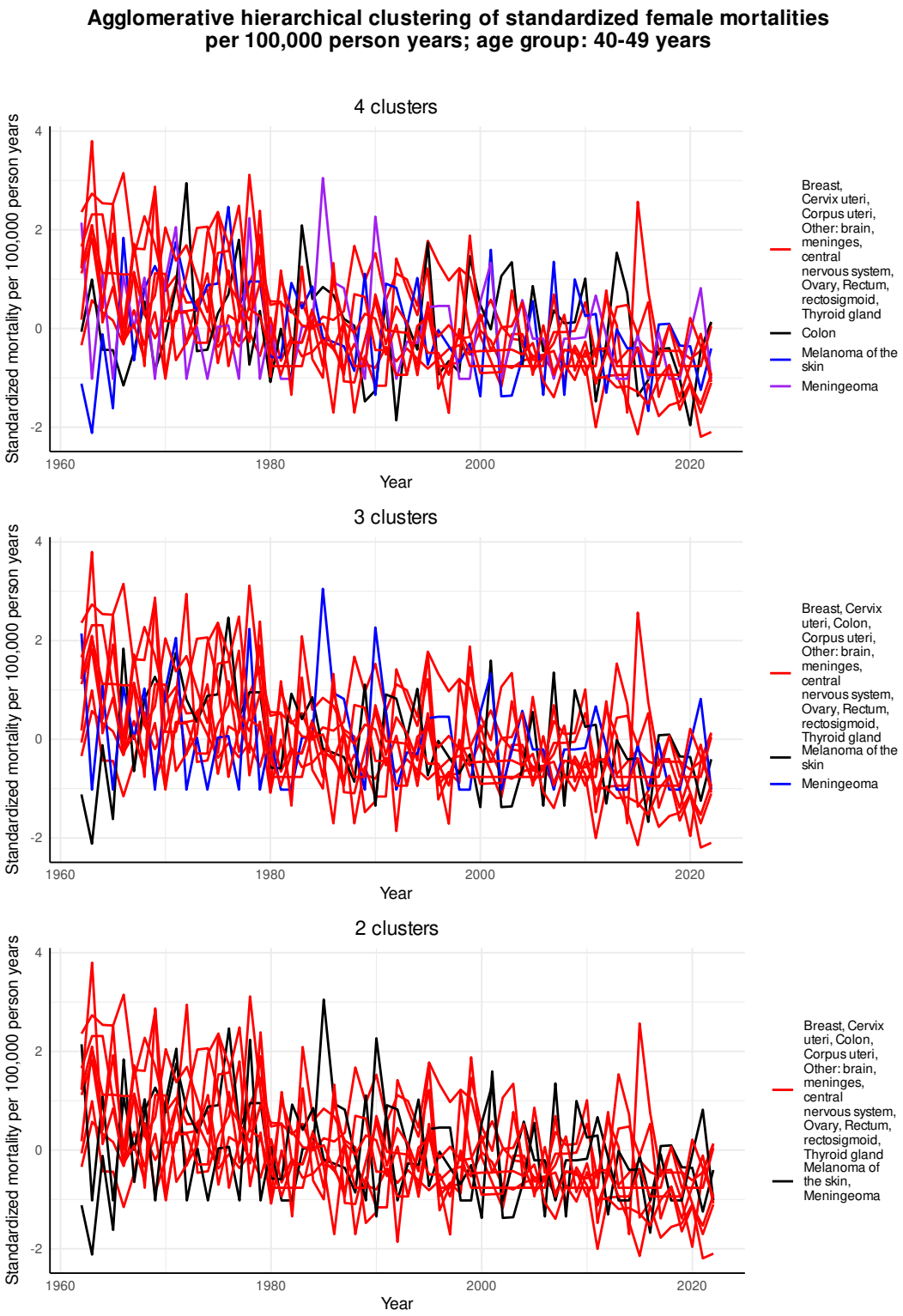


Figure B31: Agglomerative hierarchical clustering applied to the standardized mortalities per 100,000 person years of the most common cancers among females aged 40-49 years in Finland from 1962 to 2022.

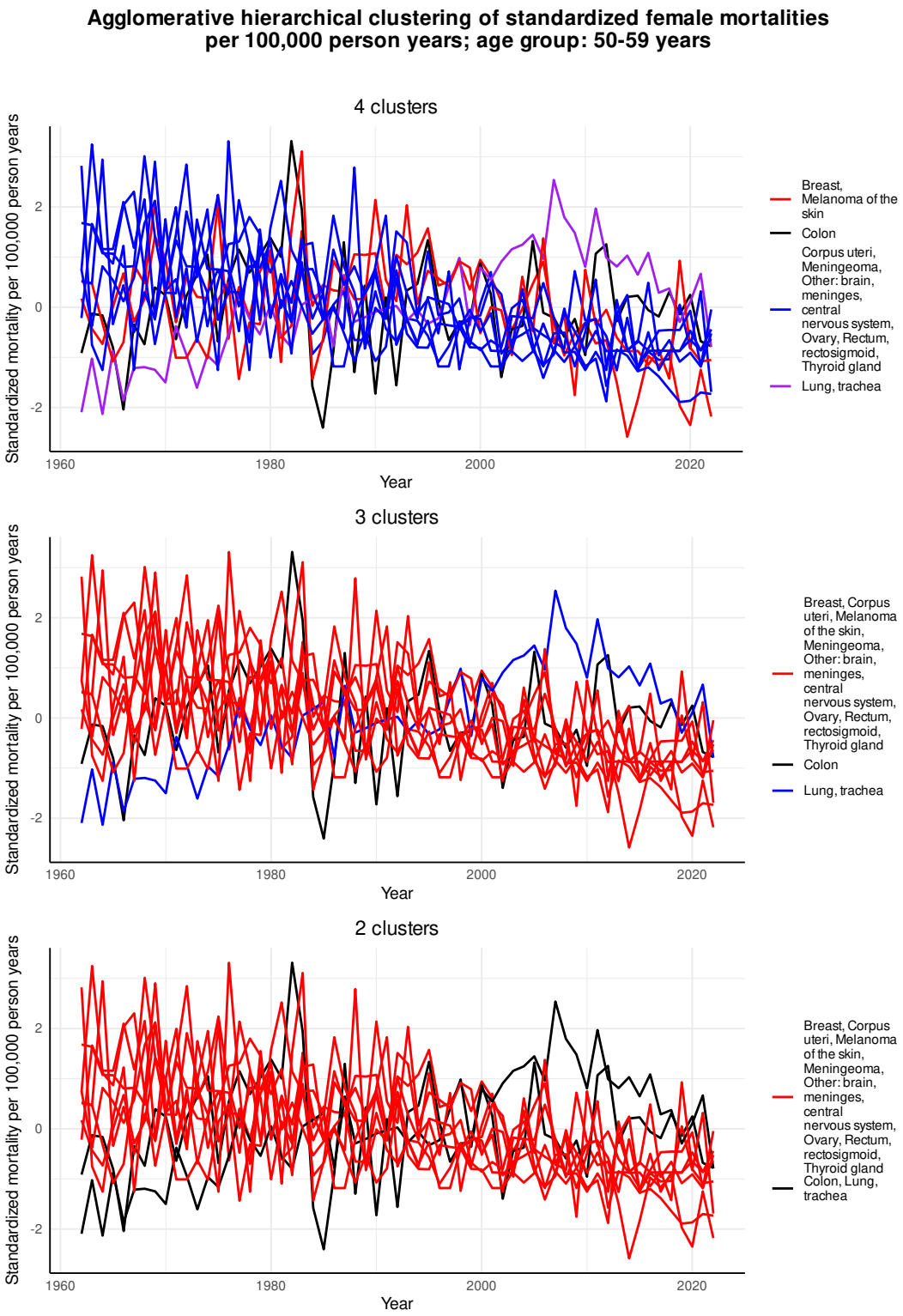


Figure B32: Agglomerative hierarchical clustering applied to the standardized mortalities per 100,000 person years of the most common cancers among females aged 50-59 years in Finland from 1962 to 2022.

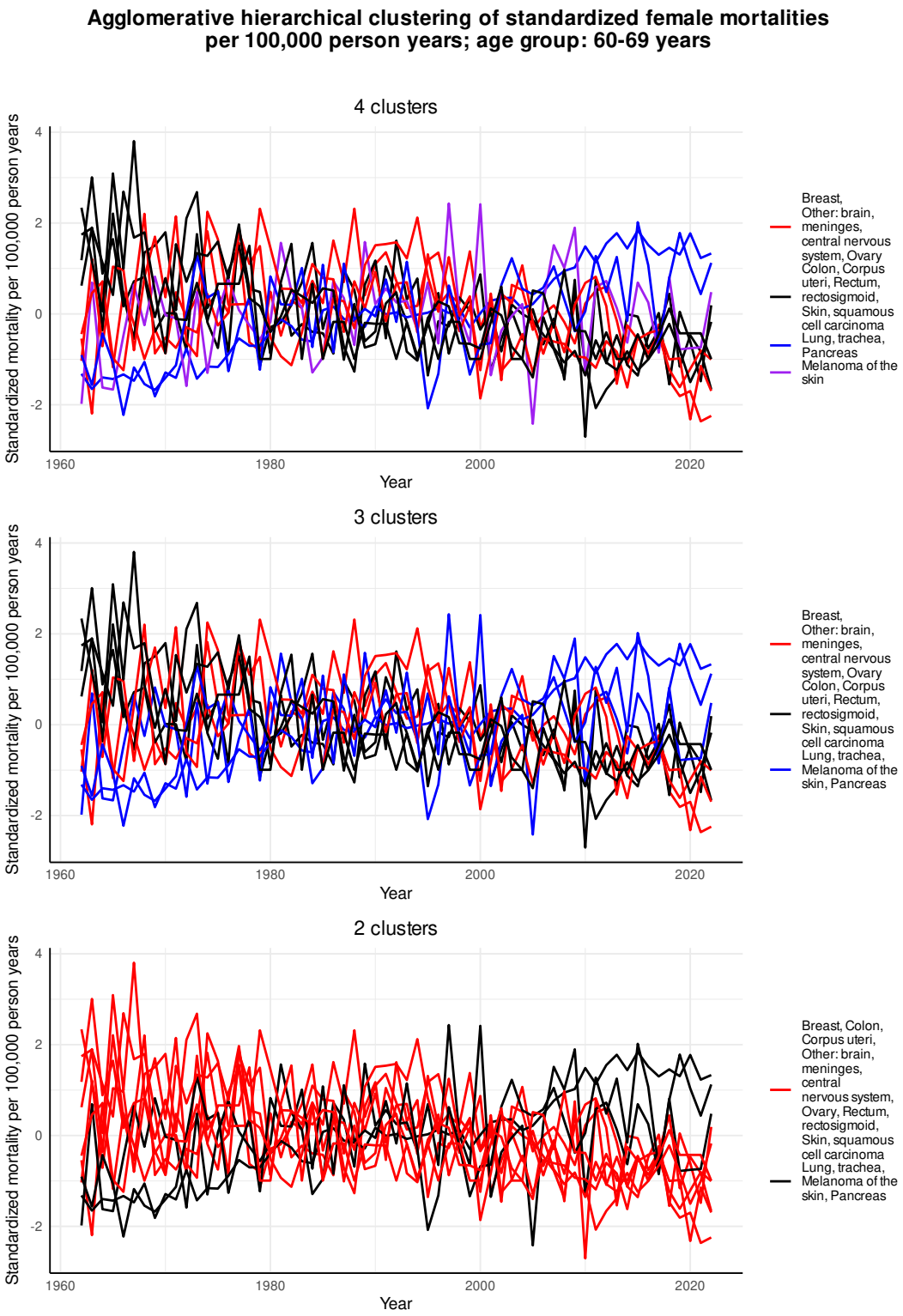


Figure B33: Agglomerative hierarchical clustering applied to the standardized mortalities per 100,000 person years of the most common cancers among females aged 60-69 years in Finland from 1962 to 2022.

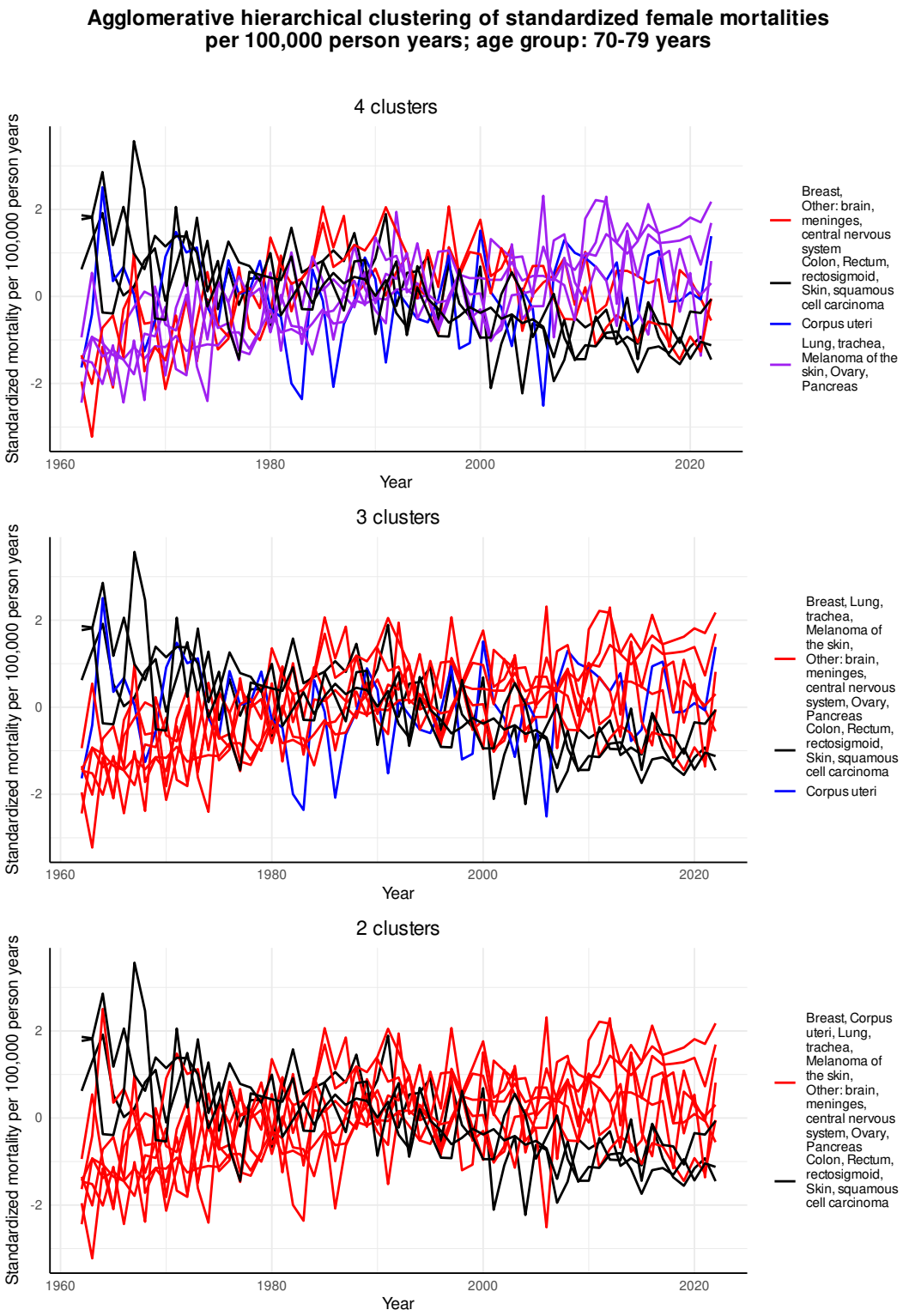


Figure B34: Agglomerative hierarchical clustering applied to the standardized mortalities per 100,000 person years of the most common cancers among females aged 70-79 years in Finland from 1962 to 2022.

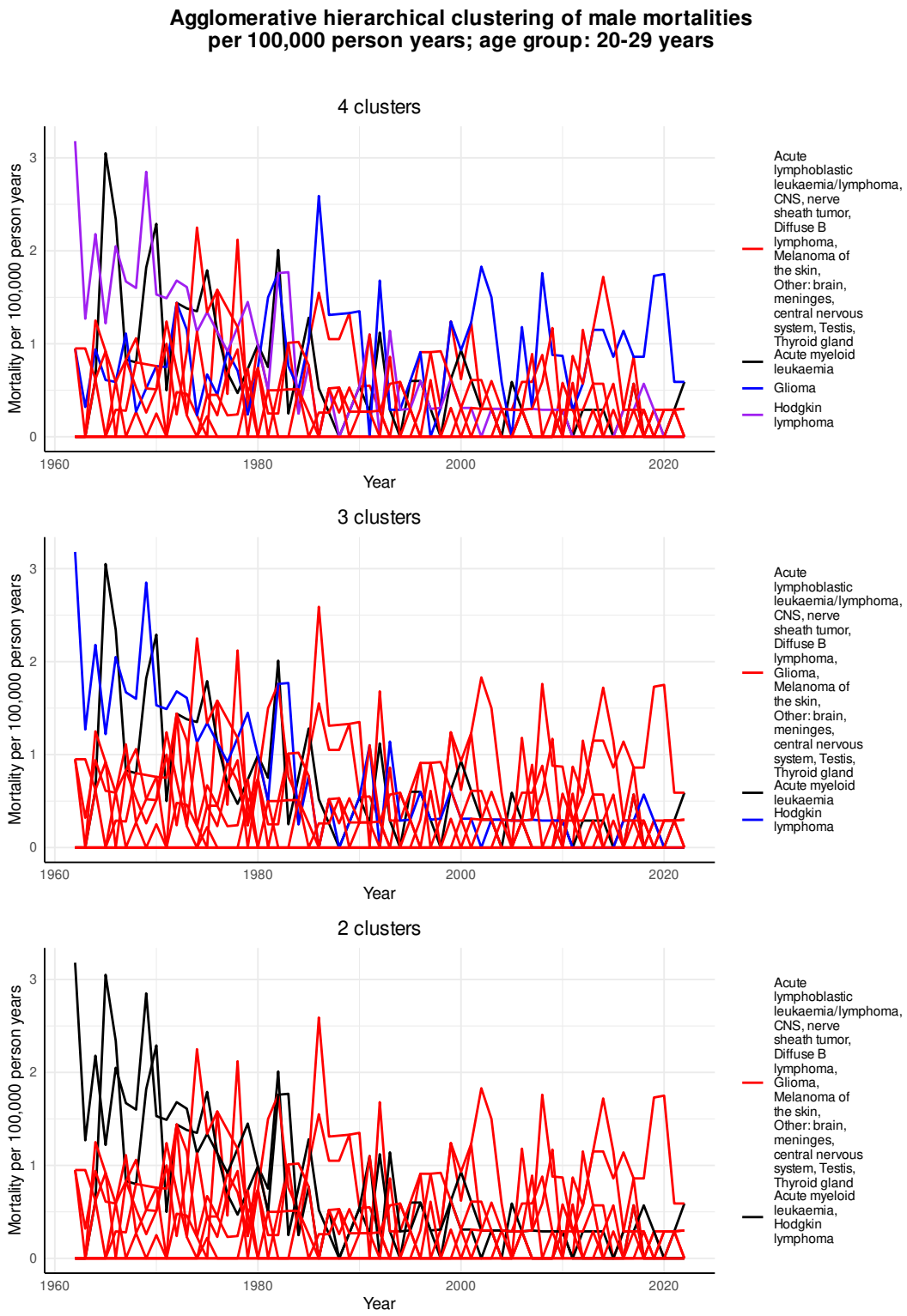


Figure B35: Agglomerative hierarchical clustering applied to the mortalities per 100,000 person years of the most common cancers among males aged 20-29 years in Finland from 1962 to 2022.

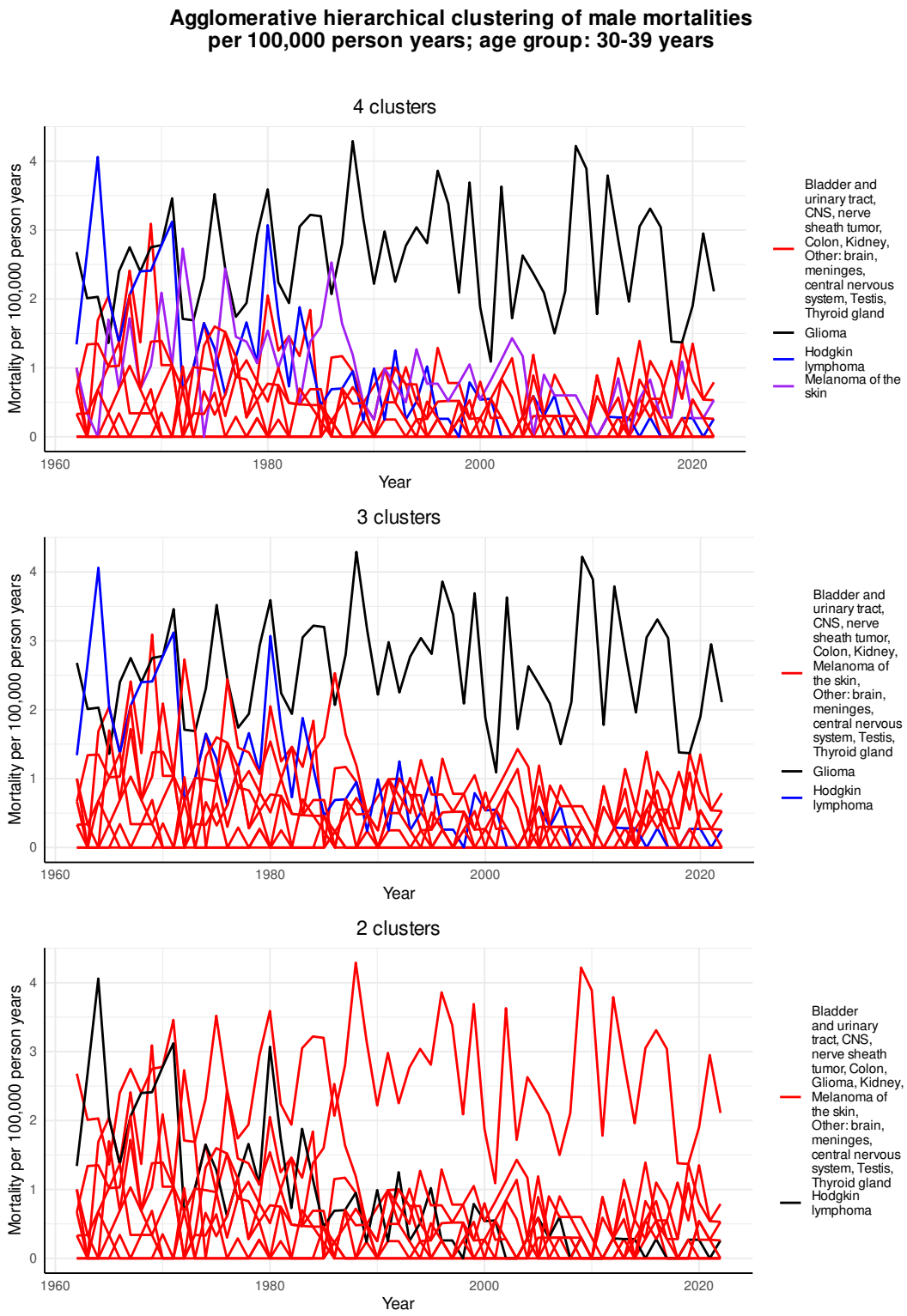


Figure B36: Agglomerative hierarchical clustering applied to the mortalities per 100,000 person years of the most common cancers among males aged 30-39 years in Finland from 1962 to 2022.

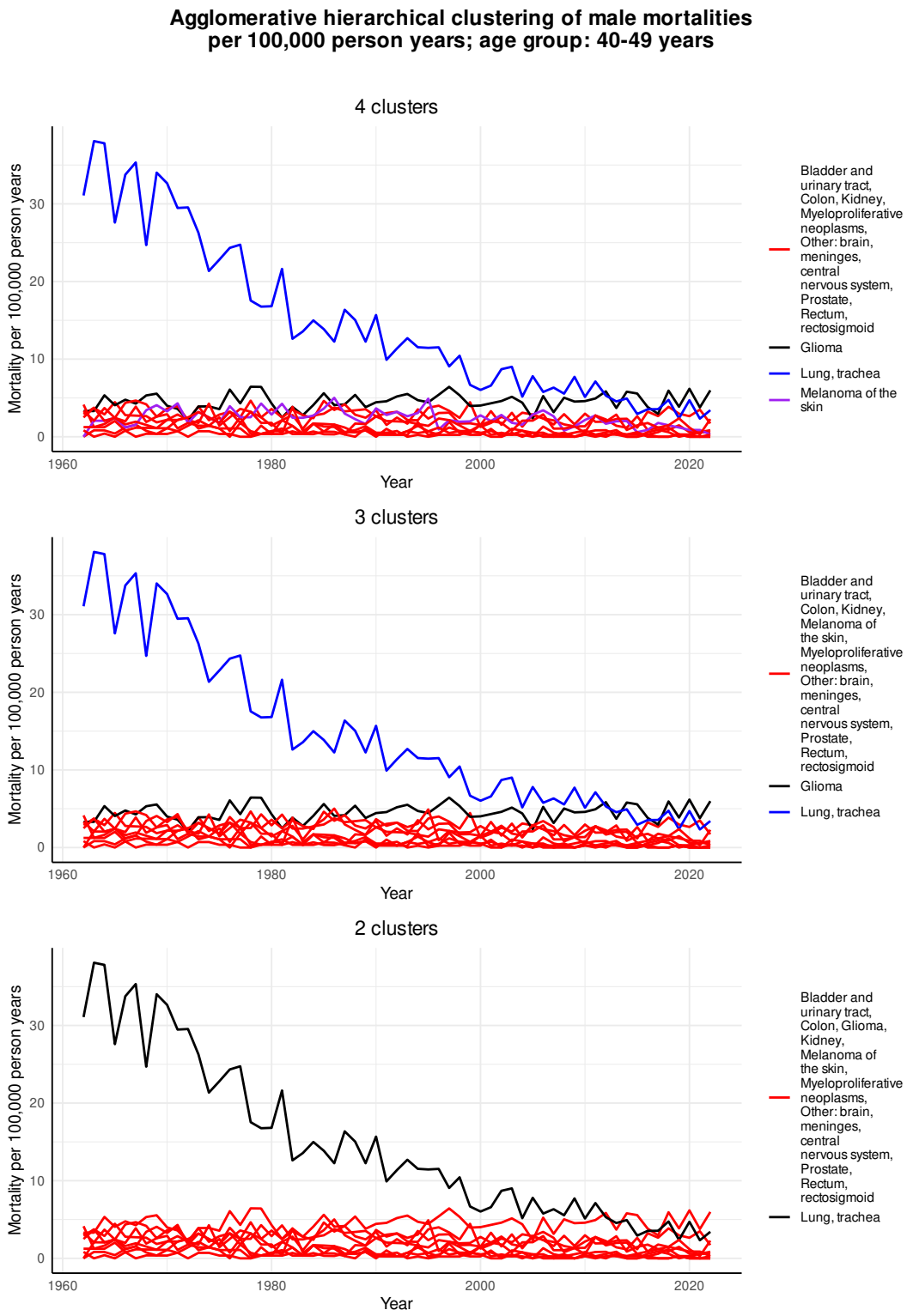


Figure B37: Agglomerative hierarchical clustering applied to the mortalities per 100,000 person years of the most common cancers among males aged 40-49 years in Finland from 1962 to 2022.

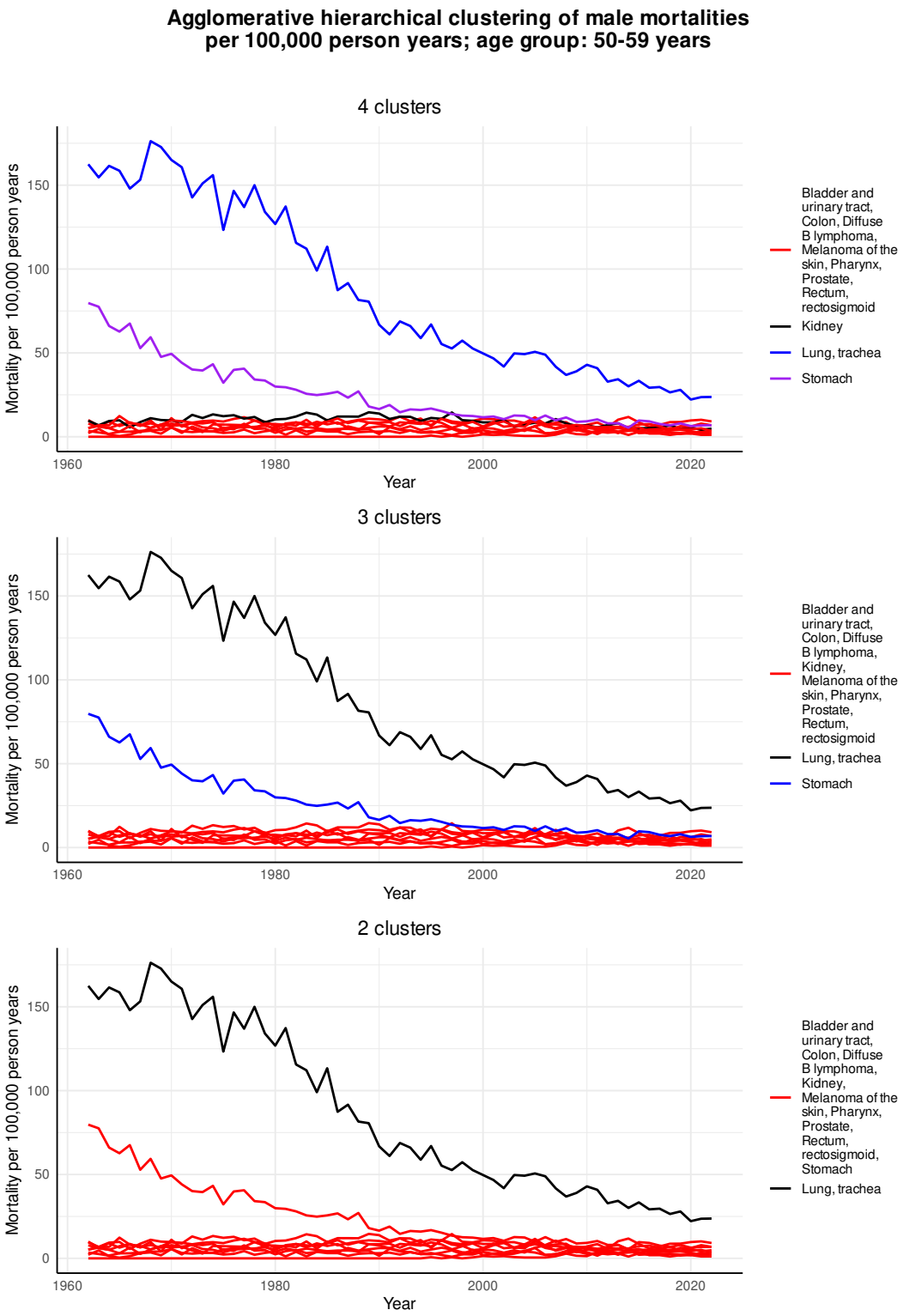


Figure B38: Agglomerative hierarchical clustering applied to the mortalities per 100,000 person years of the most common cancers among males aged 50-59 years in Finland from 1962 to 2022.

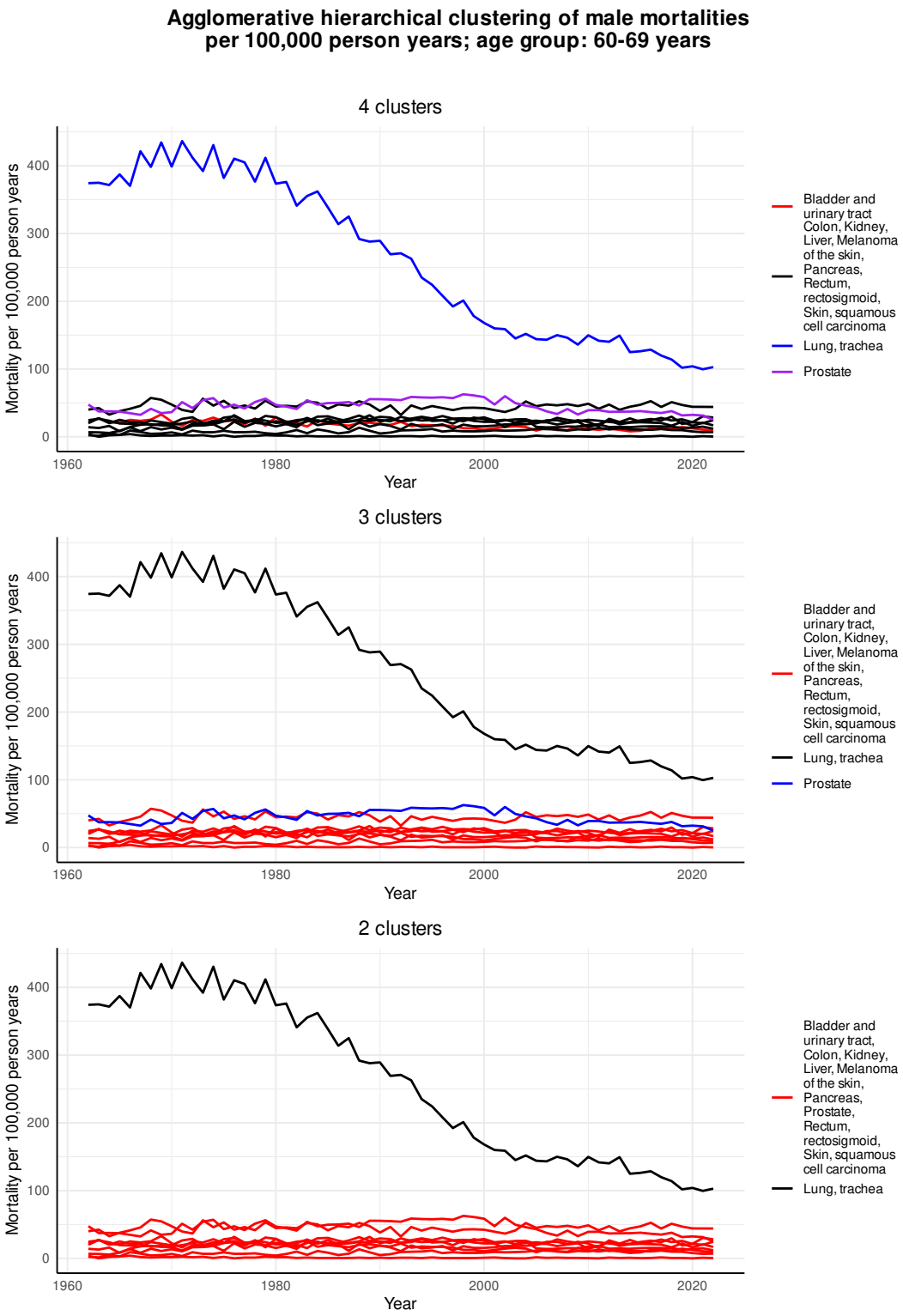


Figure B39: Agglomerative hierarchical clustering applied to the mortalities per 100,000 person years of the most common cancers among males aged 60-69 years in Finland from 1962 to 2022.

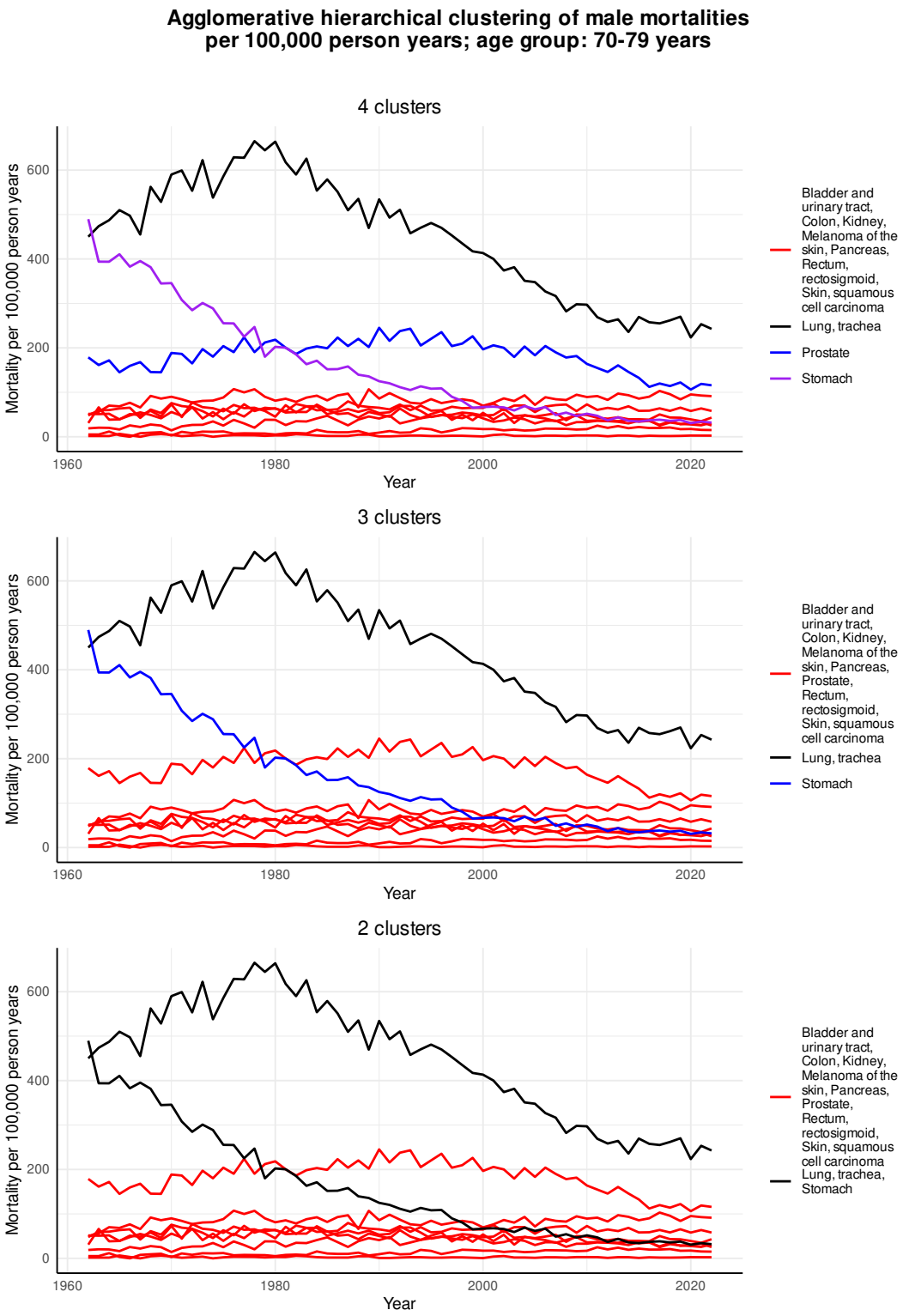


Figure B40: Agglomerative hierarchical clustering applied to the mortalities per 100,000 person years of the most common cancers among males aged 70-79 years in Finland from 1962 to 2022.

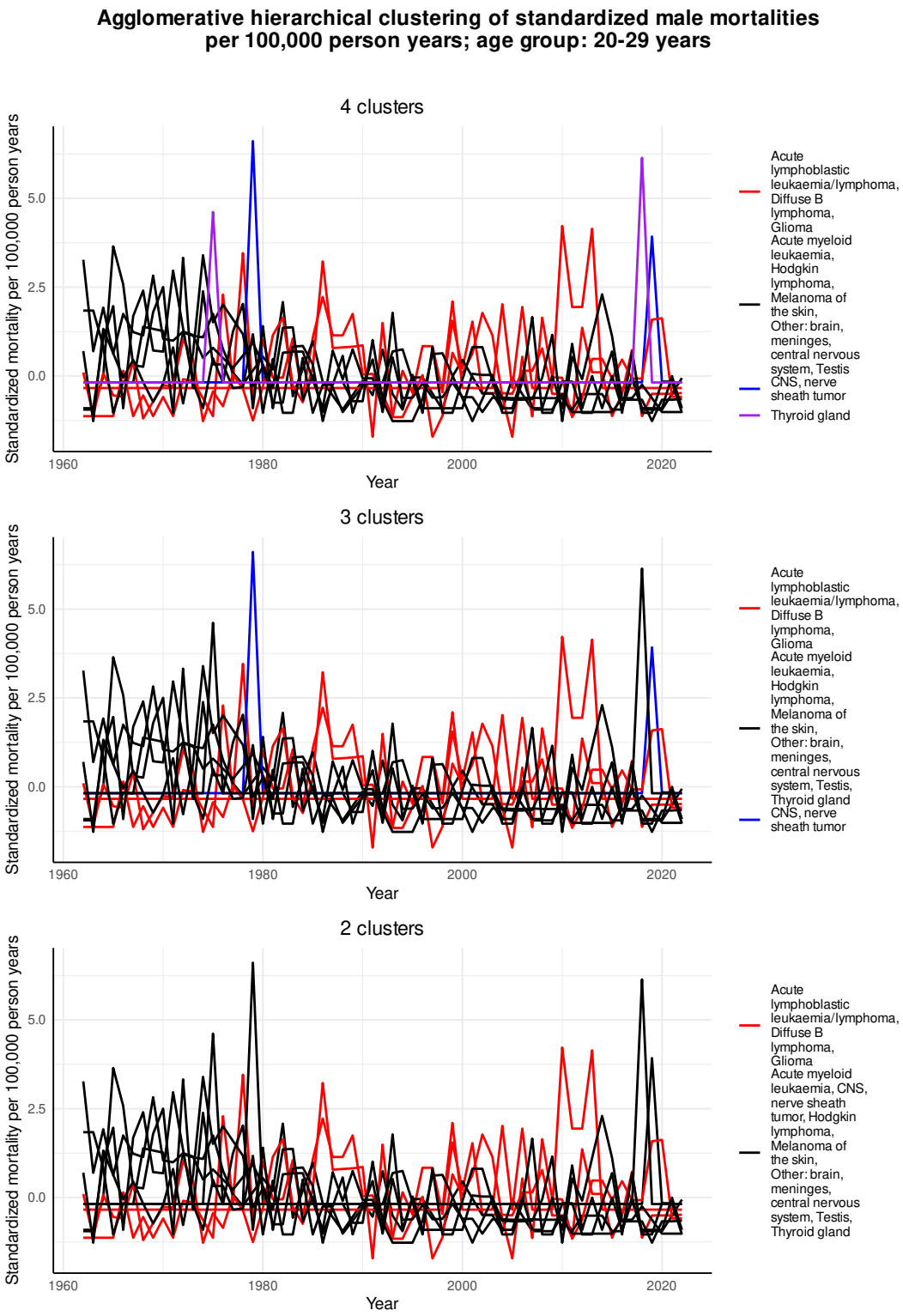


Figure B41: Agglomerative hierarchical clustering applied to the standardized mortalities per 100,000 person years of the most common cancers among males aged 20-29 years in Finland from 1962 to 2022.

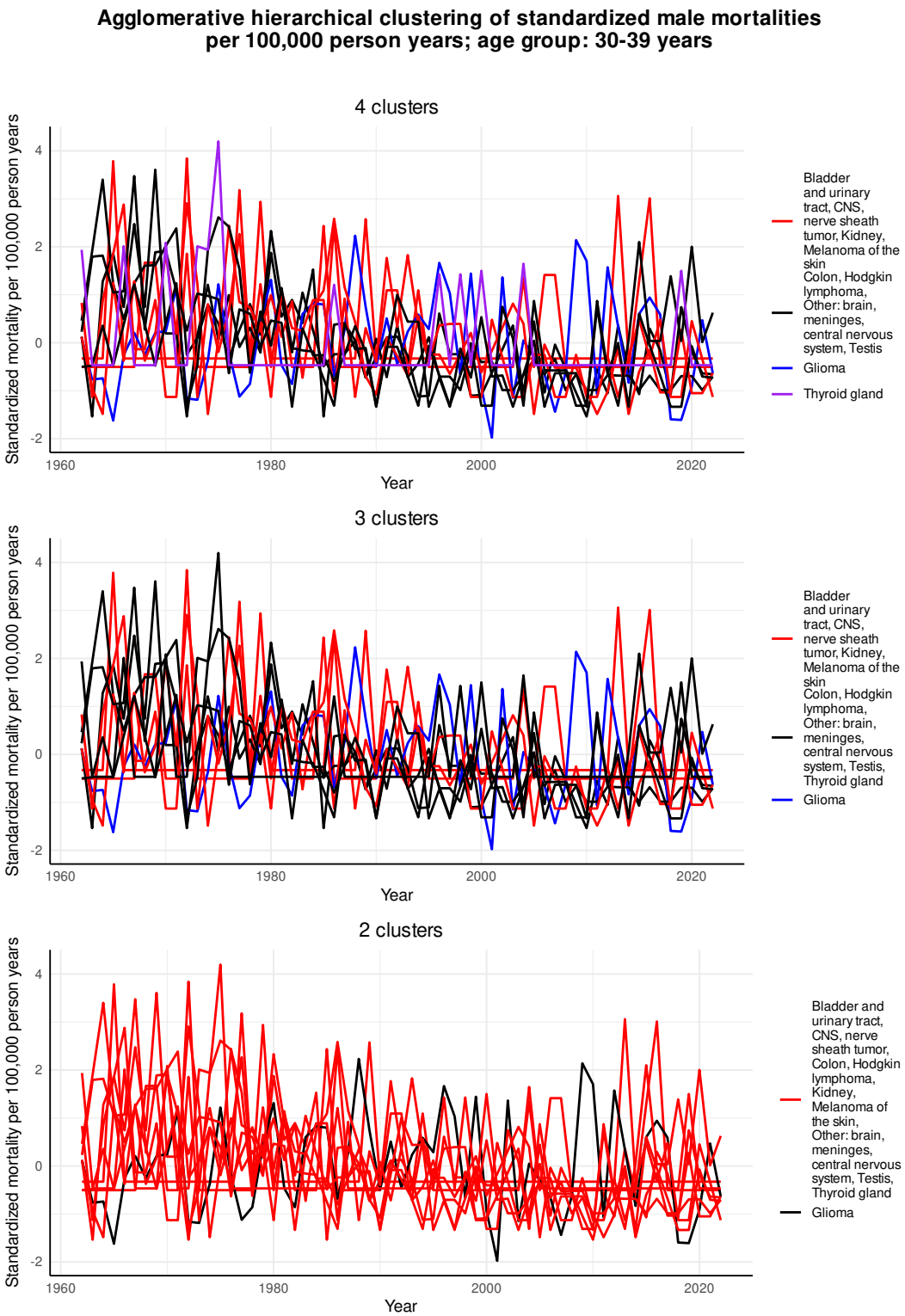


Figure B42: Agglomerative hierarchical clustering applied to the standardized mortalities per 100,000 person years of the most common cancers among males aged 30-39 years in Finland from 1962 to 2022.

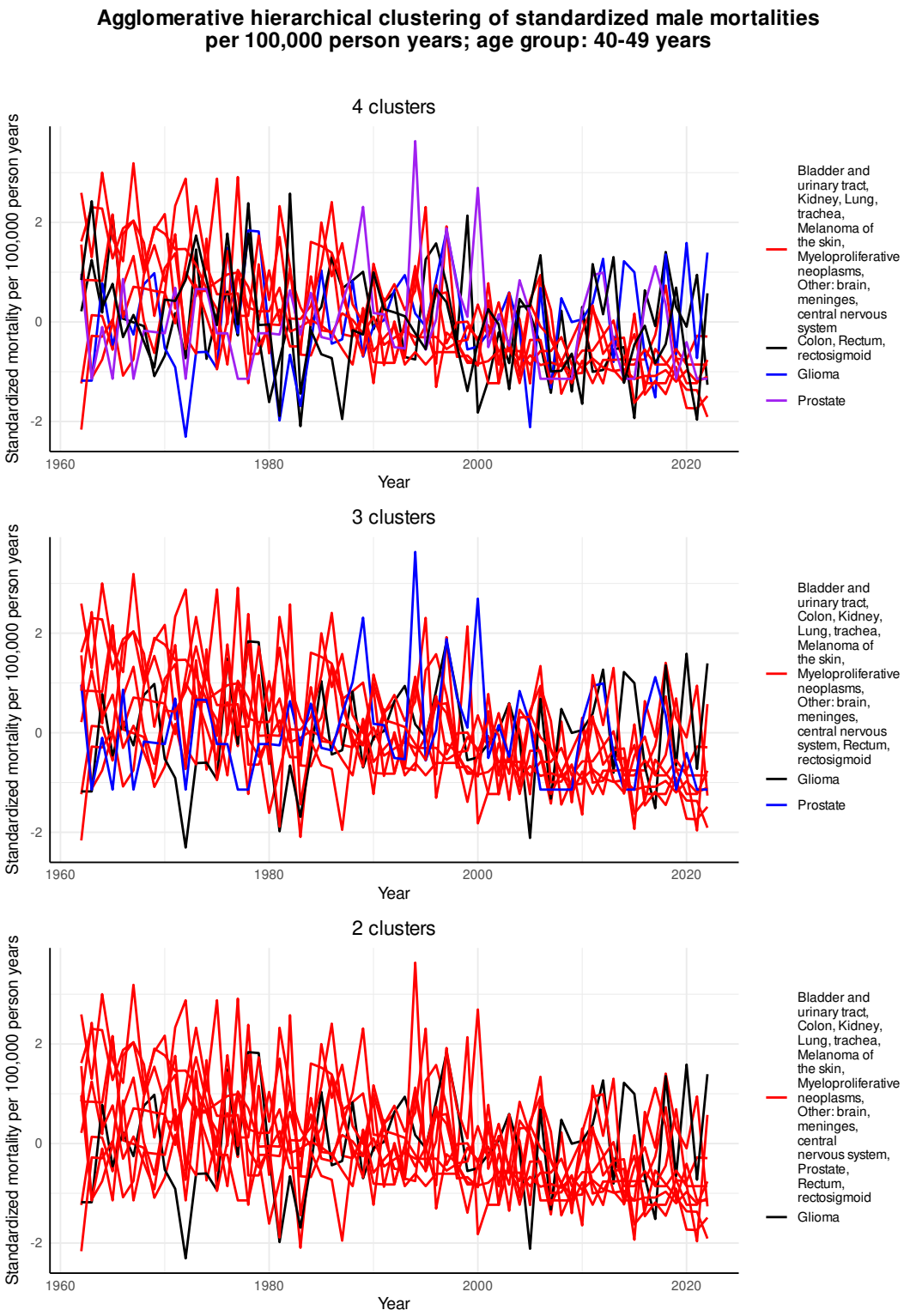


Figure B43: Agglomerative hierarchical clustering applied to the standardized mortalities per 100,000 person years of the most common cancers among males aged 40-49 years in Finland from 1962 to 2022.

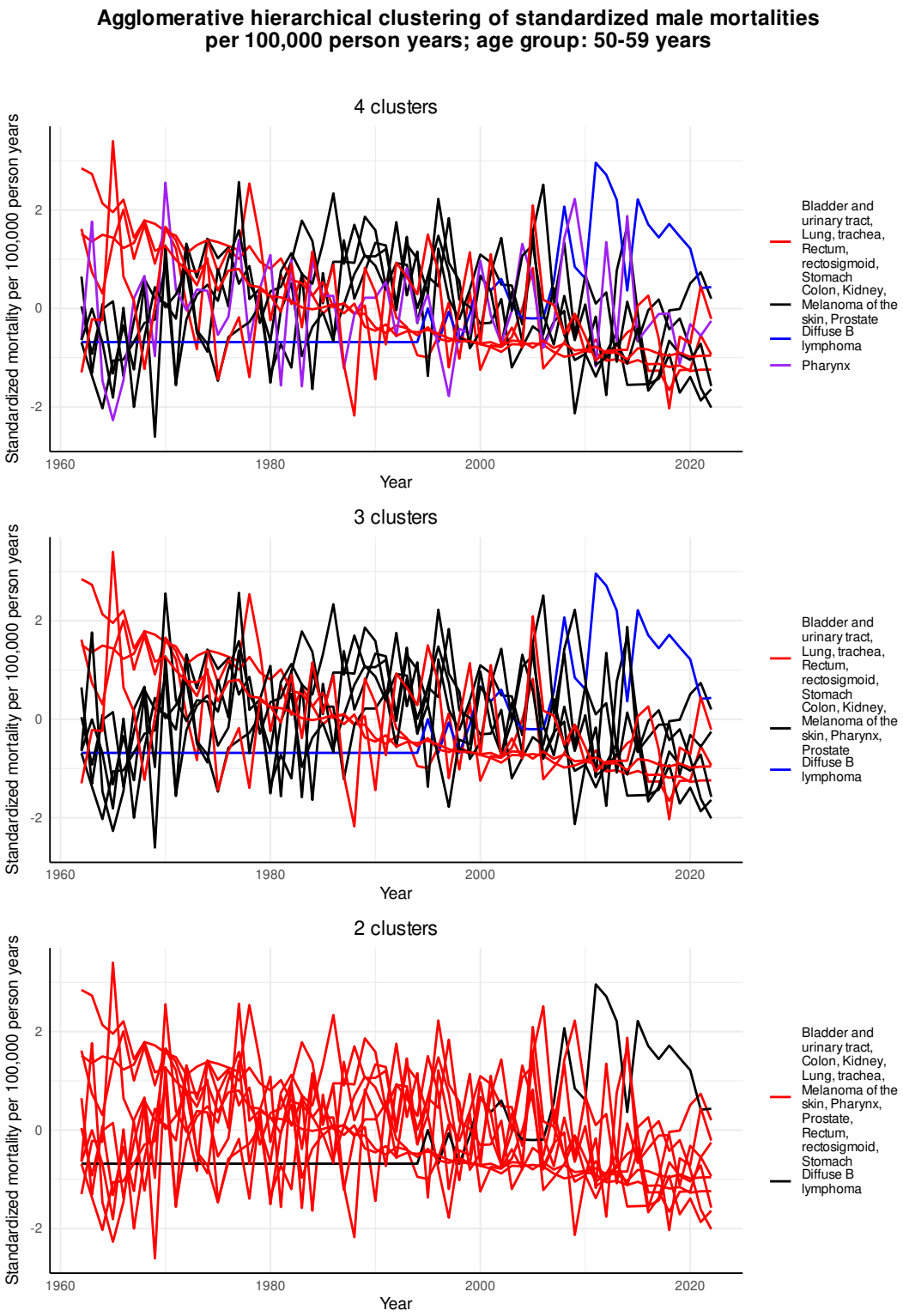


Figure B44: Agglomerative hierarchical clustering applied to the standardized mortalities per 100,000 person years of the most common cancers among males aged 50-59 years in Finland from 1962 to 2022.

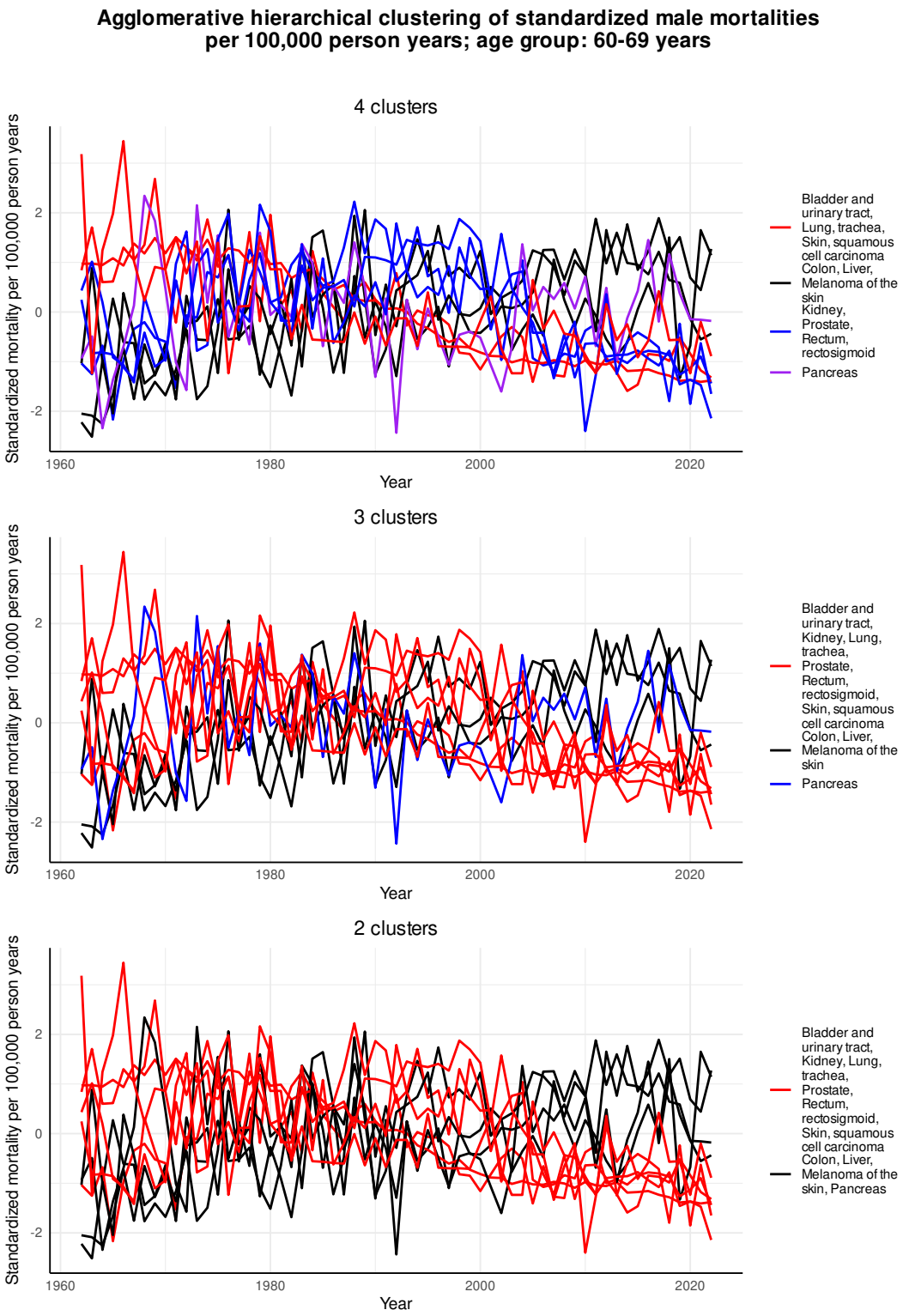


Figure B45: Agglomerative hierarchical clustering applied to the standardized mortalities per 100,000 person years of the most common cancers among males aged 60-69 years in Finland from 1962 to 2022.

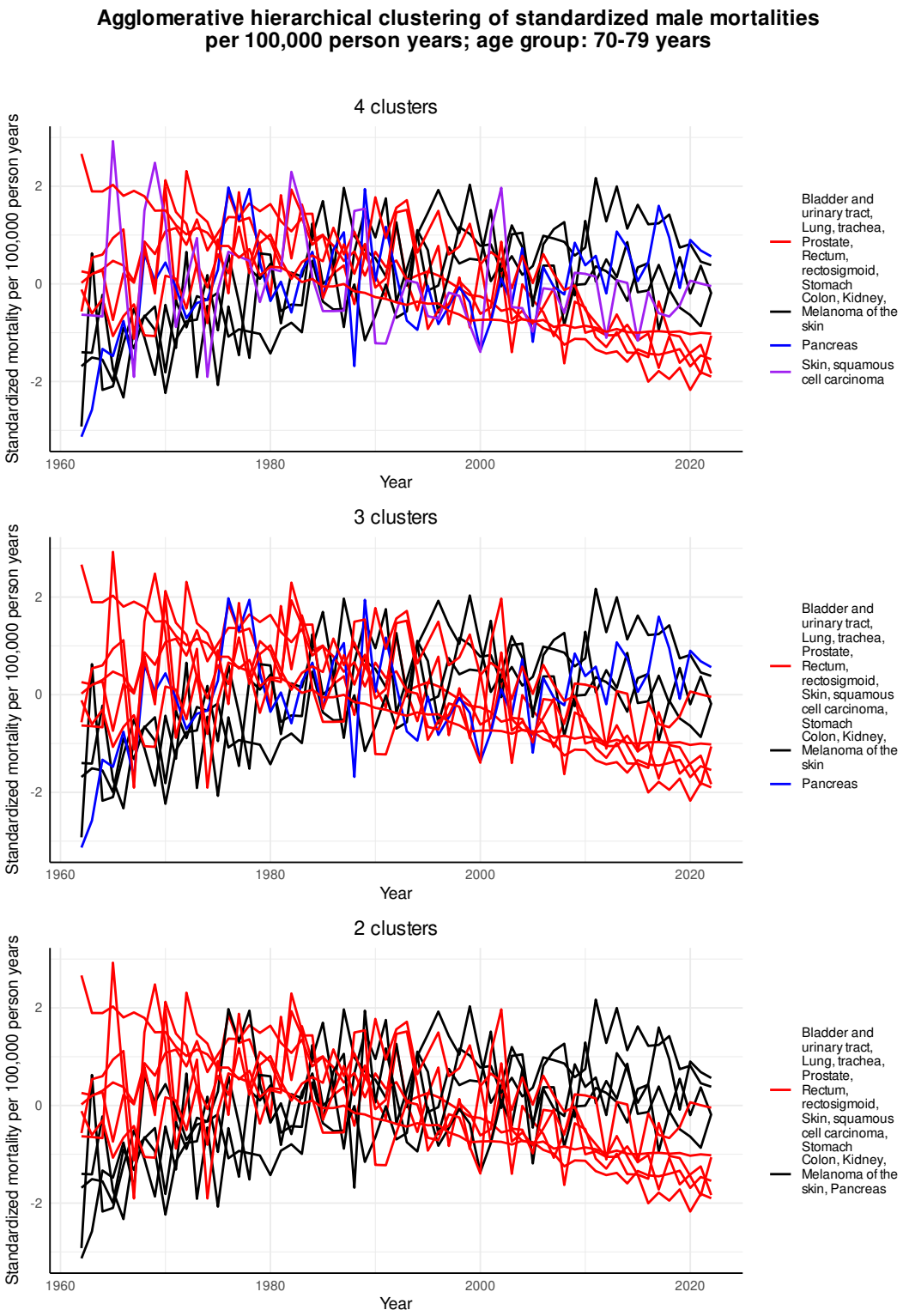


Figure B46: Agglomerative hierarchical clustering applied to the standardized mortalities per 100,000 person years of the most common cancers among males aged 70-79 years in Finland from 1962 to 2022.

C Dendrograms of the clustered cancer incidence and mortality data over time in Finland

This Appendix contains Figures C1 - C23 representing the dendrograms of the clustered cancer incidence and mortality data over time in Finland. The clustering was performed using the agglomerative hierarchical clustering algorithm, as discussed in Section 4. The dendrograms are included for both original, or unstandardized, and standardized data, covering females and males and all age groups from 20-29 to 70-79 years. The only exception are the dendrograms corresponding to the clustering process of female incidence rates, both unstandardized and standardized, for the age group of 20-29 years, which are already illustrated by Figure 13 in Section 5.1.

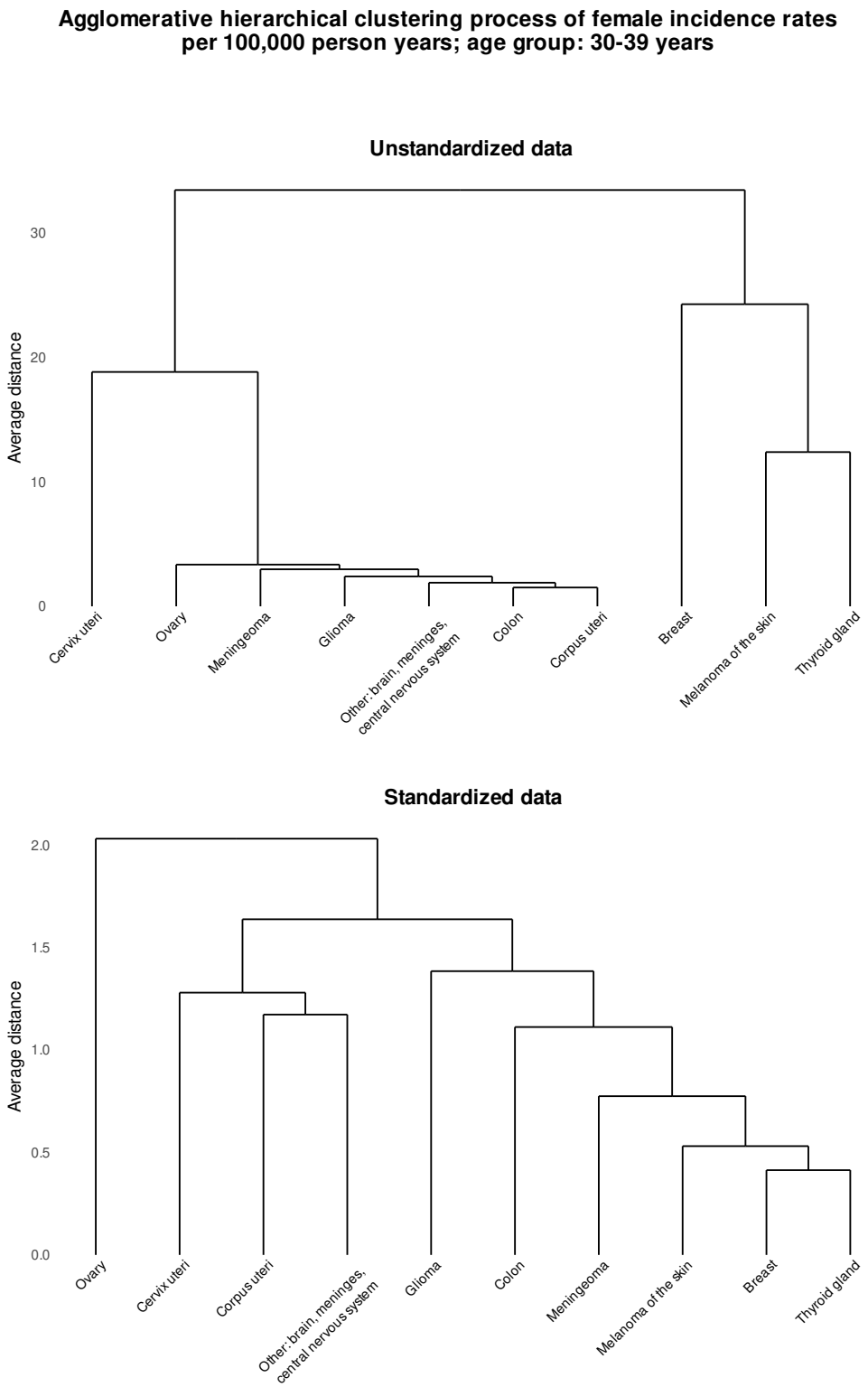


Figure C1: Dendrograms of clustered unstandardized and standardized cancer incidence data of the most common cancers among females aged 30-39 years in Finland from 1962 to 2022.

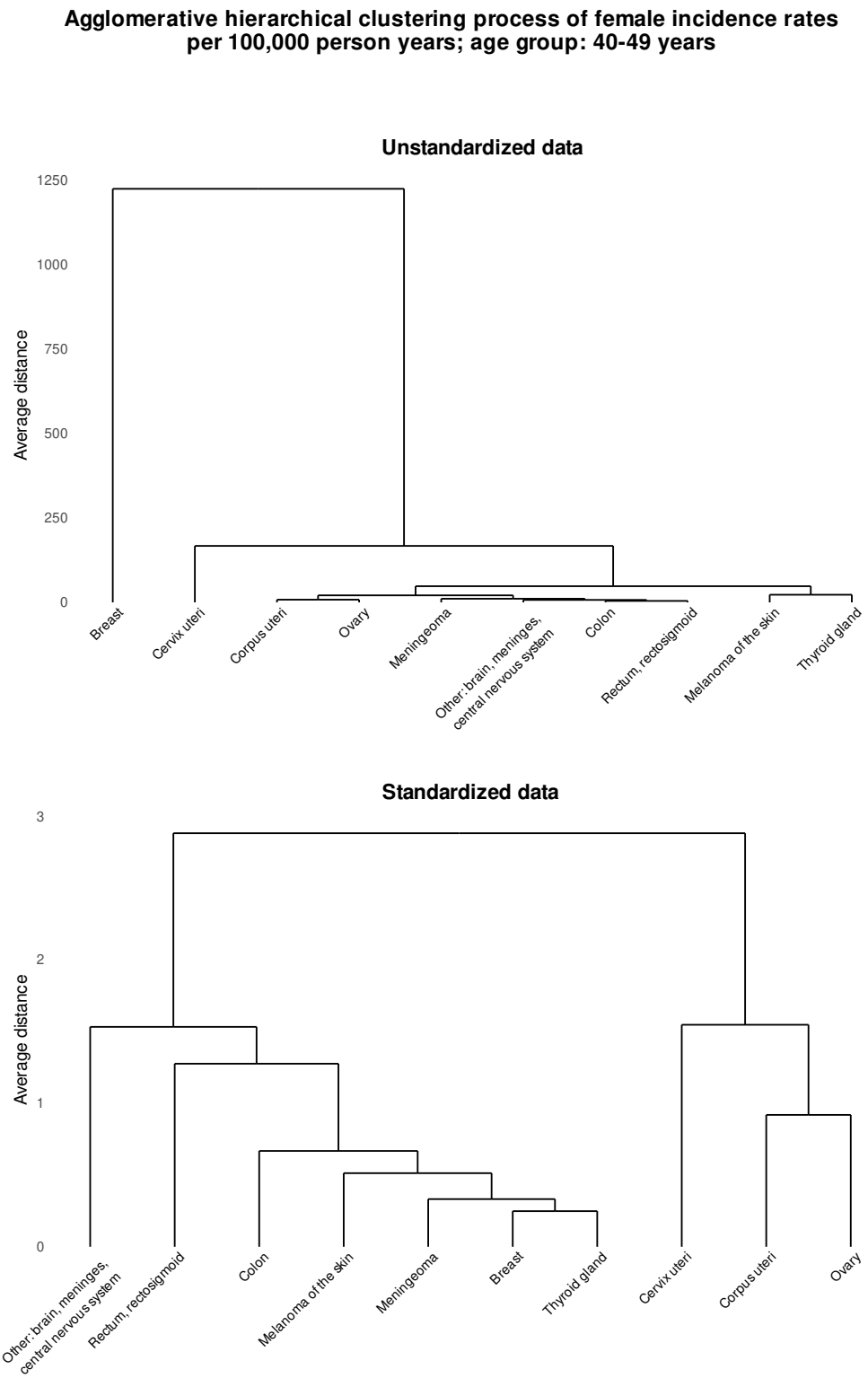


Figure C2: Dendrograms of clustered unstandardized and standardized cancer incidence data of the most common cancers among females aged 40-49 years in Finland from 1962 to 2022.

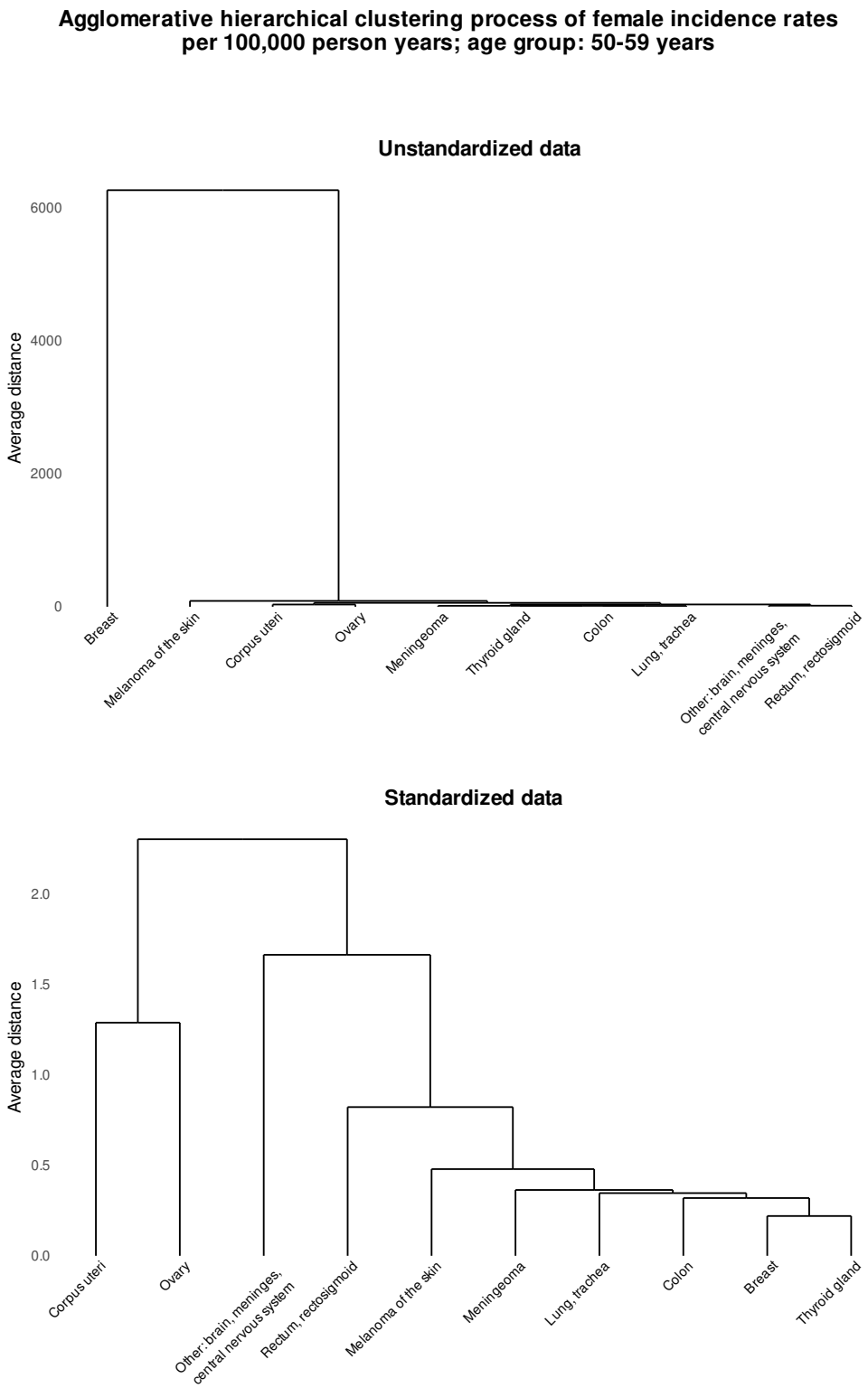


Figure C3: Dendrograms of clustered unstandardized and standardized cancer incidence data of the most common cancers among females aged 50-59 years in Finland from 1962 to 2022.

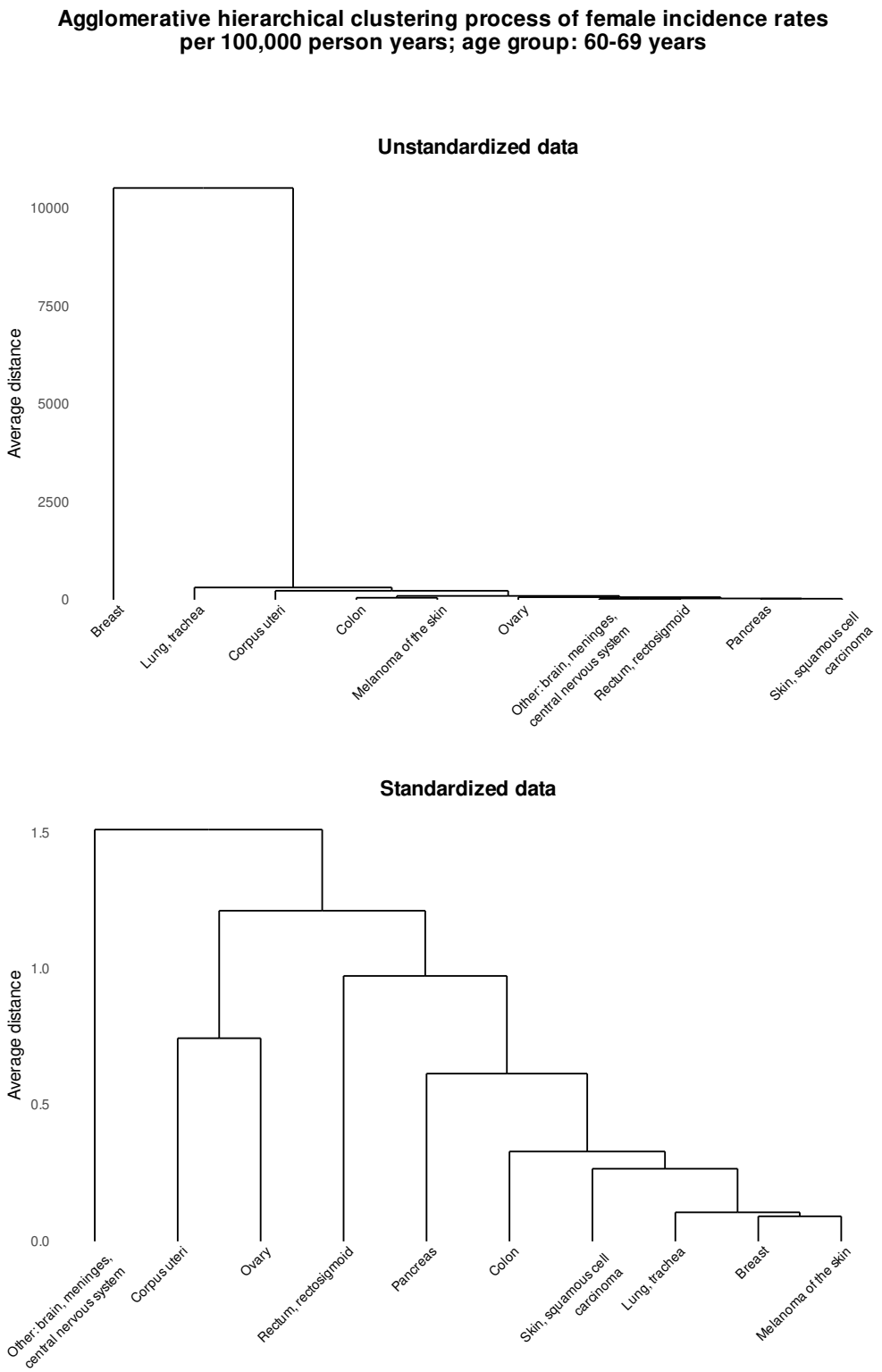


Figure C4: Dendrograms of clustered unstandardized and standardized cancer incidence data of the most common cancers among females aged 60-69 years in Finland from 1962 to 2022.

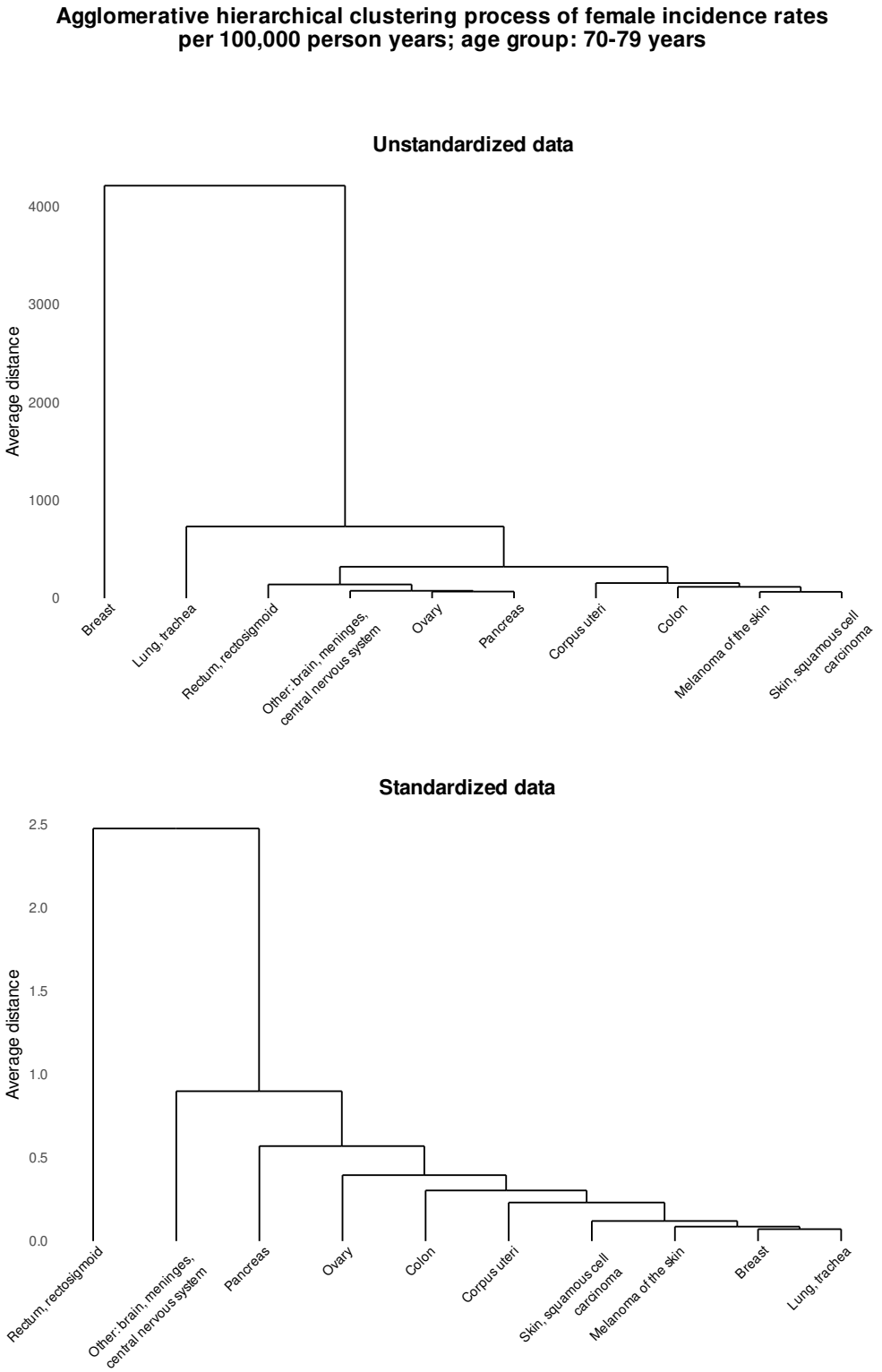


Figure C5: Dendrograms of clustered unstandardized and standardized cancer incidence data of the most common cancers among females aged 70-79 years in Finland from 1962 to 2022.

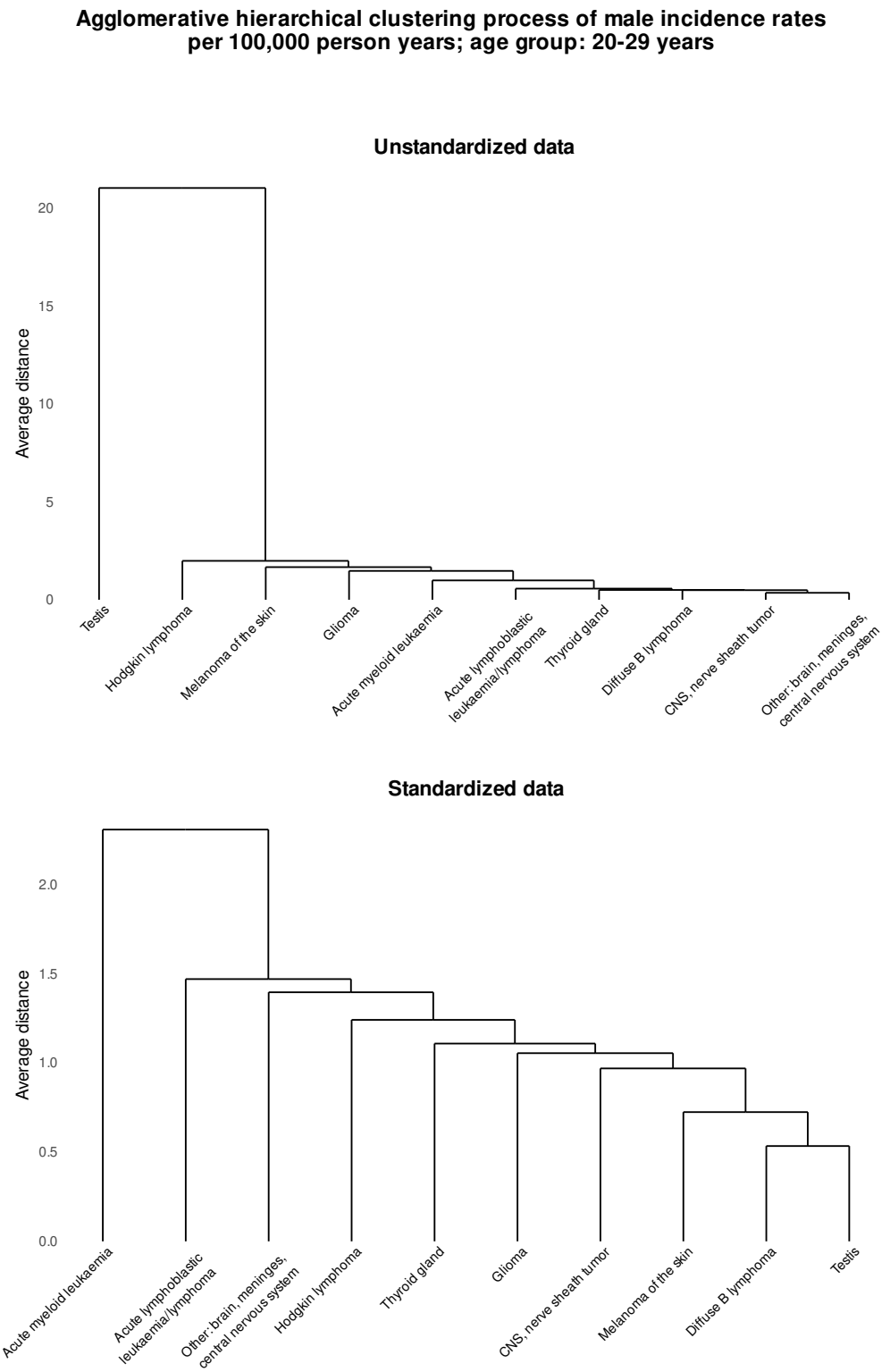


Figure C6: Dendrograms of clustered unstandardized and standardized cancer incidence data of the most common cancers among males aged 20-29 years in Finland from 1962 to 2022.

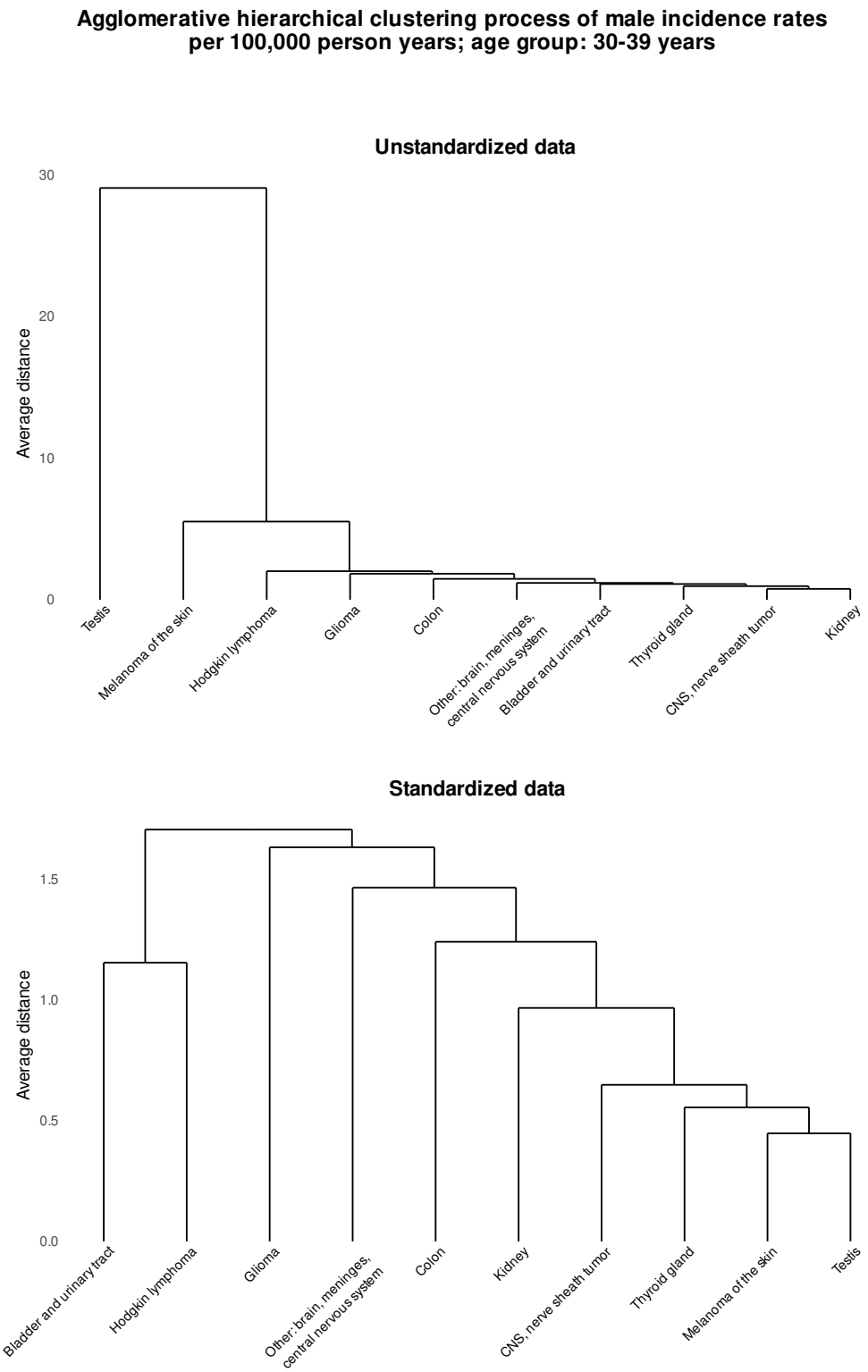


Figure C7: Dendrograms of clustered unstandardized and standardized cancer incidence data of the most common cancers among males aged 30-39 years in Finland from 1962 to 2022.

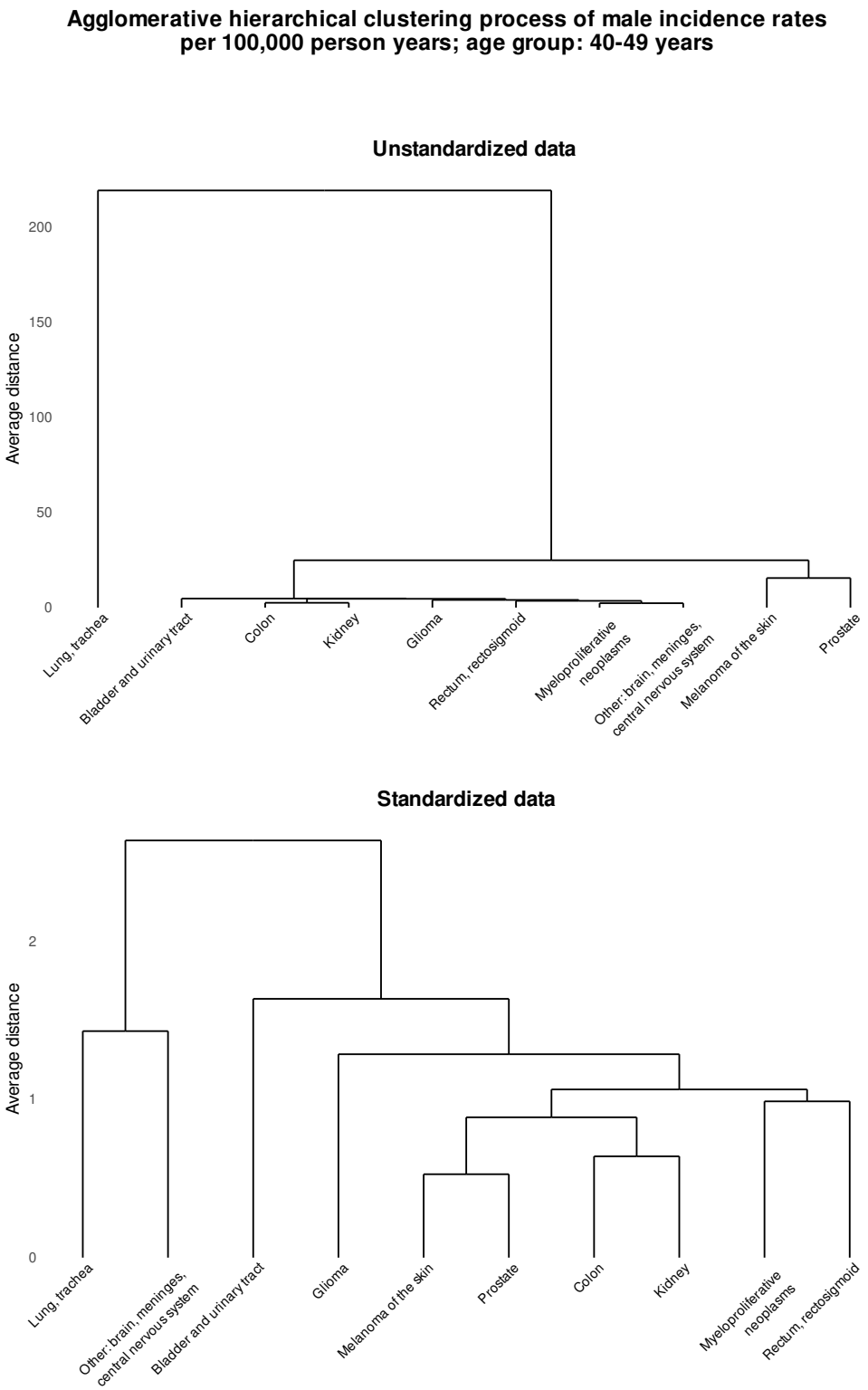


Figure C8: Dendrograms of clustered unstandardized and standardized cancer incidence data of the most common cancers among males aged 40-49 years in Finland from 1962 to 2022.

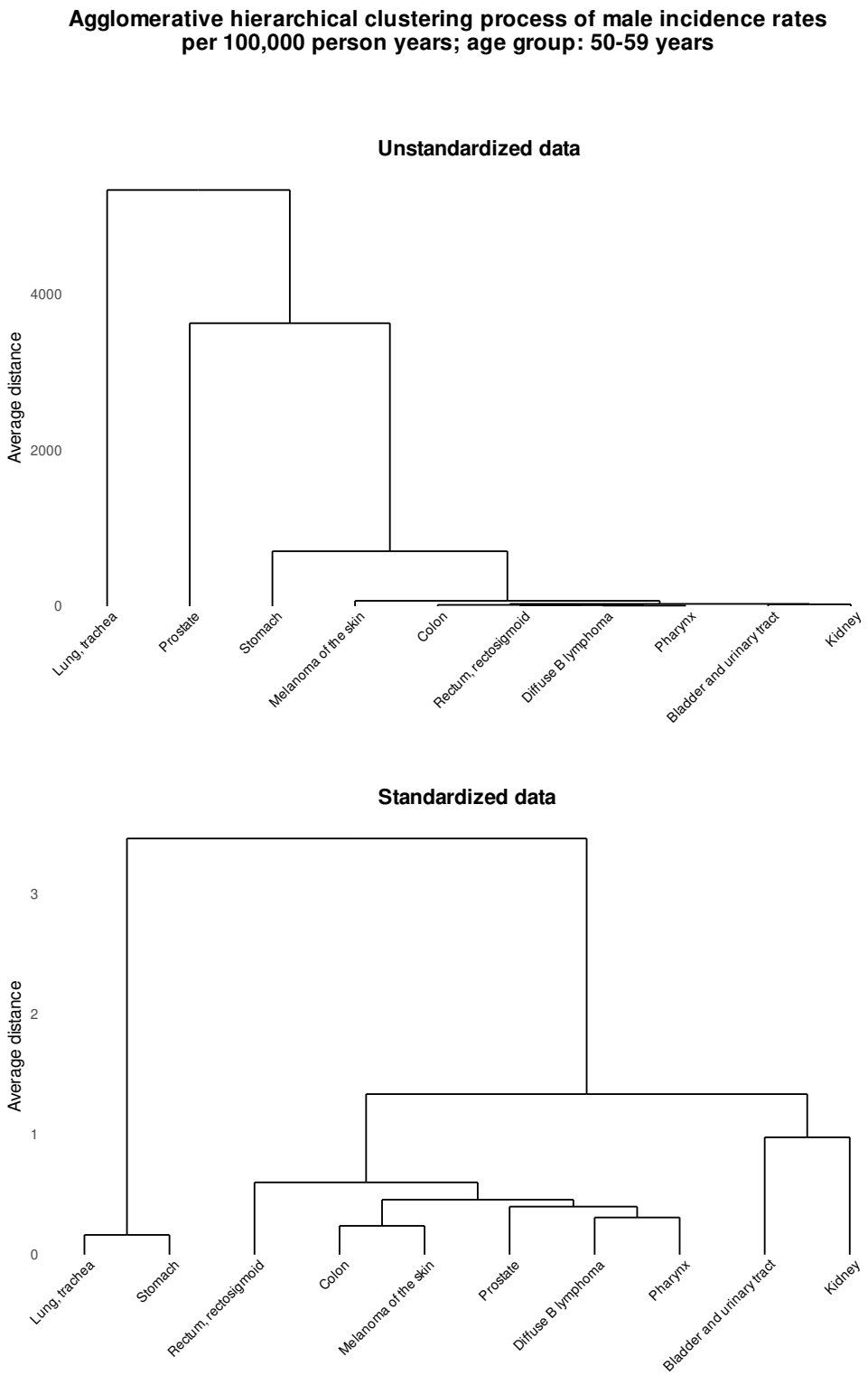


Figure C9: Dendrograms of clustered unstandardized and standardized cancer incidence data of the most common cancers among males aged 50-59 years in Finland from 1962 to 2022.

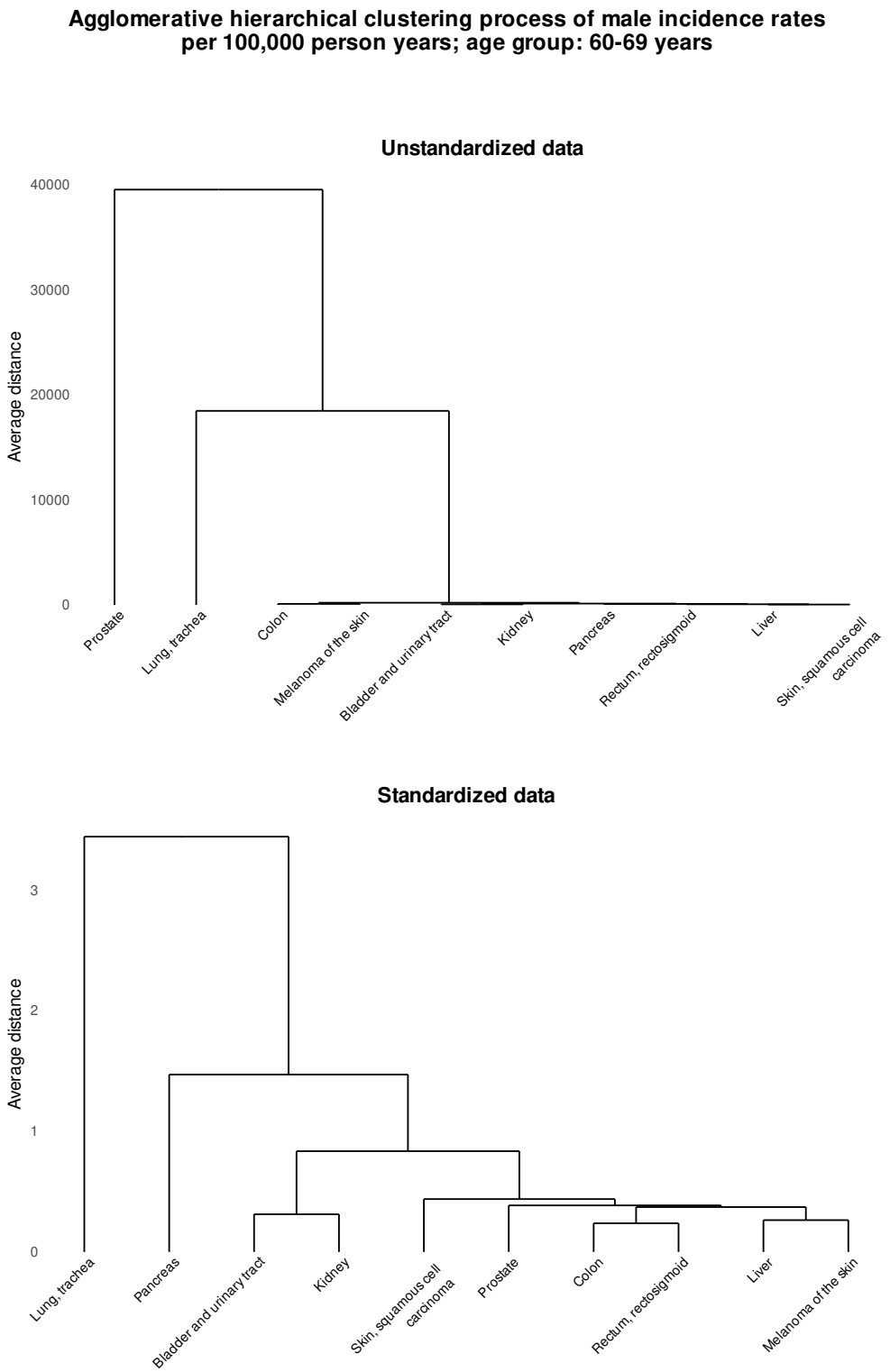


Figure C10: Dendrograms of clustered unstandardized and standardized cancer incidence data of the most common cancers among males aged 60-69 years in Finland from 1962 to 2022.

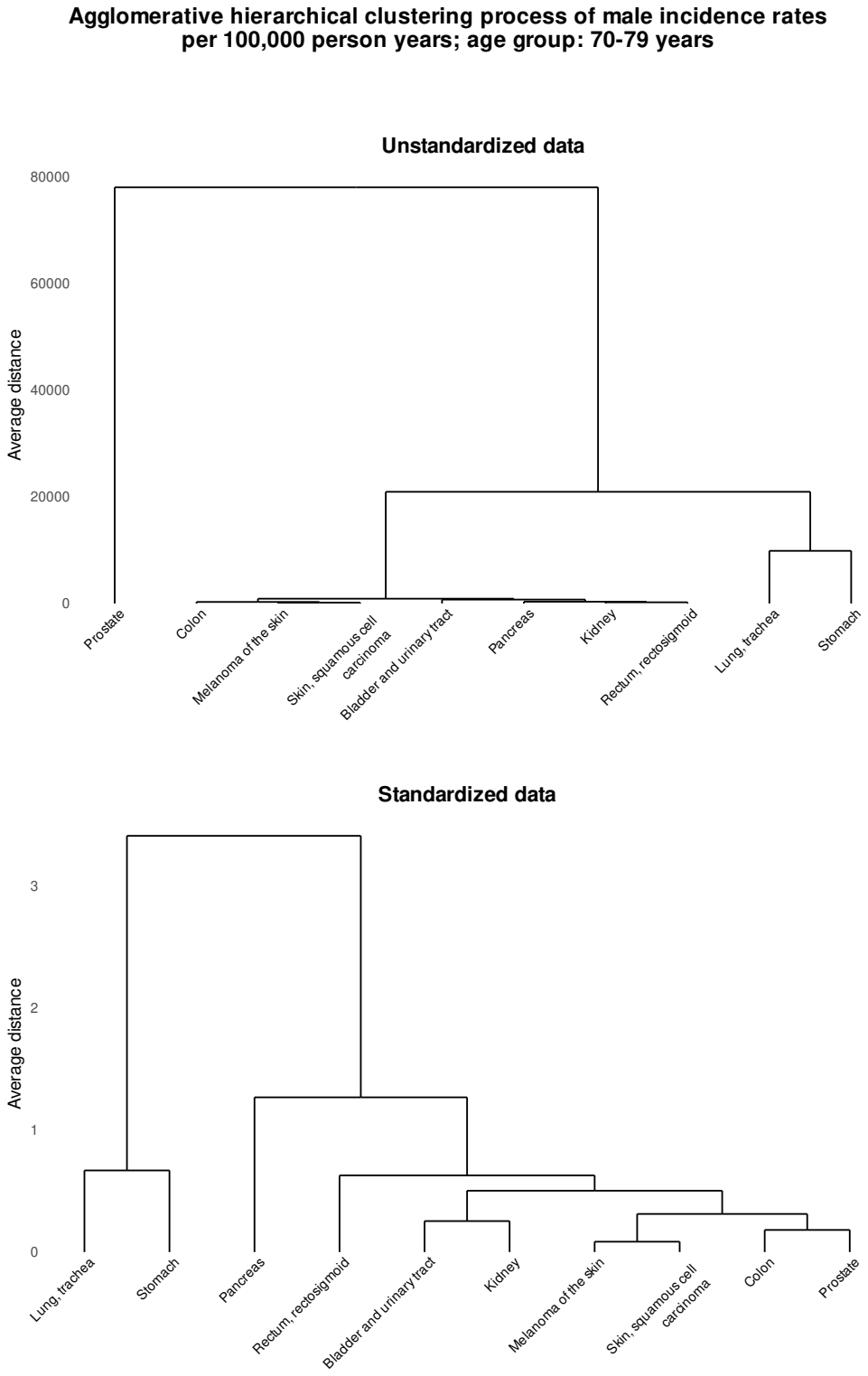


Figure C11: Dendrograms of clustered unstandardized and standardized cancer incidence data of the most common cancers among males aged 70-79 years in Finland from 1962 to 2022.

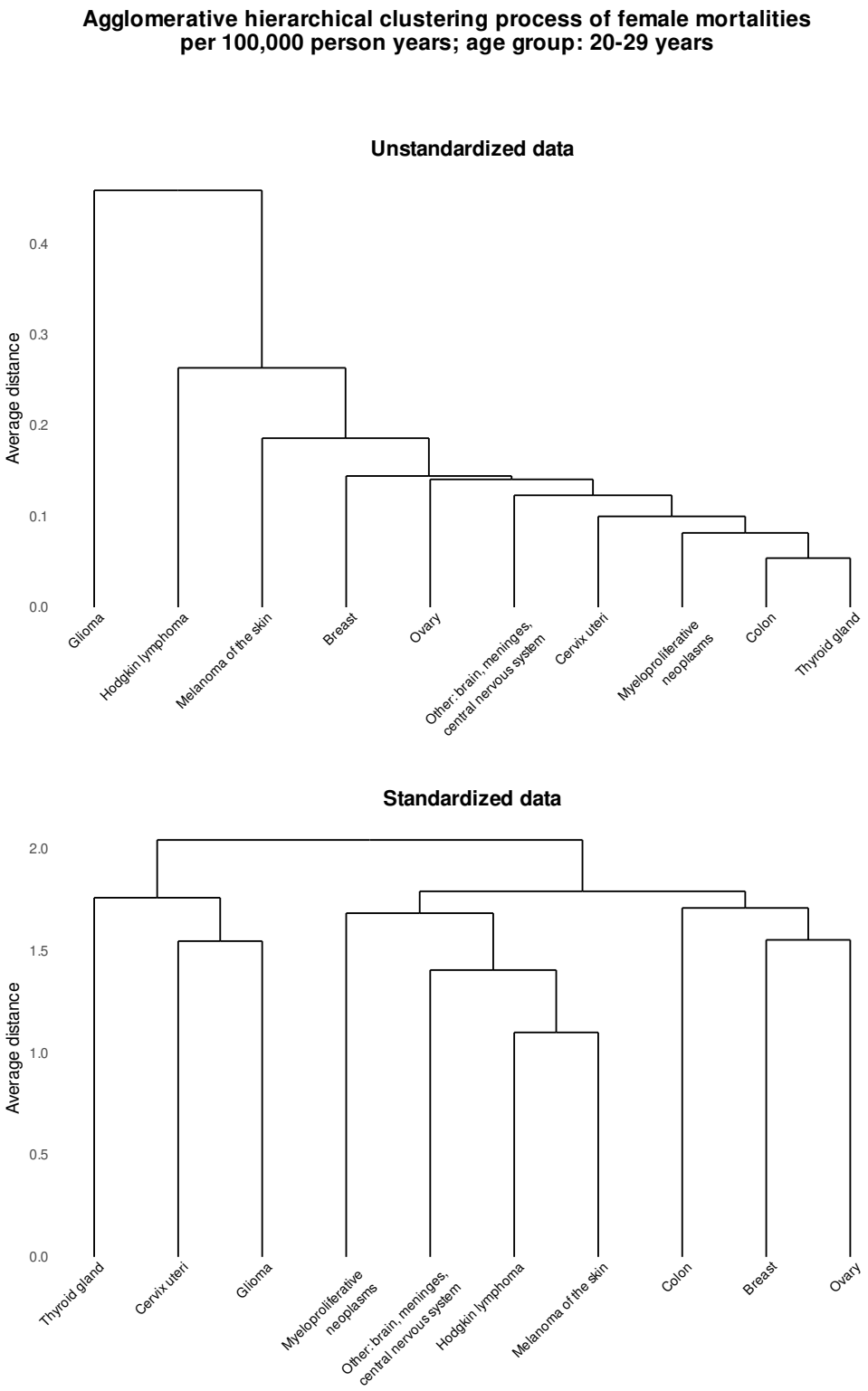


Figure C12: Dendrograms of clustered unstandardized and standardized cancer mortality data of the most common cancers among females aged 20-29 years in Finland from 1962 to 2022.

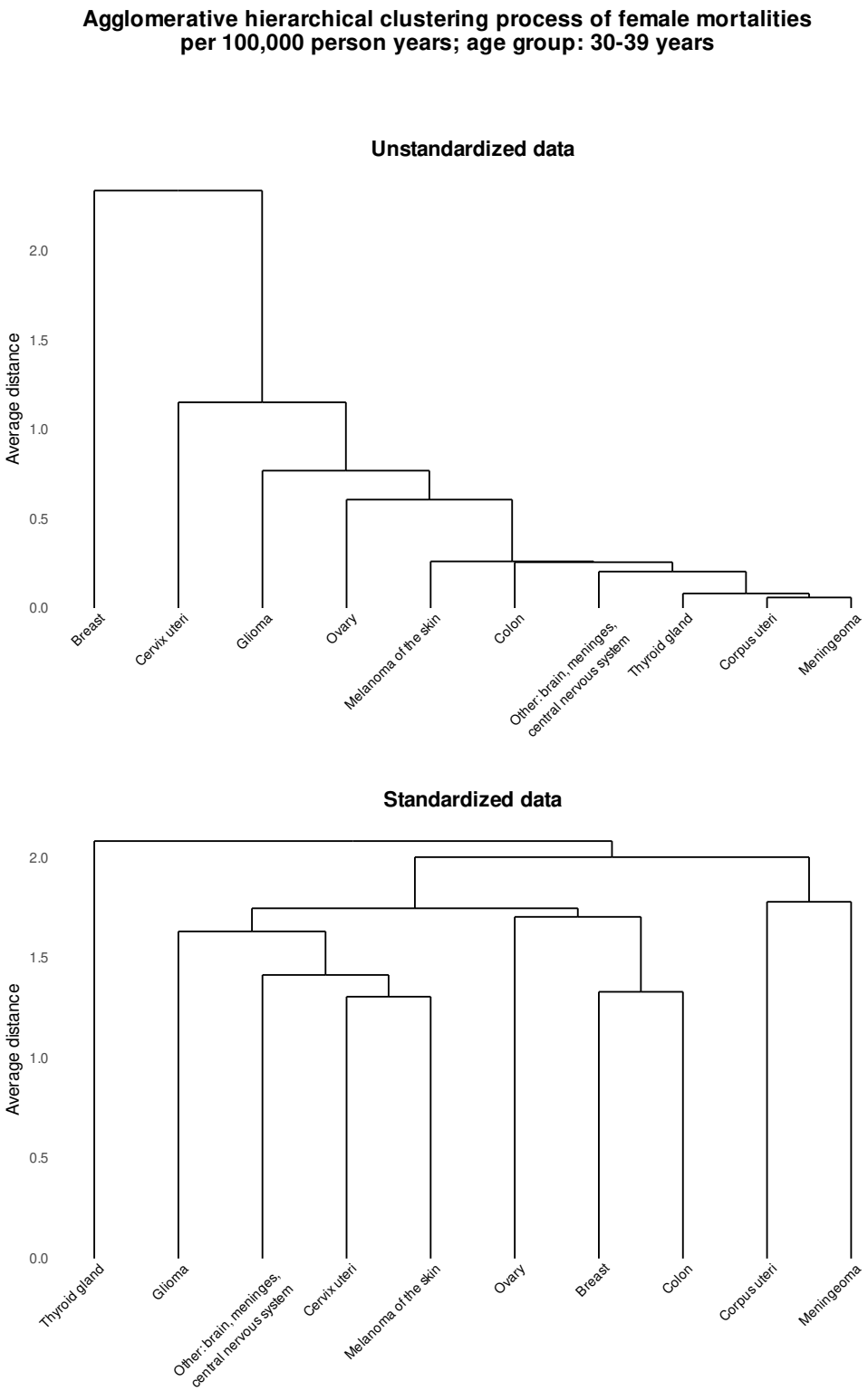


Figure C13: Dendrograms of clustered unstandardized and standardized cancer mortality data of the most common cancers among females aged 30-39 years in Finland from 1962 to 2022.

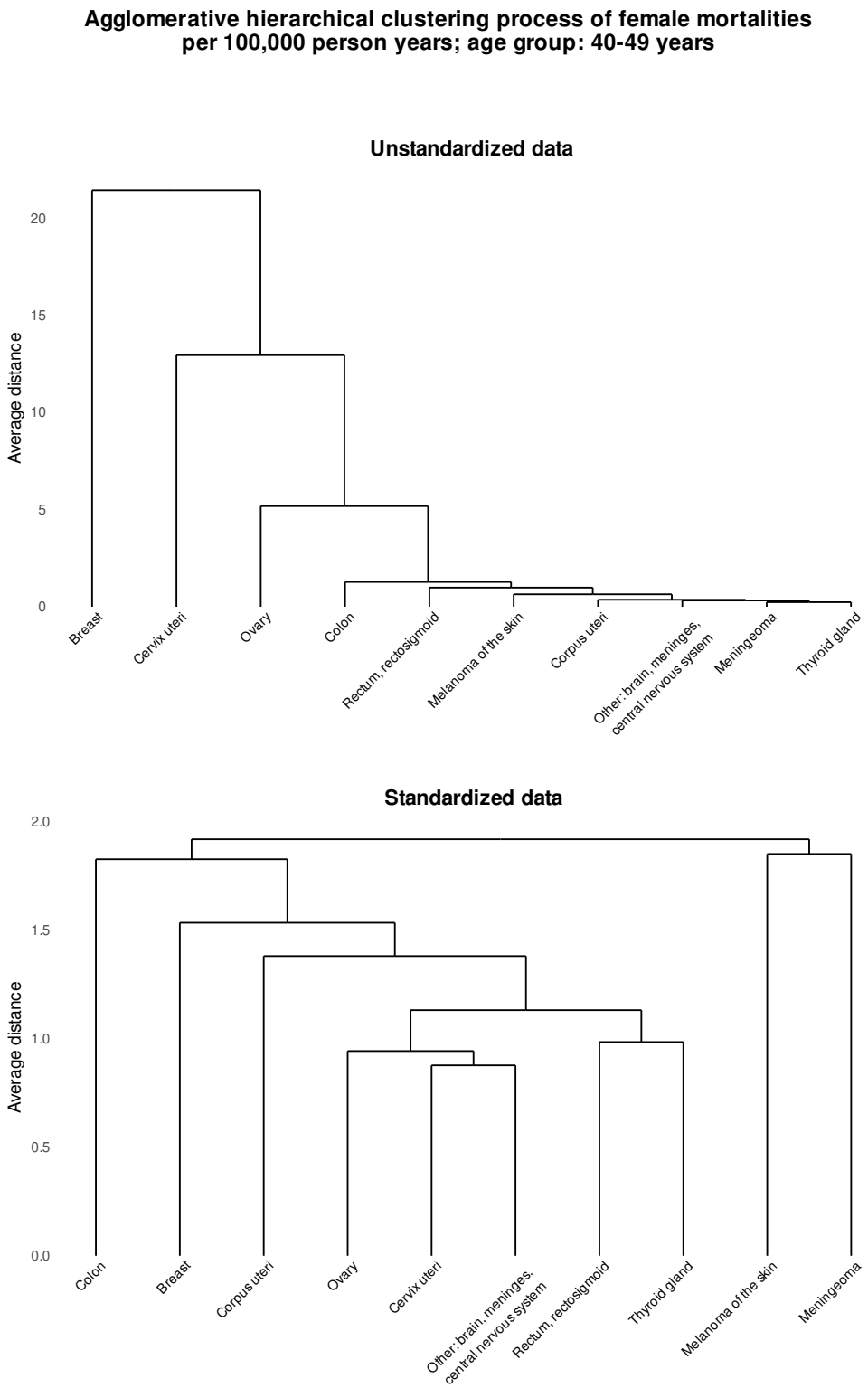


Figure C14: Dendrograms of clustered unstandardized and standardized cancer mortality data of the most common cancers among females aged 40-49 years in Finland from 1962 to 2022.

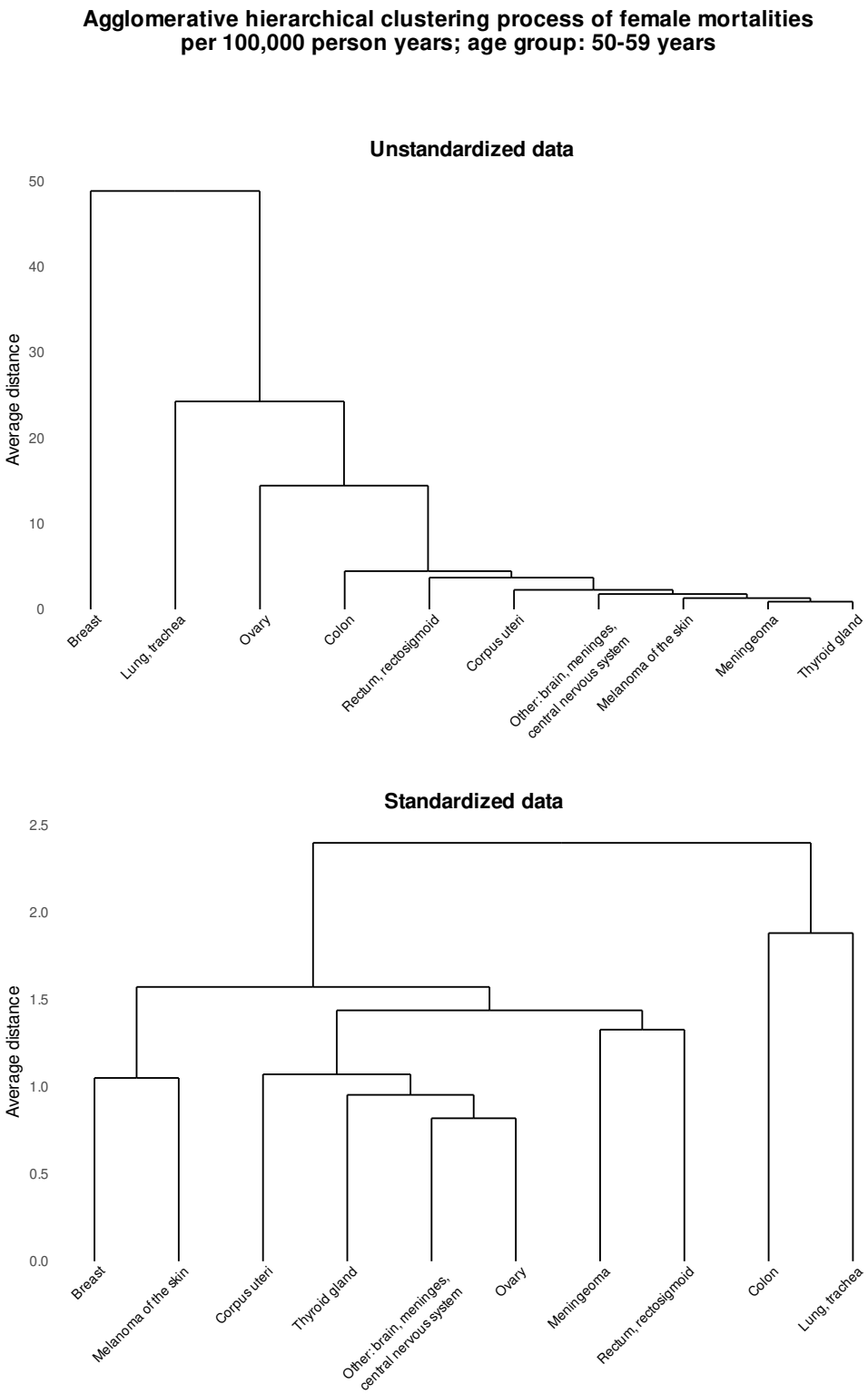


Figure C15: Dendrograms of clustered unstandardized and standardized cancer mortality data of the most common cancers among females aged 50-59 years in Finland from 1962 to 2022.

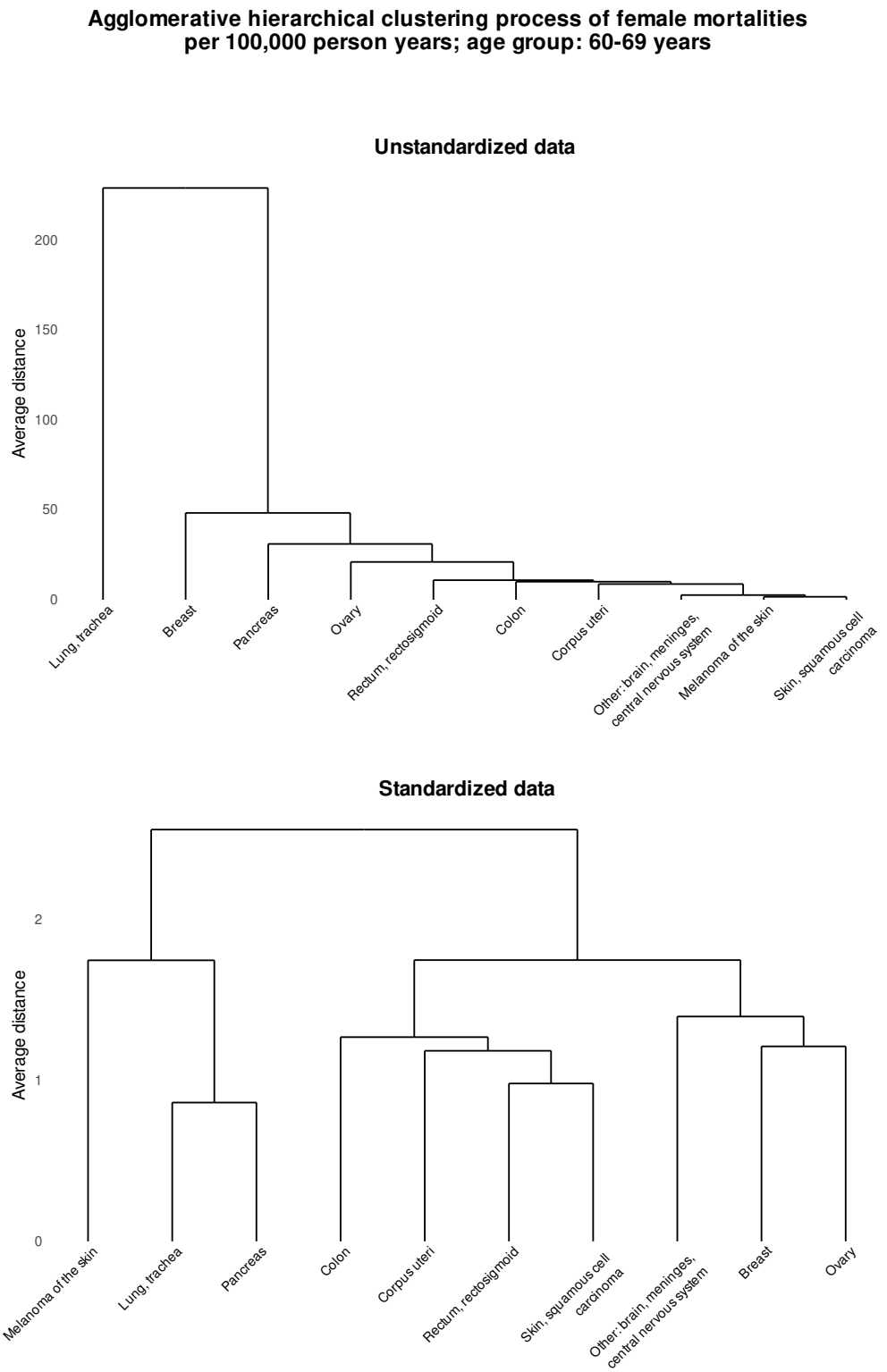


Figure C16: Dendrograms of clustered unstandardized and standardized cancer mortality data of the most common cancers among females aged 60-69 years in Finland from 1962 to 2022.

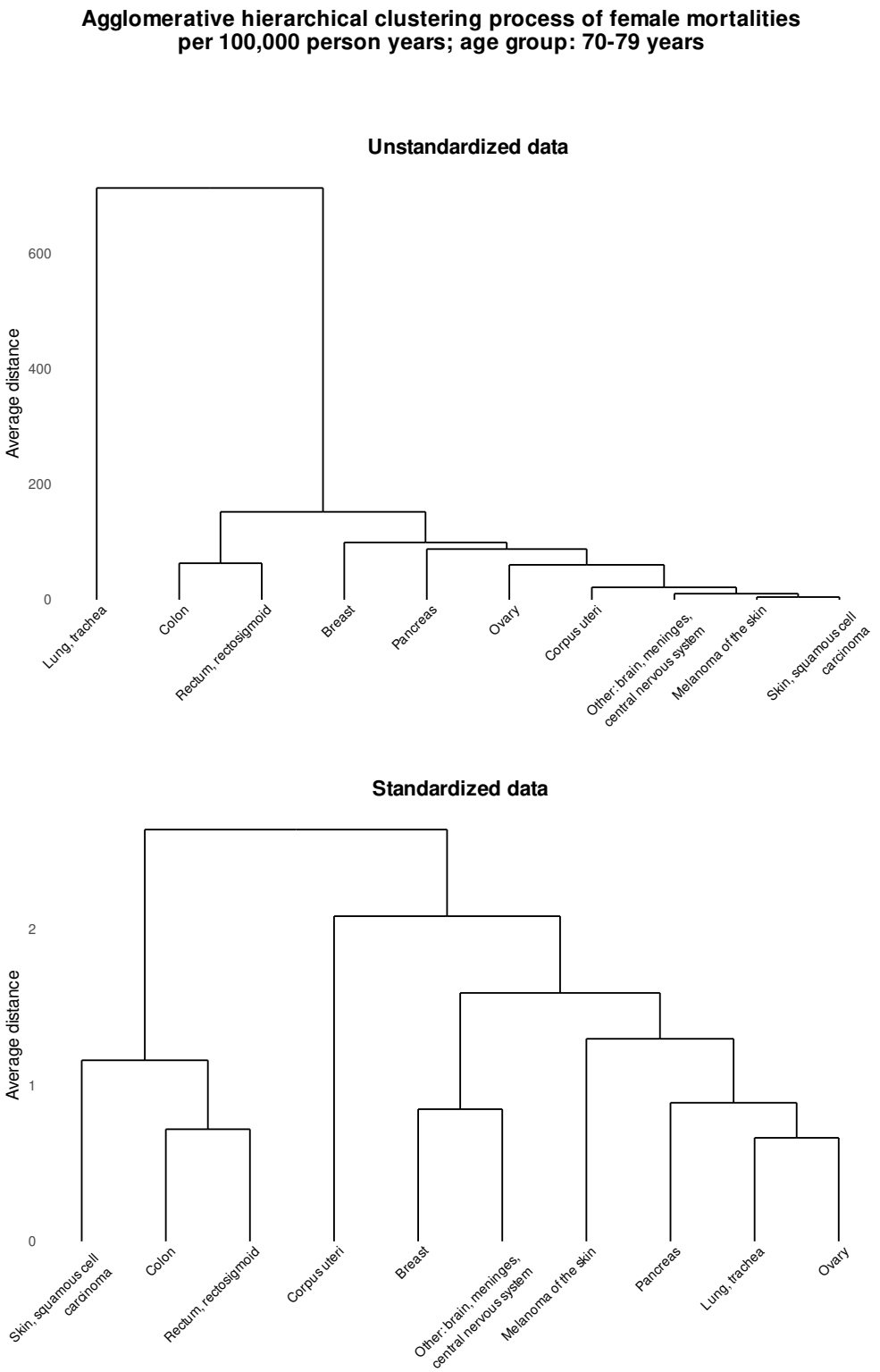


Figure C17: Dendrograms of clustered unstandardized and standardized cancer mortality data of the most common cancers among females aged 70-79 years in Finland from 1962 to 2022.

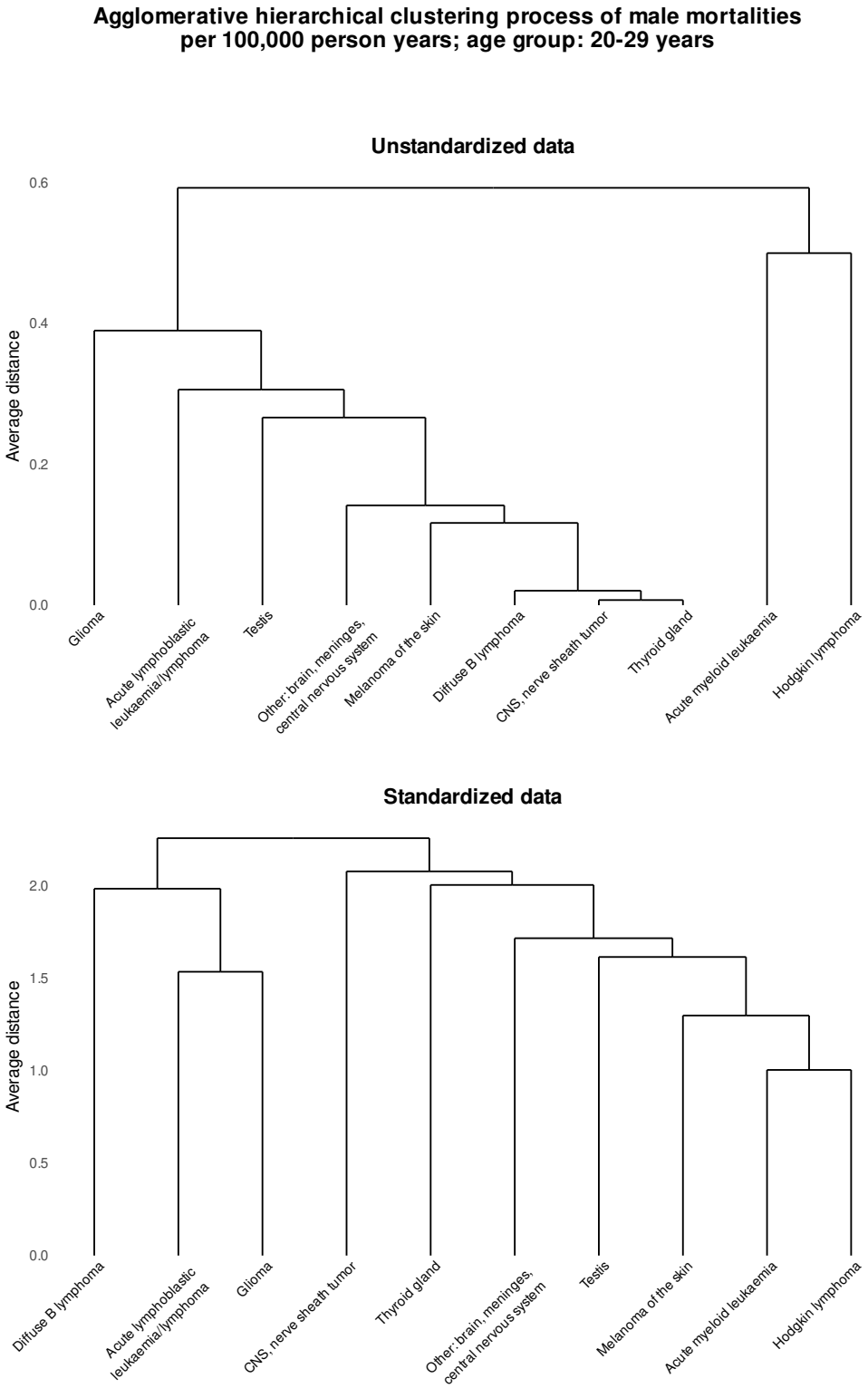


Figure C18: Dendrograms of clustered unstandardized and standardized cancer mortality data of the most common cancers among males aged 20-29 years in Finland from 1962 to 2022.

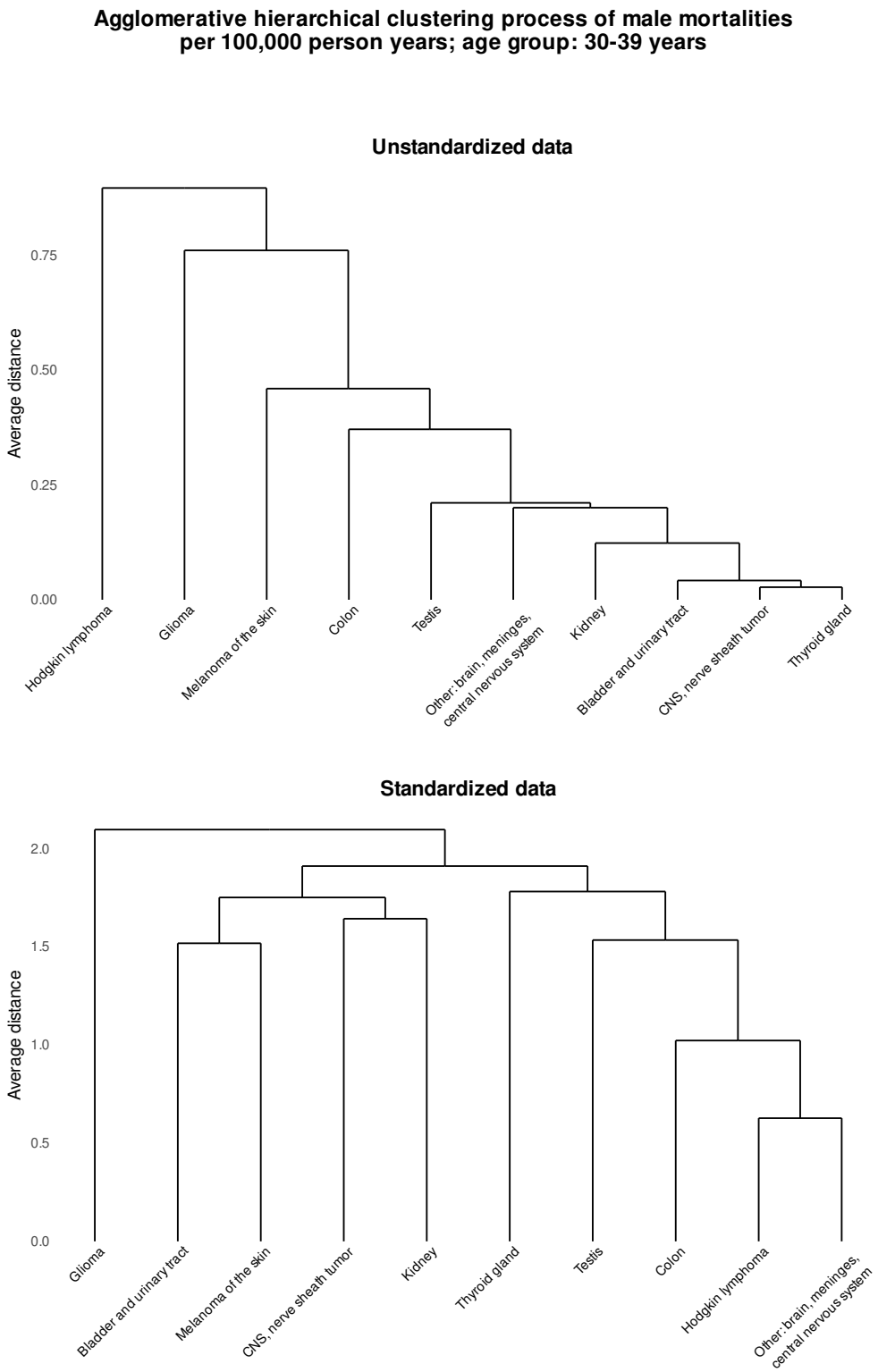


Figure C19: Dendrograms of clustered unstandardized and standardized cancer mortality data of the most common cancers among males aged 30-39 years in Finland from 1962 to 2022.

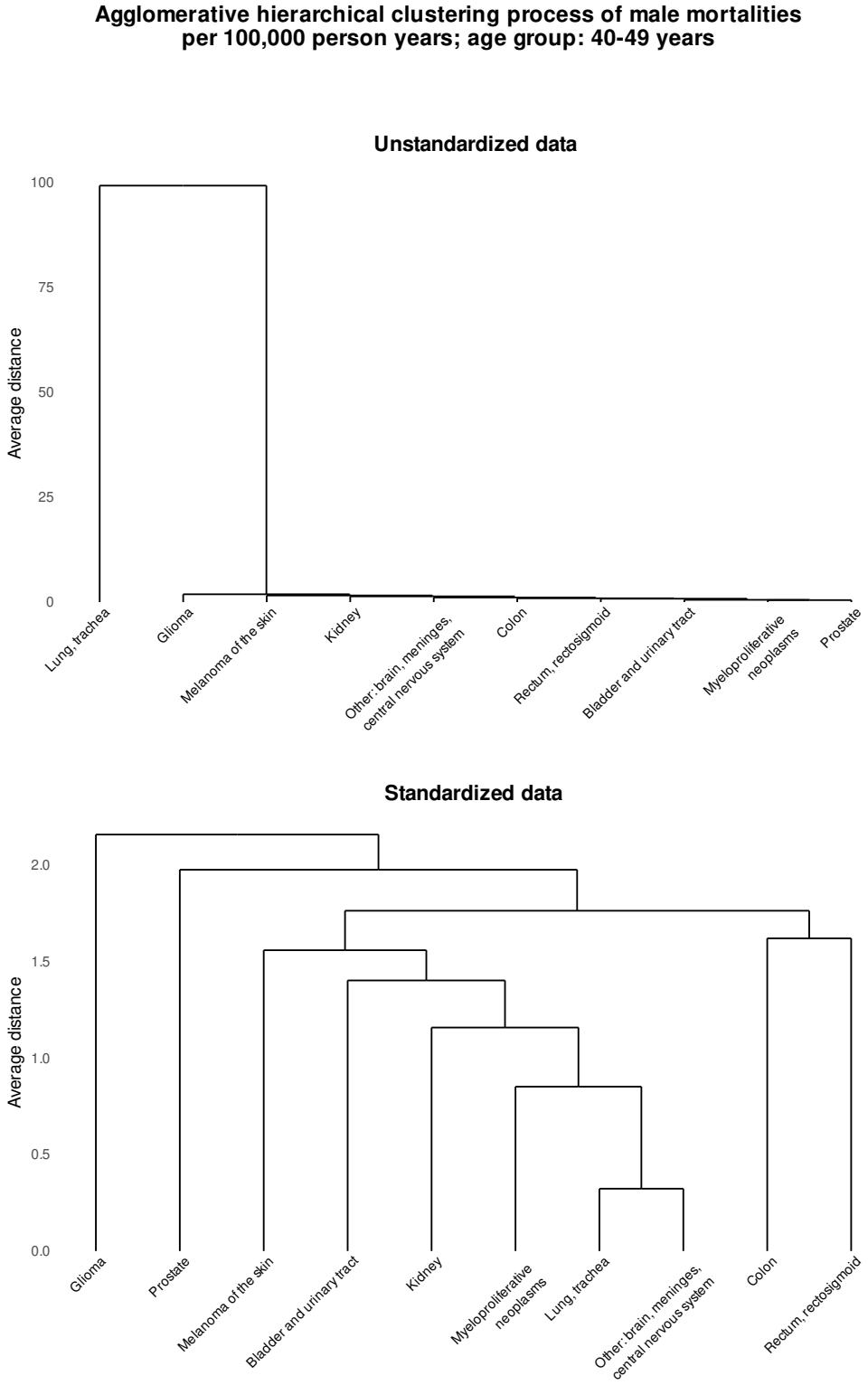


Figure C20: Dendrograms of clustered unstandardized and standardized cancer mortality data of the most common cancers among males aged 40-49 years in Finland from 1962 to 2022.

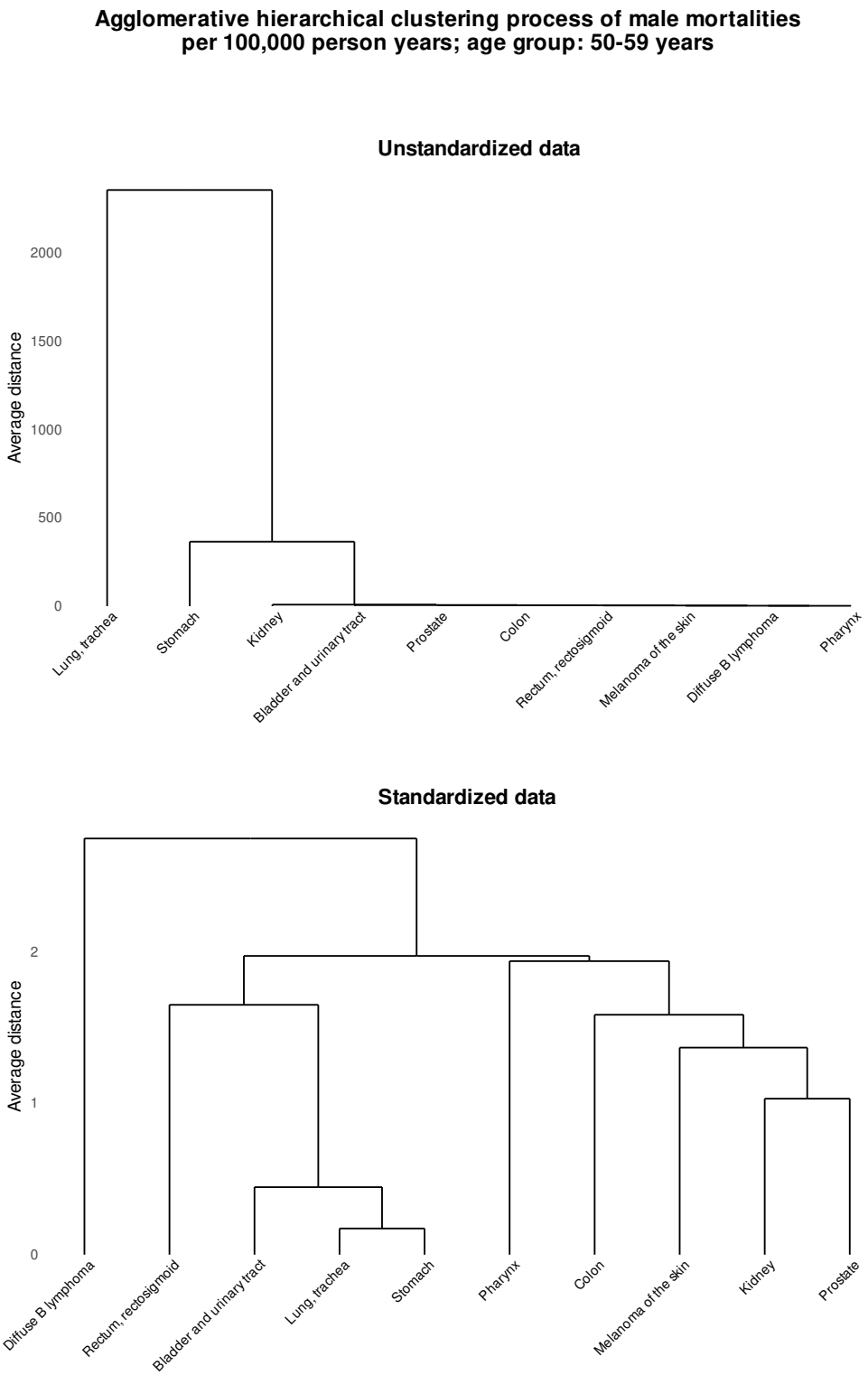


Figure C21: Dendrograms of clustered unstandardized and standardized cancer mortality data of the most common cancers among males aged 50-59 years in Finland from 1962 to 2022.

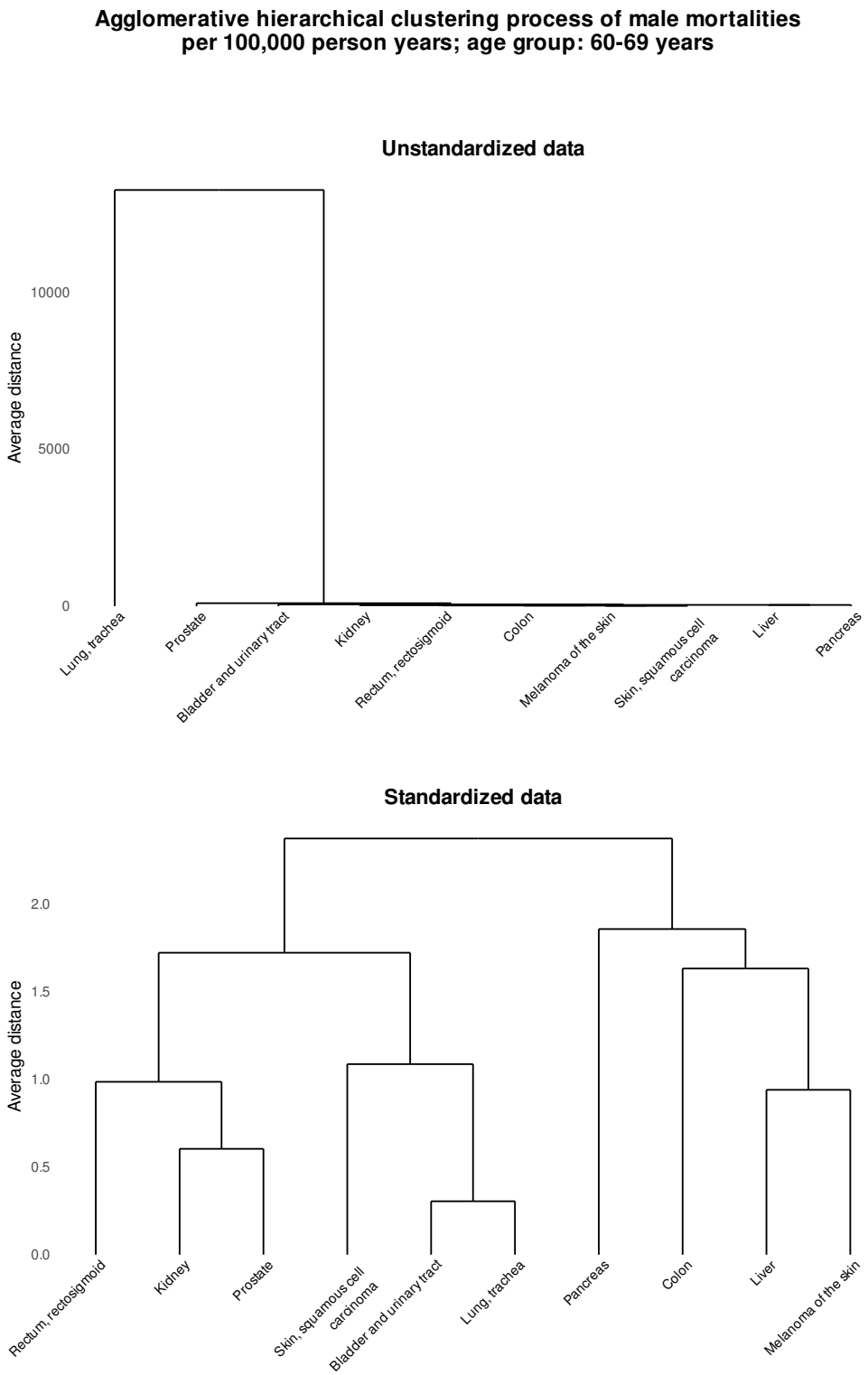


Figure C22: Dendrograms of clustered unstandardized and standardized cancer mortality data of the most common cancers among males aged 60-69 years in Finland from 1962 to 2022.

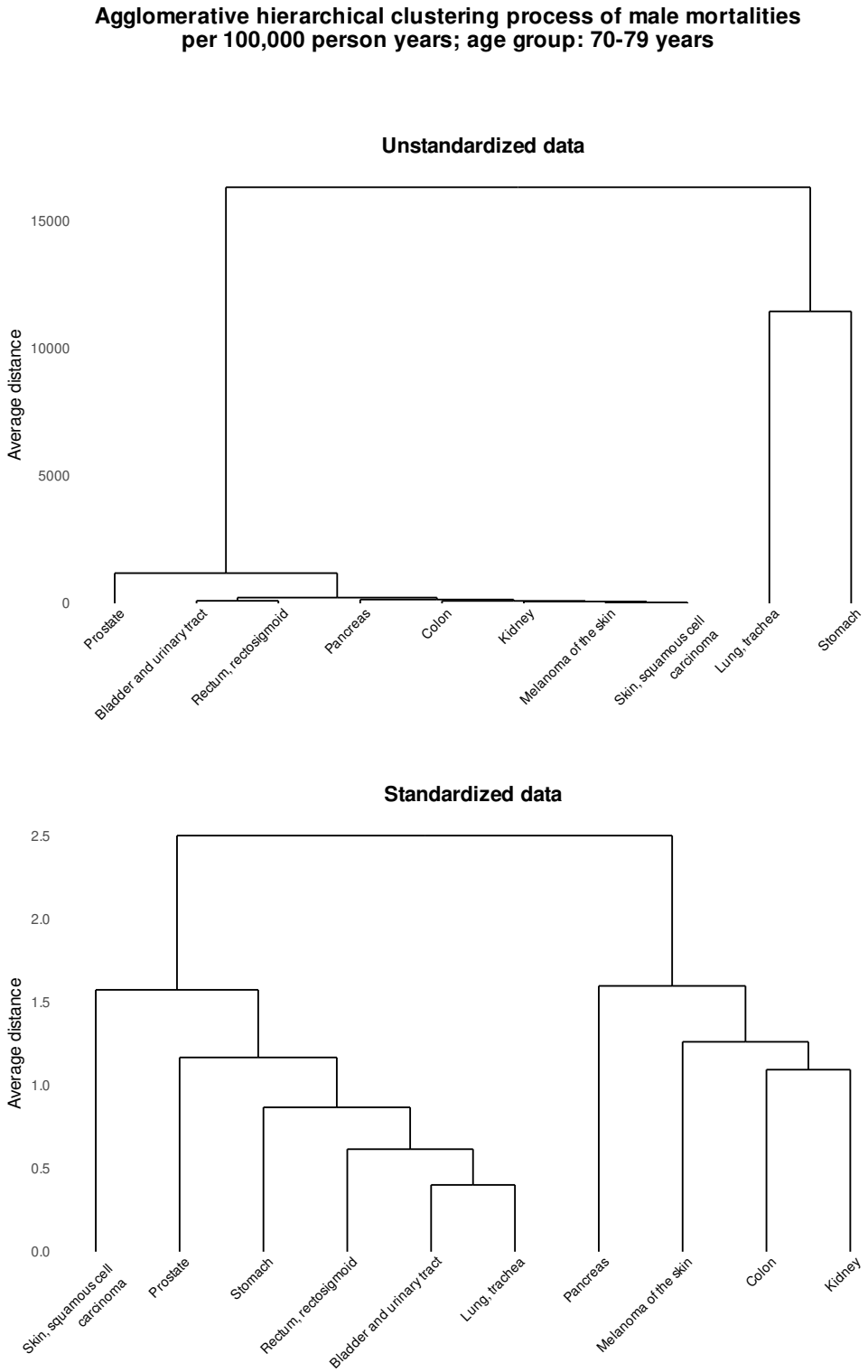


Figure C23: Dendrograms of clustered unstandardized and standardized cancer mortality data of the most common cancers among males aged 70-79 years in Finland from 1962 to 2022.