UNIVERSITY OF SOUTHAMPTON

FACULTY OF SOCIAL SCIENCES

Social Statistics and Demography

Ethnic Differences in Life Expectancy, Mortality and Cause of Death in Colombia: Analysis of Census and Death Register Data

by

John Jairo Loja Torres

Thesis for the degree of Doctor of Philosophy

June 2024

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF SOCIAL SCIENCES

Social Statistics and Demography

Thesis for the degree of Doctor of Philosophy

ETHNIC DIFFERENCES IN LIFE EXPECTANCY, MORTALITY AND CAUSE OF DEATH IN COLOMBIA: ANALYSIS OF CENSUS AND DEATH REGISTER DATA

by John Jairo Loja Torres

This thesis uses data from census and death registers to undertake a systematic investigation of mortality differentials among ethnic population groups in Colombia, comparing Indigenous, Afro-descendant, and White-Mestizo ethnic groups. The main findings of this project are summarised across three dimensions: assessment of quality of mortality data, population life expectancy, and causes of death. Additionally, the study applies machine learning methods to identify the ethnic identity of deceased people based on demographic and epidemiological information derived from mortality registers.

Social inequalities among Indigenous, Afro-descendants, and White-Mestizos in Colombia are well documented in research studies. However, there is little understanding of how those ethnic inequalities manifest in mortality statistics in terms of data collection biases and analysis. This thesis will address the aforementioned knowledge gap by analysing the ethnic differentials in mortality and causes of death data comparing the census data with mortality registration records. Four research questions are addressed in this study. (i) How do life expectancy and mortality rates vary among different ethnic groups by age and sex? (ii) To what extent under-registration of deaths influences the estimation of life expectancy and mortality? (iii) Which causes of death contribute most to the differences in life expectancy and mortality among different ethnic groups? (iv) How can we apply machine learning techniques on mortality register data to improve ethnic classification in Colombia?

The first analytical chapter of the thesis aims to address under-registration of deaths, and how they affect mortality estimations in Colombia, and particularly how they under-estimate mortality indicators of Indigenous and Afro-descendants groups. The results show that the deaths of Indigenous and Afro-descendants are registered at a considerably lower rate than those of White-Mestizos, and the level of under-registration is higher for children and residents in rural areas. The ethnic minority groups are less likely to be captured in mortality registration, attributed to various social, economic, and geographic factors such as residence in remote areas, cultural practices and inhibitions, lack of awareness and the absence of institutional systems in regions dominated by ethnic minorities.

In the second analysis, we applied statistical models to correct for mortality underestimation in the mortality registers, and revised the estimates of mortality outcomes including life expectancy and life inequality defined as the average years of life lost due to early deaths. These measurements were estimated by age, gender and ethnicity. According to Colombian vital statistics, White-Mestizos show higher mortality and lower life expectancy when compared to Indigenous and Afro-descendants. However, the results show a reversal trend when mortality data are corrected for under-registration. Life expectancy at birth for Indigenous is estimated at 66.7 years, 71.5 years for Afro-descendants, and 78.9 years for White-Mestizos. Overall, White-Mestizos live on average 12.2 years longer than Indigenous and 7.4 years longer than Afro-descendants. The higher mortality in ethnic minority groups is evident and consistent when estimated using census mortality data and corrected mortality records. Similarly, the dispersion of the age of death in Indigenous and Afro-descendants is higher than for White-Mestizos, whose age of death is narrowly concentrated in the older age groups. This suggests that life disparity within ethnic groups is higher when compared to White-Mestizos.

The third analytical chapter of the thesis identifies the main causes of death by ethnic group, age, and gender. The analysis demonstrated three different patterns based on causes of death by age, sex and ethnicity. The first is the White-Mestizo pattern in which the main causes of death are degenerative diseases such as cardiovascular diseases, cancer, and respiratory diseases. The age groups most affected by such causes are the older age groups, in which the majority of such deaths are concentrated. The second pattern is exhibited primarily by the Afro-descendants, and combines very high mortality of young men between ages 15 and 40 (principally by homicide) with an emerging ageing population with increased deaths attributed to degenerative diseases. Deaths of Afro-descendants present a bimodal distribution with high mortality at younger and older ages. The epidemiological patterns of the Indigenous group, on the other hand, present high mortality in newborns and children compared with death at other ages. Likewise, the Indigenous groups had the highest proportion of deaths in women of reproductive age due to pregnancy, childbirth, and puerperium-related causes. The causes of death more common in this ethnic group are those related to the perinatal period, and nutritional deficiencies in the case of children and circulatory diseases, cancer, and respiratory diseases in death at older ages. This pattern suggests that women belonging to indigenous groups are more vulnerable in terms of adverse living conditions with poor access to essential maternal and infant care.

The final analytical part of the thesis applies machine learning techniques on mortality records to improve the classification of ethnic identities in Colombia. We proposed a random forest classification model which considers geographical, sociodemographic, and epidemiological information by ethnicity. The results show that random forest models predict the ethnicity of deceased individuals more accurately with probabilities over 80%. The model yielded AUC = 0.913 for Indigenous people, while for Afro-descendants and White-Mestizos AUC scores were 0.811 and 0.845 respectively, suggesting a high level of accuracy in terms of correctly identifying the ethnicity of deceased people.

This research has two key limitations. First, the 2018 Census did not account for more than four million people, of which at least 1.5 million were Afro-descendants. The census omission of Afro-descendant households could lead to potential underestimation of mortality outcomes and hence the differences between ethnic groups may be biased. Second, there is lack of adequate information of the deceased people. The census asks if reported deaths were registered, but does not collect information about the causes of death. We were able to identify the level of under-registration but we could not determine if the causes of death play a role in this phenomenon. It is also possible the internal conflicts in certain geographic regions within Colombia may have contributed to the under-registration of deaths in ethnic groups.

Three recommendations are drawn from the findings of this study. First, the ethnic minority groups would need to be empowered and included in social policies and local decision-making processes and governance structures. They should be adequately trained and informed of the relevance and benefits of vital statistics for local planning, with measures to overcome cultural sensitivities and other barriers including language and geographic isolation. Second, census and surveys need to include additional questions to probe details of possible circumstances underlying the causes of death, verbal autopsies and deaths with no funerals (missing bodies). Finally, there is a dire need to consider health policy interventions to prevent mortality in women and children attributed to malnutrition and infectious diseases especially during the perinatal period, targeting the indigenous community. For Afro-descendants, policy efforts should focus on reducing homicide mortality in young populations.

Table of Contents

List of Figures

List of Tables

Declaration of Authorship

I, John Jairo Loja Torres, declare that the thesis entitled *Ethnic Differences in Life Expectancy, Mortality and Cause of Death in Colombia: Analysis of Census and Death Register Data* and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted form the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;

Signed: ....

Date: ....

Acknowledgements

I express my heartfelt appreciation and gratitude to my supervisors Professor Sabu Padmadas and Dr. Jason Hilton for their invaluable support, intellectual contributions, timely feedback, patience, and for creating a warm and familiar working environment. The last few years have been intellectually stimulating and thank you Dr. Jason Hilton for your relentless support which I will cherish for the rest of my life, and Professor Sabu Padmadas for the great leadership full of motivation and optimism to face each of the inconveniences along the project. I extend my sincere thanks to Dr. Andrew Hinde who co-supervised the first year of the project with Professor Sabu Padmadas, who encouraged me to take the challenges and help me settle well in Southampton. I express my enormous gratitude to the Professors Jakub Bijak and Peter Smith for their thoughtful comments and constructive suggestions, and to Professor Laia Becares for taking the time to review and evaluate my thesis.

This research could not have been undertaken without the support of the Economic & Social Research Council (ESRC) and the South Coast Doctoral Training Partnership (SCDTP), to whom I shall remain extremely grateful for funding my PhD project. I am also grateful to Athina Vlachantoni, Director of SCDTP, and the rest of the SCDTP Team for their support throughout the course of my project. I am deeply indebted to the Departamento Nacional de Estadistica (DANE) [Colombian National Statistics Office] and the teams of Estadísticas Vitales, Census y Demografía, and Sala de Procesamiento Especializado Externo for their encouragement and providing the information and technical support, especially thanks to Edna Margarita Valle Cabrera, Angela Patricia Vega Landaeta, y Laura Tatiana Ocampo Isaza.

I gratefully acknowledge the invaluable feedback and brilliant ideas from seminars, summer schools, and conferences where I took part: I Congreso Colombiano de Demografia, Annual Meeting South Coast Doctoral Training Partnership (SCDTP), 13th Conference of Young Demographers, European Population Conference 2022, PopFest 2022, R Programming- BaRcelona Summer School of

Demography (BSSD), Data Visualization - the art/skill cocktail, Rostock Retreat 2021: Visualizing Uncertainty, and Matrix Approaches to Modelling Kinship: Theory and Applications. I would also like to offer my special thanks to Dmitri A. Jdanov for the invitation to present and discuss my research at the Laboratory of Demographic Data at the Max Planck Institute for Demographic Research.

I am also thankful to staff working across University, including in cleaning, library, iSolutions, and secretarial units for facilitating the smooth running of the project. In particular Jane Parsons and Claire Caffrey for their outstanding job. My sincere thanks to colleagues and friends for the interesting discussions and the cherished time spent together at Southampton. In particular, many thanks to Dr. Julie Vullnetari, Andrea Obando, Dr. Giuseppe Troccoli, and Michele Zadra for their insightful comments, suggestions, and for their generous support during these years.

Finally, my gratitude goes out to my family for their love, constant encouragement and unconditional support. To all those who contribute to this intellectually stimulating and exciting journey, my love and sincere gratitude.

CHAPTER 1

Introduction

This thesis contributes to a systematic understanding of trends, differentials and causes of mortality in Colombia for the period between 2008 and 2019, and how mortality patterns for Indigenous, Afro-descendants, and White-Mestizo groups, differ from general mortality trends in the country.

1.1 Research Problem

The existence of mortality differences by regions in Colombia is an indicator of ethnic health inequalities. However, although people in Indigenous and Afro-descendant regions are at higher risk of dying prematurely, the true population level of mortality, life expectancy and associated causes of death are largely unknown or not systematically investigated. The lack of scientific evidence in this regard implies lack of evidence-based effective public health policies for reducing health inequalities in Colombia, and more importantly the lack of affirmative action policies directed to prevent premature deaths in the most vulnerable groups.

The analysis of ethnic mortality is key to understanding how socioeconomic inequalities are reflected into health differences in the Colombian case. This research tackles three fundamental aspects that examine the problem from different perspectives: ethnicity related data problems, mortality estimation, and causes of death. A serious challenge for ethnic health studies has been the lack of information and poor quality data on ethnic minority populations. Therefore, the quantification of non-registered deaths by ethnic groups and the ethnic identification of deceased people who died prior to 2008 when ethnic classification was introduced in the mortality registers are major challenges to properly understanding overall mortality trends and patterns. Improving the quality of ethnicity data is vital for public health policies in multi-ethnic countries where official statistics are fragmented or incomplete.

On the other hand, a systematic examination of the mortality trends and causes of death of the Colombian population, particularly of Indigenous and Afro-descendants, will allow a better understanding of the population-level changes driven by mortality patterns and leading causes of deaths among ethnic groups, as well as help identifying the most affected geographic regions in the country. Addressing this evidence and knowledge gap is crucial for the design and improvement of social and public health policies.

In recent years, racial justice movements have drawn public attention and significance in terms of debates on social policy and anti-discriminatory programmes. At the same time, national statistics offices have started to collect and improve ethnic information in population censuses and demographic surveys. In Colombia, although the Censuses from 1993 and 2005 collected data on ethnicity, the variables were included in mortality registers only from 2008 onwards. This research aims to address the issues described above, and provide robust model-based estimates to quantify the ethnic differences in mortality in Colombia. Moreover, the findings of this research will enable us to better understand the importance of ethnicity in quantifying inequalities in population health outcomes. Policy makers, therefore, can directly benefit from the findings of this study in formulating appropriate national health policies and building a fair and inclusive multi-ethnic society.

Overall, this thesis contributes to a better understanding of ethnic mortality differences and provides an earnest effort to strengthen evidence on health and ethnicity in Colombia by implementing innovative statistical methodologies for improving data quality in societies where mortality data of ethnic groups are incomplete, defective, or unavailable.

1.2 Background

Since colonial times in the late 15th century and post-independence in 1819, race and ethnicity have been identified as key determinants of social inequalities across Latin American societies. In Colombia the Indigenous and Afro-descendant population have been not only socially excluded, but also geographically segregated in outlying regions and poor urban areas (Barbary and Urrea, 2004). According to National Statistics Office, DANE, the departments of Vaupés, Guainía, Vichada, and Amazonas have the highest proportion of Indigenous population with 81.7%, 74.9%, 58.2%, and 57.7% of their population reporting Indigenous identity, respectively. In contrast, the largest number of Afro-descendants are found in Chocó, Bolivar, Valle del Cauca, and Antioquia with 63.1%, 15.4%, 14.5%, and 4.9% of the people reporting Afro background, respectively (Departamento-Administrativo-Nacional-De-Estadística, 2019a,b), see Figure 1.1.

Figure 1.1: Colombian Regions with Higher Proportions of Ethnic Population

Figure 1.1 shows the Colombian regions with higher proportion of Afro-descendant and Indigenous population. According the 2018 Census, the lined department Vaupes, Guaninía, Vichada and Amazonas have the largest Indigenous population in the country. While the majority of Afro-descendants are located in Valle del Cauca, Chocó, Bolivar, and Antioquia. These estimations are based in absolute numbers that did not considered regions with percentual majority of Indigenous or Afro-descendants but smaller populations.
Source: Own elaboration based on 2018 Census data.

Colombian society comprises three main ethnic groups whose socioeconomic statuses were established during colonial times. White-Mestizos, who are descendants of Spaniards and more recently of other European migrants, have been at the top of the social pyramid. In contrast, Indigenous peoples, who are descendants of the aboriginal population, and Afro-descendants, who are descendants of enslaved populations brought from Africa, have historically occupied lower positions. Indigenous and Afro-descendants groups are socially and economically more deprived than White-Mestizos (Viáfara, 2017; Telles et al., 2015; Flórez et al., 2003; Wade, 1993a). For instance, infant mortality is higher along the Pacific coast, in the north and in the south of the country where ethnic minority groups are dominant. In the same way, ethnic populations are significantly less likely to have health insurance, almost 50% of them do not have any access to health care system (Bernal and Cárdenas, 2007), and according to self-reported health, ethnic minorities rated worse than the rest of the population (Agudelo-Suárez et al., 2016).

The lower socioeconomic conditions among ethnic groups have likely been influenced by various factors, including structural racism and ideologies of racial superiority, which classify Indigenous and Afro-descendant populations as "uneducated, primitive, and inferior." (Baquero, 2015; Urrea-Giraldo and Hurtado, 2002; Bernard, 1987; Nengwekhulu, 1986; Wade, 1985). Nevertheless, class and race hierarchies are not always concomitant, since there are also non-whites in the Colombian middle-class (Wade, 1993a). But the historical intersection of ethnic groups and lower social classes have been sufficient to create the basic hierarchy of the racial order that structures society at present, in both life and death (Wade, 1993a, p.20).

Social and health inequalities in Colombian society stem from various factors contributing to the existing mortality gap among ethnic groups, with racism and ethnic discrimination standing out as significant drivers. Racial discrimination affects access to economic resources and diminishes opportunities for a healthy life by creating barriers in both public and private institutions. Furthermore, it introduces a racial bias in policy-making processes. The higher rates of poverty, unemployment, and lower educational attainment in ethnic populations significantly impact healthcare access. Ethnic regions historically receive lower levels of public investment and infrastructure, with agriculture and mining being the primary economic activities in these areas. However, these sectors also pose environmental challenges to living conditions.

Indigenous and Afro-descendant communities may reside in areas with poorer environmental conditions, exposing them to air and water pollution, inadequate housing, and challenging geographical and climatic characteristics that increase the risk of vector-borne diseases. Additionally, cultural factors specific to ethnic groups, including traditions related to medicine, diet, and social practices, can significantly influence health outcomes. Language barriers among populations that do not speak the dominant language of the country can affect attitudes towards healthcare institutions and may hinder access to information, medical advice, and services.

The factors contributing to ethnic health inequities in the country have historically been neglected.

The lack of attention from central governments towards the needs of ethnic minorities may stem from the perception that these communities are not a priority for social policy. Additionally, there persists a belief that these communities are uncivilised and accustomed to living in a pre-stage of development. This racist view of ethnic groups could be one of the key drivers of the ethnic social inequalities among ethnic groups in the country.
Indigenous and Afro-descendants are thus at higher risk of dying earlier than White-Mestizo groups. The extent and trends of the mortality differences are, however, unknown, which suggest that public health policies could be neither evidence-based nor effective in reducing health inequalities in Colombia, and interventions to avert premature deaths in the most vulnerable groups are unlikely to be effective without any targeted and affirmative action.

1.3 Research Context

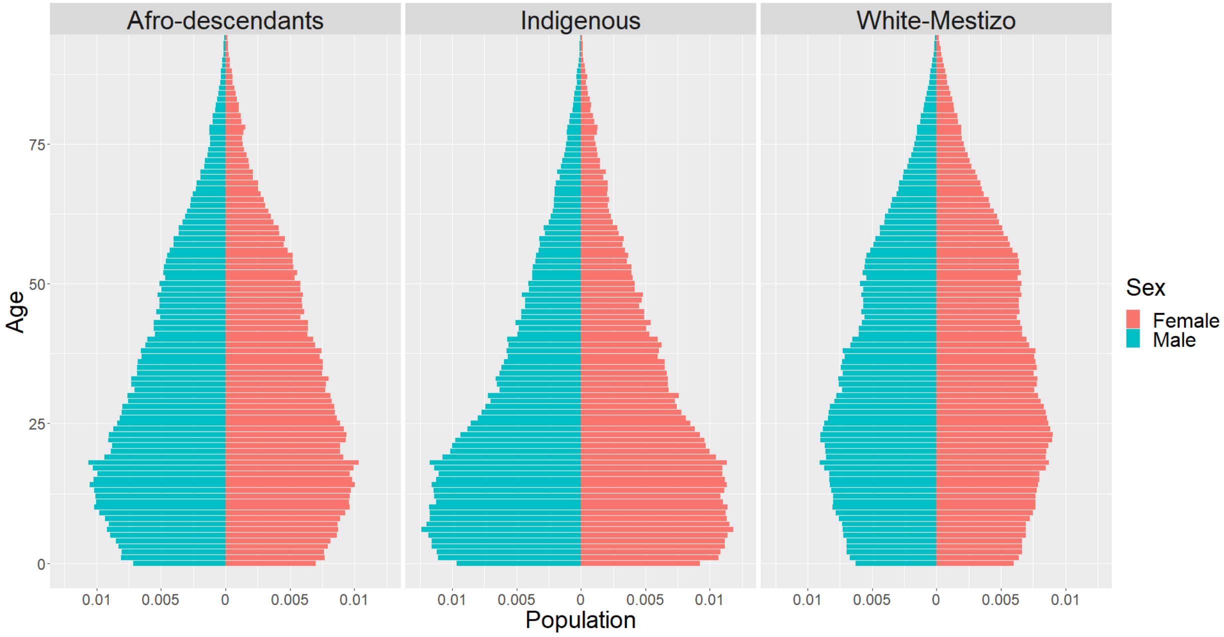
Latin American society was divided during the Spanish colonial period into a racial order in which Indigenous and Africans were at the bottom, although with some exceptions as new ethnic categories emerged, including *Criollos*, *Zambos*, *Mulattos*, and *Mestizos* (McKinley, 2012; Wade, 2010a, 2009)

Population censuses conducted during the Colony in the 16thcentury and after, included the ethnic population but only in order to estimate the workforce of Indigenous and the number of slaves for commercial reasons (Andrews, 2016; Rodríguez, 2010). At the beginning of the 19th century, Colombian censuses started counting the Indigenous population in the country, and subsequently the 1993 Census introduced a question to estimate the Afro-descendant population, after a gap of 150 years since Afro-descendants were last counted in the year 1843 (Bodnar, 2006). However, some researchers contend that Afro-descendants were counted for the first time during the Republic era in 1918 (Rodríguez, 2010, 92). The question on self-reported ethnicity was implemented firstly in 1993 Census and modified for the 2005 and 2018 Censuses.

Colombian Afro-descendants originated mostly from today's Congo, Angola, Gambia, Nigeria, Ghana, Ivory Coast, Guinea, Sierra Leone, Senegal, and Mali; and constitute the second largest Afro-descendant population in the region after Brazil. This geographical dynamic laid the foundations for the current ethnic population distribution in Colombia, in which a higher proportion of White-Mestizos Spanish descendants are located in Bogotá and other main cities in the centre of the country. Afro-descendants populated the Pacific and Atlantic coast, while Indigenous are based in the periphery, the regions of the Andes and the Amazon. The 2018 Census, after correcting for omissions[1] counted a total population of $48,258,494$ inhabitants in the country, from which $1,905,617$ are Indigenous belonging to 115 Indigenous communities, and $4,671,160$ were identified as Afro-descendants, according to self-reported ethnicity (Departamento-Administrativo-Nacional-De-Estadística, 2019a,b). Figure 1.2 presents the population structure by ethnicity and their characteristics in terms of age distribution.

[1]The Census 2018 had an omission of 4,094,077 persons according to official reports

Figure 1.2: Population Pyramid by Ethnic Group



Population structure of the different ethnic groups in Colombia, 2018.
Source: Own elaboration based on Colombian 2018 Census.

The Indigenous population is believed to have arrived in American territories more than 30,000 years ago, as a result of migrations from the last glacial period. In pre-Columbian America, Colombia was inhabited by various Caribe cultures. In the north of the country, there were cultures such as Tayrona and Zenú, while in the central region, there were cultures like Muiscas and Quimbayas. The western region was populated by Tumaco-Tolita and Chocó cultures, and the Amazon in the south was inhabited by numerous small indigenous communities (DNP, 1989; Jaramillo, 1964). According to the 2018 Census, 4.4% (approximately 1,905,170 individuals) of the Colombian population is Indigenous, representing more than 100 communities and speaking around 64 Amerindian languages. These figures position Colombia as the 7th country with the highest percentage of Indigenous population in the region, following Bolivia, Chile, Ecuador, Guatemala, Peru, and Panama. However, Indigenous leaders in Colombia argue that the proportion of Indigenous people in the country could be greatly underestimated.

Colombia is one of the most important economies in the region and has experienced a general mortality decrease in the last decades. Consequently, life expectancy in Colombia has increased more than 20 years since 1960, Figure 1.3 shows life expectancy trends by gender and year. However, life expectancy by ethnic groups has not been estimated due to data quality problems. On the other hand, the health and social conditions among ethnic groups in the region are significantly different from the general trend in the whole population (Hopenhayn et al., 2006). According to the Colombian National Statistics Office (DANE) the educational level and the social living conditions of the ethnic minorities are below the national average; 30.6% of the Afrosdescendants experience multidimensional poverty, and in rural areas, this figure increases to 50%. On the other hand, 45%

of indigenous population achieve only primary school levels of education or lower. This clearly highlights social heterogeneity, suggesting that population changes are happening at a different rhythm among different groups in the interior of the country. Indigenous and Afro-descendant populations present lower living conditions than White-Mestizos, and census data show that the regions where they are located are the poorest in the country.

Figure 1.3: Life Expectancy in Colombia from 1960 to 2019

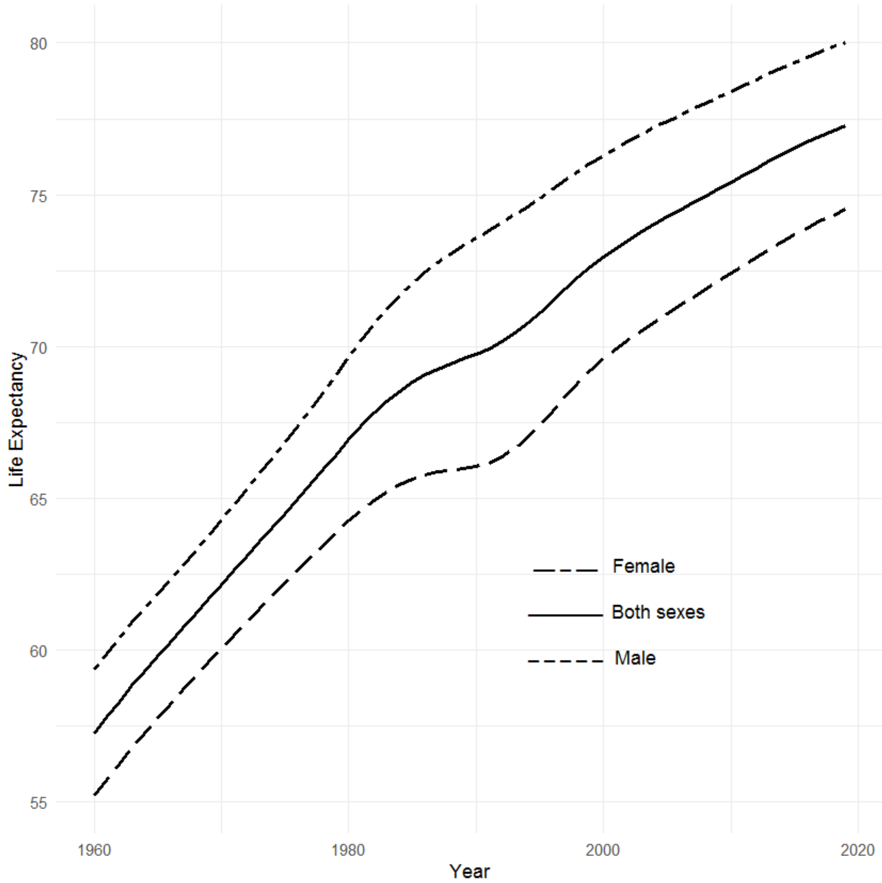


Figure 1.3 presents the life expectancy at national level which has grown more than 20 years since 1960.
Source: World Bank.

Clear evidence exists that demographic behaviour differs considerably across ethnic groups. The fertility rate in Colombia, for example, was 1.8 per woman in 2018 according to the World Bank. Although fertility of the individual ethnic groups is unknown, the child-woman ratio gives a general idea of the different patterns among ethnic groups. The child-woman ratio is defined as the number of children under 5 years, divided by the number of women at childbearing age, and was 28.6; 42.1 and 24.6 for Afro-descendants, Indigenous, and White-Mestizos respectively, see Table 1.1. In the same way, ethnic groups migrate at a lower rate than White-Mestizos groups, and their migration is mostly from rural to urban areas (Del Popolo et al., 2011). This has been driven mostly by the higher intensity in rural areas of Colombia's long-running internal conflict. According to the United Nations, the size of the displaced population in Colombia increases every year.

Table 1.1: Demographic Indicators by Ethnicity, Census 2005 and 2018.

Demograhic Indicator	Indigenous		Afro-descendant		White-Mestizo		Total pop.	
	2005	2018	2005	2018	2005	2018	2005	2008
Male %	50.5	49.9	49.7	48.8	48.9	48.7	49.0	48.8
Female %	49.5	50.1	50.3	51.2	51.1	51.3	51.0	51.2
Child-Woman ratio	62.2	42.1	42.2	28.6	34.8	24.6	36.8	25.6

Source: Departamento-Administrativo-Nacional-De-Estadística (2019a,b)

However, the real size of the displaced ethnic population is unknown because of the high level of under-registration (Soledad and Jiménez, 2011).

Indigenous and Afro-descendants territories are principally located in the Amazon Rainforest and along the Pacific Coast. These areas represent most of the Colombia's unexplored areas. In this context, forced displacement of these communities have been determined not only by reasons of the internal armed conflict, but in a considerable proportion too by megadevelopment projects such as African oil palm crops, as well as the existence of rich natural resources and the mining industry, and by the spread of illicit crops (Cárdenas, 2018; Oslender, 2007; Escobar, 2003). The High Commissioner for Refugees from United Nations (UNHCR) reported that between January and November 2020, more than $21,000$ displaced people from which 55% were Afro-descendants and 20% were Indigenous (UNHCR, 2020).

Colombia is a country with many contradictions in different social conditions. Ethnic groups have played a crucial role in the history and in the construction of a national identity; Colombia had presidents with Indigenous and Afro-descendant background in 1854 and 1861 respectively. But political leaders of ethnic minorities have been marginalised and erased from national history. The importance of this study thus consists precisely in starting to make visible the information that is missing in the national narrative. The analysis of ethnic morbidity and mortality represent an important step in this direction, in terms of understanding health conditions and the epidemiological patterns of Indigenous and Afro-descendants that have remained invisible in the aggregate national statistics. The present analysis will therefore yield better insights of ethnic differentials in mortality and morbidity over time in Colombia.

1.4 Research Questions and Objectives of the Study

This thesis aims to analyse the mortality difference and identify mortality patterns among Indigenous, Afro-descendants, and White-Mestizos in Colombia between 2008 and 2019. To achieve this aim, it will assess the quality of mortality data collected by census and registration sources as well as compare mortality data from the national vital statistics and the 2018 Census to investigate possible under-registration and how it affects the different ethnic groups.

This research will address the following four inter-related questions.

1. How do life expectancy and mortality rates vary among different ethnic groups by age and sex?

2. To what extent under-registration of deaths influences the estimation of life expectancy and mortality?

3. Which causes of death contribute most to the differences in life expectancy and mortality among different ethnic groups?

4. How can we apply machine learning techniques on mortality register data to better understand and improve ethnic classification in Colombia?

To address the research questions, estimating the ratio between registered deaths in official statistics and census-reported deaths is a central objective. This estimation will facilitate, first, the design of a methodology to improve data quality in mortality registers; second, a reliable estimation of mortality trends by ethnicity and the identification of the main causes of death in ethnic minority groups. The leading causes of death in these groups have remained underrepresented in national statistics due to the low proportion of ethnic registered deaths. Additionally, the study aims to apply machine learning techniques to identify the ethnicity of deceased individuals for the years in which this information is missing. Reconstructing ethnic information will enhance and complete the available data on ethnic mortality, enabling the analysis of mortality differentials for years prior to 2008.

Life expectancy and mortality rates could exhibit significant variation among different ethnic groups by age and sex as Indigenous and Afro-descendant populations generally experience lower life expectancy and higher mortality rates compared to White standard populations. These disparities are pronounced in specific age and sex categories, reflecting underlying socioeconomic inequalities, access to healthcare, and exposure to risk factors. For instance, young Afro-descendant men are disproportionately affected by violence-related deaths, while Indigenous populations may have higher mortality rates from preventable diseases and inadequate healthcare services. These variations highlight the intersection of ethnicity, age, and sex in shaping health outcomes in Colombia. Additionally, understanding how under-registration of deaths impacts the estimation of life expectancy and mortality rates, particularly among ethnic minority groups, will allow more accurate death counts to analyse health disadvantages in ethnic groups. This under-registration is often more prevalent in marginalised communities, such as Indigenous and Afro-descendant populations, due to factors like geographic isolation, socioeconomic barriers, and systemic neglect. Addressing under-registration is crucial for obtaining accurate demographic and health statistics, which are essential for effective policy-making and resource allocation aimed at reducing health disparities.

The analysis of the different causes of death will contribute to explain possible disparities in life expectancy and mortality among ethnic groups. For Indigenous populations, infectious diseases, malnutrition, and lack of access to medical care could be significant contributors. In Afro-descendant populations, violent deaths, chronic conditions related to socioeconomic disadvantages could be more prominent compared to other groups. These causes reflect broader social determinants of health, including poverty, limited access to education, and systemic discrimination.

The relationship of life expectancy, causes of death, ethnicity, and additional demographic information could represent a great potentical for the use of Machine learning techniques to enhance the accuracy and completeness of ethnic classification in Colombia. Techniques such as random forests, neural networks, and support vector machines could identify patterns and correlations between demographic variables and ethnic identity. This approach not only will fill gaps in historical data but also improves the reliability of current records. Enhanced ethnic classification enables more precise analysis of health disparities and supports the development of informed health policies aimed at addressing inequities.

This research is the first of its kind to systematically measure ethnic differentials in mortality in Colombia, and will thus address questions about the extent of inequalities in mortality among various population groups and across different geographical regions in Colombia; how the mortality gap has changed over time; and what causes of death are the largest contributors to the gap within and between different ethnic groups. The study will likewise address questions about what factors contribute more to potential under-registration of mortality among various ethnic groups, and how we can reconstruct mortality data by ethnicity prior to 2008.

Ethnic inequalities are not easily measurable: the gradual linear increase in life expectancy at the national level masks the heterogeneity in mortality levels among different population subgroups and across different geographic regions. Based on evidence from the existing literature (Chapter 2) and informed by the research gaps identified from the literature review, this study will test two interrelated research hypotheses. First, we believe that mortality data from registration sources are likely to be underestimated in areas with high concentration of Indigenous and Afro-descendant communities. Second, Indigenous and Afro-descendant communities experience higher mortality rates and lower life expectancy, and they display distinct patterns across age, sex, and cause of death, particularly with respect to child mortality and infection diseases.

1.5 Structure of the Thesis

The thesis is organised as follows: Chapter 2 defines the concepts and meaning of the categories skin color, race, and ethnicity for the present study, and the importance of the correct understanding of those categories in population health research. It further outlines the different theories and models for the analysis of health inequalities in ethnic populations. Chapter 3 presents the data employed in the project and describes the variables and data structure of the Census 2018 and the Colombian mortality records and how information about ethnicity is collected.

Chapter 4 analyses how mortality under-registration affects the different ethnic groups and what regions in the country present the highest level of mortality underestimation. Further, it estimates the probability of a deceased person of ethnic origin being omitted from vital statistics and how the ethnic background and sociodemographic characteristics make one more prone to being excluded from registration. In Chapter 5, a correction method for mortality under-registration for the period $2008 - 2019$ is presented, along with estimates of the ethnic mortality gap and life expectancy with corrected and uncorrected data. The method improves mortality information by considering age at death, gender, and ethnic bias.

Chapter 6 analyses all causes of death by ethnicity and their contribution to life expectancy and life disparity among ethnic groups. This chapter also analyses how the proportion of deaths by causes has changed over time. Chapter 7 proposes the use of machine learning methods to identify the ethnicity of deceased people. The method can be employed to construct the ethnic variable for years in which this information was not collected. Finally, Chapter 8 summarises the main findings, key contributions, and limitations of the study.

CHAPTER 2

Ethnicity and Mortality in Colombia and Latin America

2.1 Understanding Skin Color, Race, and Ethnicity

There is much confusion about the category race/ethnicity in epidemiological research. The use of ethnic ascription has not been rigorous in most studies due to inappropriate methodologies to define ethnicity (Moubarac, 2013; Gravlee, 2009; Bhopal, 2004; Bhopal and Donaldson, 1998). One study suggests that about 64% of the studies are using race and ethnicity incorrectly, and that this misunderstanding has often led to wrong classification or identification of ethnic communities (Moubarac, 2013). Categories such as skin color, race, and ethnicity are used often interchangeably in population health studies. However, there are important differences in these categories and how they are defined, measured, and interpreted in addressing health inequalities across different ethnic groups. In order to examine mortality by ethnic group, we must first understand, and contextualise the ethnic categories. As race and ethnicity are social constructed concepts, their understanding could change from one society to another. It is therefore fundamental to define and specify what we mean by ethnicity.

Skin color as a biological category is of little interest in social sciences for several reasons. First, the use of White and Black as social identifiers is quite vague and they are broad categories that amalgamate groups of people that differ socially and culturally (Moubarac, 2013; Bhopal and Donaldson, 1998). Additionally, it hides the remarkable heterogeneity within the same category. The Office of Management and Budget (OMB) Standards from the US defines White as a person having origins in any of the original populations of Europe, the Middle East, or North Africa (Office of Management and Budget, 1997). In England and Wales, according to the Census 2011, the category White includes White-British, White-Irish, White-Gypsy/Traveler, and White-other, the latter made up mainly by the Polish and other European populations. These groups, however, have shown significant differences in terms of social and health conditions (Office for National Statistics, 2021, 2020). The use of the category Black based on the skin color presents a similar problem. In Colombia, for instance, groups of Black-Maroon, Black-Caribbeans, and Blacks from the Pacific Coast coexist with clear cultural and linguistic differences (Wade, 2012; Jordan, 2004). Additionally, there is significant number of Afro-descendant people living in the main cities of the country that do not identify themselves as part of those groups. Therefore a Black category based on the skin color will transform social heterogeneity into a biological category.

Race on the other hand, considers skin colour, but it is more complex than a simple skin colour, more controversial and perhaps a more blurred concept due to the fact that it considers phenotypical and cultural elements in the perception of race. Two main ideas have been proposed in the literature to better understand the concept of race. The first is the biological paradigm based on the traditional North American approach of race as a division of humankind through natural selection (Smedley, 2007; Keita and Kittles, 1997). The biological idea of race has been quite successful due to the sociocultural reality of race and racism in North America that presents social inequalities as a biological inferiority, while at the same time epidemiological evidence for racial inequalities reinforce public understanding of race as a result of genetic superiority and perpetuate a racialised view of human biology (**?**Powe et al. 2013; Gravlee 2009, p.48; Gonzáles et al. 2003).

Bamshad and Olson (2003) attempted to answer the question whether race does exist in biology. They concluded that genetic variations among human beings do not depend on skin color or facial features, but they follow a pattern which varies across space. This means that groups with similar physical characteristics can be quite different genetically if they are in a different geographical space. For instance, Afro-descendants from the Caribbean island, North America, and West Africa present variations in terms of prevalence to specific diseases (Cooper et al., 1997) which represent differences in terms of biological adaptation. The concept of race based on a biological differentiation has been systematically proved to not have any scientific foundation. Skin color represents a non-concordant genetic variation in the human body, which suggests that there is no relationship between the tone of the skin and a better or worse biological constitution (Jorde and Wooding, 2004; Goodman, 2000; Ehrlich and Holm, 1964; Wilson and Brown, 1953). Therefore, it has no value for predicting other aspects of biology as those characteristics are independently distributed. A racialisation based on skin color is thus an arbitrary criterion, similarly as it would have been a classification based on the colour of eyes and hair, or short and tall people; but colonial encounters privileged more evident phenotypical differences such as skin color, principally in the context of unequal power relations (Wade, 1993a, p.4). The critique to the racialisation of the human biology should not be understood as a denial of genetic human diversity that has been proved to be determined by geographic distance (Li et al., 2008; Manica et al., 2005). Rather, the central argument is that the race as a concept is inadequate for describing the complex structure of human genetic variation (Gravlee, 2009, p.50).

The second perspective is the idea of race as a social construction in which there is a general consensus in the social sciences. Race as a social construction should be understood as an idea that evolved through social interaction, and that has been internalised as an objective reality (Berger and Luckmann, 1966) In other words, the social construction of race implies that people acquires the knowledge to understand that races can be classified by certain particularities in the type of hair and skin colour, or cultural traits, but not by anthropometric measures such as short and tall people. This construction is humanly produced and it has therefore an artificial burden that can change from one society to another, as the perception of the determinants of a particular race can be more or less flexible depending on the society. In North America and South America for example, blackness and whiteness are perceived differently, and the same person could be ascribed to different racial groups depending on the country they are located.

Race, therefore, can vary as a social construction, but skin colour as a biological category does not. Skin colour, however, is an important factor in the construction of the idea of race, but it is not a *sine qua non* of race. That is, the idea of race goes beyond skin colours and can be understood even in a colour-blind context. For instance, the voice and the way of speaking in a phone conversation can be truly indicators of race, skin colour is therefore not indispensable, and it can even be in contradiction with the colour we expect for a particular race if a person has Afro-descendant features but their skin colour is not what is considered black. Skin colour could be sometimes ambiguous to determine race, as racialisation implies a characterization of the other in which not only the skin colour is considered but also characteristics of the body and culture are used as racial markers (Groothuis, 2020; Garner, 2017; Fanon, 1961). Racialisation, in this sense, works as a kind of fetishism that allow to perceive race in social practices and objects which do not have "skin colour". Blackness, or Indigenousness can be seen even through racial markers when phenotypical information is not available or even when that information is in contradiction with what we expect for a specific race.

Race differs from ethnicity, which refers to the *ethnos*: the collectivity and how we are related with it (Østergard, 1992). It should be understood as a collective identity with putative common ancestry that shares cultural symbols and practices like language, religion, and norms, and characteristics such as type of diet, living style, and kinship (Cornell and Hartmann, 2007). In this sense, ethnicity is a self-identity, and believed to be part of belonging to a group while race refers to how the person is identified by others. In other words, race is the result of colonial encounters, whereas ethnicity is connected with the formation of the state-nation and the idea of national identity (Bonilla-Silva, 1999).

Race and ethnicity can be therefore opposite categories if a person is ascribed to a group to which they do not have an ethnic identity and sense of belonging. One does not have to be or self-identify as part of a racial group in order to be classified as such (Kaufman, 2018), and vice-versa it is also true that a person does not need to look like a black to report an Afro-descendant ethnicity. Ethnic-identity changes can be seen as a problem by researchers that conceive ethnicity as a biological immutable characteristic; however, as a social constructed variable, ethnic identification is susceptible to variations over time. An indigenous person, for instance, who migrates to the city, has family with a non-indigenous person and consequently changes lifestyle could eventually identify themselves as non-indigenous as they could develop feelings of belonging to a different ethnic group, socioeconomic status, educational level, or place of residence that could affect ethnic identities. Likewise, in multi-ethnic families, people can preserve language, religion, living styles and identities even when they do not share anymore the general physical traits of their ancestors or the group they belong to.

The concepts of race, ethnicity, and skin colour refer to distinct categories and must be understood within a broader sociocultural context. Although race and ethnicity may include skin colour, they are primarily socially constructed concepts based on culture, beliefs, and language, among other physical characteristics. This social construction allows for racial and ethnic identities to exist without common biological traits among their members. In contrast, skin colour alone is a biologically determined category that can be imprecise as an identifier when considered in isolation. It functions as a racial marker in combination with different factors. For example, an Indigenous person, an Afro-descendant, and a White-Mestizo could have the same skin colour, but their identities become clearer when additional physical traits, culture, language, and other factors are considered. In Colombia, ethnicity is preferred to determine cultural identities and it is generally self-reported, but for deceased people, the ethnic identity is stated by their family. Since 2008 the availability of ethnic mortality data and the improvement in the quality of the information have allowed a better understanding of ethnic health inequalities in the country.

2.2 Racism as the Root Cause of Health and Social Inequities

The definition of race, as previously mentioned, remain contested, resulting in a blurred understanding of racism. Nevertheless, Racism is a concept that has been defined in various ways, depending on the context of analysis. Some authors has defined it as a systematic forms of discrimination in which non-White groups has been excluded from the access to power and resources (Jones, 2000; de Benoist, 1999; Hall, 2018), while others view racism and racialisation processes as mostly an ideology and psychological process, a type of "madness" or irrational behaviour; the Frankfurt School understood racism as a syndrome characterised by frustration where prejudice and stereotype play significant roles (Adorno et al., 1950, p.765). Still, others see it as encompassing a variety of discriminatory forms at institutional, cultural, and individual levels (Jones, 2000; Memmi et al., 2000; Krieger, 1999), and more recently, the problem of racism has been analysed as an independent *racialised social system* (Bonilla-Silva, 1997). In this document, Racism is conceptualised as a complex social system rooted in notions of racial superiority, which operates through various mechanisms to produce social inequities. This occurs across multiple levels, including individual, institutional, and cultural domains, and through diverse means such as cultural norms, language, social conventions, and stereotypes. See Figure 2.1 for a representation of the system of racism.

Racism, in this sense, is a social phenomenon and represents the foundation for a particular form of social organisation. It relies on three key elements -*assumptions*- that legitimise a specific racial order: First, the notion of a superior "race" is established. Second, this superiority is determined by biological or bio-anthropological factors. Third, this supposed "superior biological constitution" is reflected in social inequalities, often perpetuating a form of social Darwinism. (see de Benoist (1999, p.14) and Memmi et al. (2000, p.20) for additional assumptions).

It is not an exaggeration, therefore, to assert that racialised social systems have been a fundamental form of social organisation in modern societies. Race is the major mode of social differentiation that cuts across and takes priority over social class, education, occupation, gender, age, religion, culture, and other differences (Smedley, 2007, p.19). Understanding racism as a social fact enables us to distinguish its ideological aspects from its factual components and comprehend the mechanisms through which an ideology of racial superiority translates into social inequities. This perspective also clarifies why a racialised order may appear as a "natural" organisation in which racist ideologies appear disconnected from a decision-making contexts which is assumed impartial, leading to the emergence of social Darwinism as a consequence of presumed innate advantages. The proposed definition of racism makes it clear why societies present high levels of racism in the way of racial inequities, and at the same time it is not easy to identify the perpetrators, we are in a society of racism without racists to the extent that no one see themselves as a racist [Bonilla-Silva (2006), Memmi et al. (2000, p.19)]. This paradox arises from the fact that when we are in a system of racism, perpetrators are no longer required (Jones, 2000; Bonilla-Silva, 1997).

Racism and colonialism are very closely related in the sense that colonialism has been presented by racist ideologies as the means by which inferior races, presumed to be in an "arrested state of development" and incapable of self-governance, were purportedly assisted by a superior race to attain higher levels of civilization (de Benoist, 1999, p.18). This interpretation suggests that racism can be perceived not solely as a manifestation of hostility, but also as a form of "altruism" from a paternalism perspective. Paternalistic racism is conceived as the belief that individuals from certain racial groups require the intervention and guidance of a knowledgeable -typically White- person. This belief stems from the assumption that individuals who are uninformed or infantilised -typically non-White- individuals, cannot be trusted to make decisions in their own best interest (Baker (2015), VanDeVeer (2014, p.22)). Albert Schweitzer articulated this idea as follows:

> "The Negro is a child, and with children nothing can be done without the use of authority. We must, therefore, so arrange the circumstances of daily life that my natural authority can find expression. With regard to the Negroes, then, I have coined the formula: "I am your brother, it is true, but your elder brother." (Schweitzer, 1947, p.185).

Figure 2.1: Social System of Racism

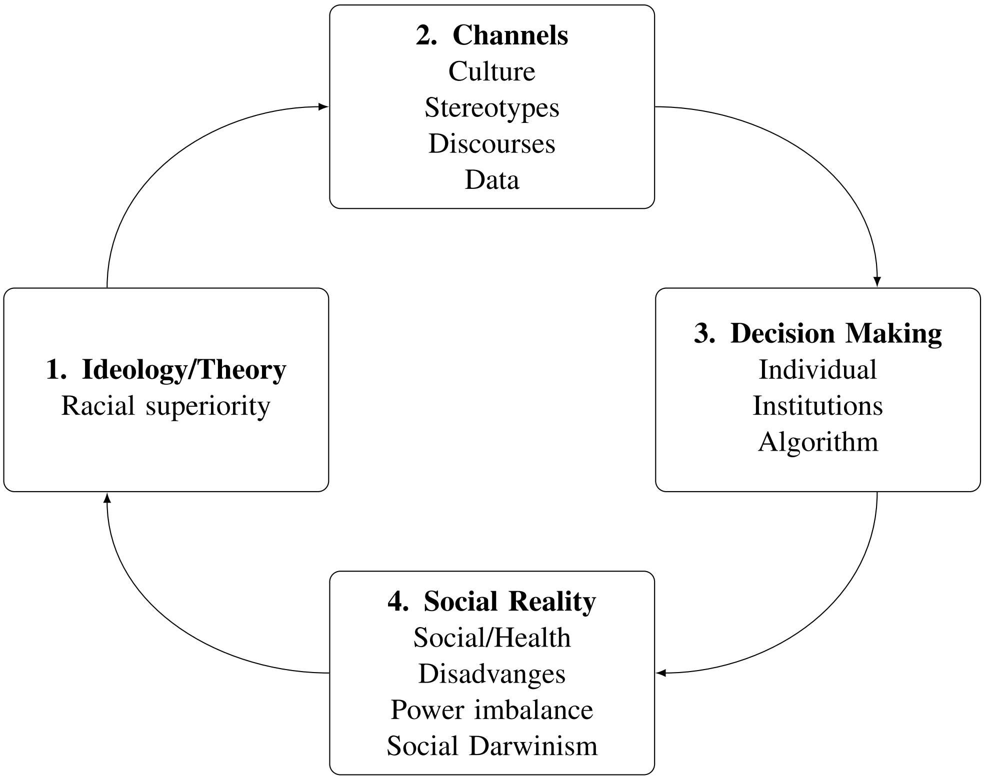


Figure 2.1 depicts the mechanism through which racism functions as a social system. The concept of inferior/superior races is disseminated through various channels, legitimising a racial hierarchy by promoting a dichotomy of social standards based on notions of "good" and "bad" ideals in terms of religion, knowledge, beauty, language, art, among others. This information racially filtered is then utilised by institutions, individuals, and algorithms to make "the best" distribution of resources and selection of candidates: the choice of the chosen ones. This process operates across all levels of society (parthnership, education, labor market, politics, justice, health, residential space, etc.). It is presented as a "natural" selection mechanism, wherein individuals consistently belonging to the same racial groups emerge as either winners or losers. Consequently, this lead to a systematic and consistent racial inequalities in the social reality which is perceived as "proof" of racial superiority, thereby perpetuating and reinforcing the racial ideology in a cyclical manner. Source: Own elaboration based on literature described in Section 2.2.

Racism can manifest in various ways, impacting Afro-descendants and Indigenous populations across different contexts. Section 2.6 discusses the main perspectives proposed to explain racial/ethnic health inequities. While racial health disparities can arise from diverse pathways, it is essential to acknowledge that all these factors stem from a social system of racism and ideologies of racial superiority. In essence, racism acts as the cause of the causes of health and social inequities, even when health disadvantages may seem attributed to cultural practices or geographical factors; it is evident that central governments show little interest in implementing measures to mitigate cultural and environmental adversities. In 2012, during discussions about the Plan of Social Investment for Antioquia, one of Colombia's most significant regions, a White-Mestizo member of the legislative assembly vehemently opposed social investments in Afro-descendant regions. They expressed that "investing money in Afro-descendant regions is like putting perfume on shit"[1], suggesting that racial inequities are "natural" and inevitable. This statement underscores how racism is a fundamental factor underlying health and social inequities across all contexts in Colombian society.

2.3 Mortality Transition in Latin America

The mortality transition, understood as the process whereby human societies have moved from scenarios of high mortality to very low mortality, has been somewhat a standard process in human populations; however, not all societies have experienced the changes at the same speed. According to the theory of epidemiological transition put forward by Omran (1971), the main attribute of the mortality transition is related to changes in the causes of death in which there is a characteristic shift from high prevalence of infectious diseases towards high prevalence of degenerative and man-made diseases. From Omran's perspective, there are three clear stages in the transition through which societies have to pass: first, the "age of pestilence and famine", during which the level of mortality is high and fluctuating, life expectancy ranges between 20 and 40 years, and population growth is stagnant; second, the "age of receding pandemics", during which mortality declines, epidemics and infections becomes less frequent, and life expectancy improves considerably to around 50 years; and third, the "age of degenerative and man-made diseases", in this stage, mortality continues decreasing with an stable trend, life expectancy continues increasing, and fertility becomes a crucial factor in population growth. The changes and the transitions in itself, will depend on improvements in socioeconomic conditions, improvements in public health, and in medical science.

Changes in human populations over time have also been analysed from the perspective of the "Health Transition" (Frenk et al., 1991a). This differs from the Epidemiological Transition Theory, in the sense that the idea of a Health Transition has broader implications than the epidemiological transition, and the characteristics of the observed changes are conceived as a dynamic process in which health conditions are continuously being transformed. In other words, the Health Transition

[1]Here is a brief report of assembly session: https://www.elespectador.com/colombia/mas-regiones/la-plata-que-uno-le-mete-al-choco-es-como-meterle-perfume-a-un-bollo-article-344843/

perspective considers that diseases disappear, appear or re-emerge, rather than a simple unidirectional transition over time beginning with infectious diseases and ending with non-communicable and chronic diseases dominating the causes of death. The Health Transitions Theory considers a context much more complex, in which the social environment plays a fundamental role via economic structure, political institutions, science and technology, and culture and ideology. Additionally, the genetic constitution, living conditions, and the individual behaviour constitute a multicausal system that will determine health status as a final result of the balance between exposure to disease agents and individual susceptibility.

The formulation of a more general demographic transition composed of three stages had been proposed earlier in order to build a theoretical basis for the understanding of demographic changes (Notestein, 1945; Landry, 1934; Warren, 1929). However, beyond the discussion about the determinants of mortality change, there are three stages of the mortality decline that can be distinguished (Kirk, 1996). First, the appearance of the modern state and the monopoly of force allowed, on one side, investments in agriculture and infrastructure that improved the production and distribution of food, quality of water, hygiene, and nutrition; and on the other hand, tribal wars disappeared and deaths from local conflicts were reduced significantly. Second, the discoveries of Pasteur and Koch led to the treatment of diseases such as diarrhea and tuberculosis, which increased child survival significantly. Likewise, progress in medicine and health education during the inter-war period had a positive effect on survival chances. And third, the discovery of the penicillin and the use of antibiotics represented a dramatic reduction in epidemic and contagious diseases. Since the second world war and particularly in the last few decades, life expectancy has continued improving, and a fourth stage in the mortality transition have been identified although there is not a consensus in the interpretation of the change in mortality patterns, and the new fourth stage has been called in different forms: "Cardiovascular revolution" (Vallin and Meslé, 2004), "The age of delayed degenerative diseases" (Olshanksy and Ault, 1986), "The Hybrisitic Stage" (Rogers and Hackenberg, 1987). Regardless of the different definitions of a possibly new stage in the mortality transition theory, these studies confirm that the demographic transition is an unfinished process and that even a fifth stage could be determined by the re-emergence of infectious and parasitic diseases in the form of outbreak of epidemics such as the Ebola virus and more recently COVID-19 pandemic (Vilella and Trilla, 2021; Olshansky et al., 1998).

The mortality transition in Latin America has been an heterogeneous process with clear differences among countries, but also within countries among social and ethnic groups, where there are important differences. The co-existence of patterns from different stages of the demographic transition in the same society, have been called "long-term polarization model" [Modelo de polarización prolongada], (Frenk et al., 1989, p.31); "structural heterogeneity" [heterogeneidad estructural], (Possas, 1991); and "regional heterogeneity" [heterogeneidad regional], (Di Cesare, 2010, p.7). While in countries such as Peru and Bolivia infectious diseases are still among the most important causes of death, in Argentina and Uruguay the incidence of these causes is minimal. On the other hand, countries like Colombia and Venezuela present excessive mortality by external causes, mainly homicide (Di Cesare, 2010).

Table 2.1: Economic Modernization, Fertility Transition, and Mortality Profile for some Latin American Countries

Economic modernization	Fertility transition	Profil of mortality
I ADVANCED	II VERY ADVANCED	I ADVANCED
Argentina	Argentina	Argentina
Uruguay	Uruguay	Uruguay
Cuba	Cuba	Cuba
Chile	Chile	Costa Rica
Costa Rica		
Venezuela		
II PARTIAL AND RAPID	II ADVANCED	II MIXED
Panamá	Costa Rica	Venezuela
Brasil	Venezuela	Brail
Colombia	Brasil	Colombia
México	Colombia	México
Rep. Dominicana	México	Rep. Dominicana
Ecuador	Rep. Dominicana	Ecuador
Perú	Ecuador	
	Perú	
III INCIPIENT	II EARLY	III INCIPIENT
El Salvador	El Salvador	Perú
Guatemala	Guatemala	El Salvador

Source: (Frenk et al., 1991b, p.97)

A classic work by Frenk et al. (1991b) classified the Latin American countries in different stages of the demographic transition according to its economic development, fertility, and mortality indicators. The economic indicators consider aspects of urbanization, education, and the characteristics of the national economy in terms of the major activities. The fertility indicators used include four categories: $TFR < 3$; $3 < TFR < 4.5$; $4.5 < TFR < 5.5$; and $TFR > 5.5$. For the mortality analysis, it used a ratio of the causes of death, in which the mortality rate of infection diseases was divided by the mortality rate of non-communicable diseases. The analysis found that countries such as El Salvador are in an incipient stage across all compared categories, whereas countries like Argentina, Uruguay, and Cuba exhibit more advanced socioeconomic indicators and epidemiological transitions, as shown in Table 2.1. However, the study highlights significant disparities between rural and urban areas within the same countries. For instance, Venezuela was classified as economically advanced within the region, yet its mortality indicators displayed a mixed pattern characteristic of both incipient and advanced transitions. El Salvador and Guatemala lag behind in terms of the transition. Guatemala historically presented a higher level of mortality than the rest of the countries in the region; 61% of its deaths were caused by communicable diseases in the 90's, compared to 22% in Mexico, 13% in Chile, and 7% in Uruguay (Albala et al., 1997).

The eradication of inequalities in the region is unlikely to occur in the near future and will require a considerable amount of time. Life expectancy at birth in Costa Rica, Chile, and Cuba was approximately 78 years in 2007. In contrast, Haiti, which was not included in the analysis by Frenk et al. (1991b), had a life expectancy of less than 60 years. Given that the average increase in life expectancy at birth (e^0) in the region is 0.4 years per year, Haiti lags behind Costa Rica by as much as 50 years (ECLAC 2010, p.38; ECLAC 2007, p.28). Significant disparities persist in the Latin American region and are anticipated to endure over the coming decades. Addressing this issue will necessitate structural changes in the health system. While the region has made notable progress in reducing mortality, and some countries exhibit indicators comparable to those in developed regions, a substantial portion of the population in these same countries continues to experience unfavorable health conditions.

2.4 Mortality Transition in Colombia

Demographic changes in Colombia began in the 1950's when fertility and mortality started to show a decreasing trend (Banguero and Castellar, 1993; Florez, 1987). The urbanisation process and the availability of contraceptive methods were among the main determinants of changes in the population structure (Miller, 2005; Florez et al., 2000; Flórez, 1995). The rural population in the country, for instance, was 61.4% in 1951, and in 1993 it had decreased to around 25%. This led to a higher participation of woman in the educational system, in the labor market, and in general, it led to a more emancipated role of women in society and consequently to a significant fertility decrease (Flórez, 1995). The population changes associated with the demographic transition occurred in Colombia with unusually accelerated speed, and it is not totally clear the role of the internal conflict on the demographic transition, particularly on the role of young woman and families that lost their male members. Such a situation could push women out of the house towards more active roles, thus also accelerating the urbanisation process in rural areas.

In the 1950s, Colombia experienced a period known as *La Violencia* [The Violence] (1946–1958), one of the most intense periods of internal conflict in Colombian modern history (Bailey, 1967; Guzmán et al., 1962). Despite this turmoil, mortality began to decrease during the same period, and at a faster rate than the decline in fertility. By the end of the 20th century, Colombia had shifted from a situation where the main causes of death were infections and parasitic diseases (Banco-De-La-República, 1951), to one where cancer, cardiovascular, and degenerative diseases became the major causes of mortality. The effects of *La Violencia* significantly altered the main components of mortality causes, and in the 1980s, with the rise of paramilitary groups, drug cartels, and the intensification of the internal conflict against guerrilla groups (Palacios and Safford, 2002; Florez et al., 2000), homicide emerged as one of the leading causes of death, as illustrated in Figures 2.2 and 2.4. "This social and epidemiological situation slowed the country's mortality transition from communicable to chronic-degenerative diseases in the country" (ECLAC, 2010, p.63).

During the first half of the past century, almost 50% of the total deaths in Colombia were children (Florez et al. (2000), see also Figure 2.3). The introduction of an extensive vaccination programme

Figure 2.2: Main Causes of Death in Colombia Between 1946 and 2001 (rate * 100,000)

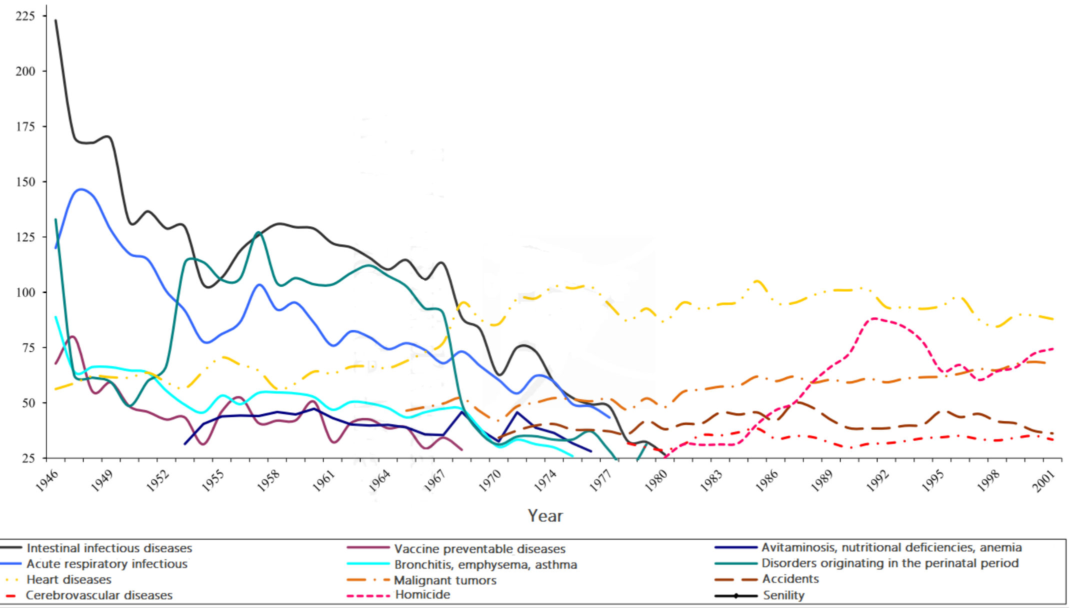


Figure 2.2 shows changes of the causes of death in Colombia between 1946 and 2001. Infectious and nutritional diseases decrease significatively while homicides, cancer, and cerebrovascular diseases increased from 1980's onwards.
Source: (Jiménez-Peña, 2014, p508).

across the country had therefore a considerable impact on mortality decline. Improvements in the diet as result of a higher purchasing power had a similar effect, as these allowed families to introduce a higher proportion of proteins in children's nutrition. For instance, in 1953, dairy products, meat and eggs represented around 40% of the average income of a Colombian family, while in 1998 the same products constituted only 11% of the income (Kalmanovitz and López, 2007, p.164). The urbanisation process likewise played a fundamental role in mortality decline as it allowed better housing and hygiene conditions; and in particular, the population migration from rural to urban areas extended the coverage of public health campaigns and allowed more effective measures to reduce malaria related deaths (Sarmiento, 2000; Florez et al., 2000; Flórez, 1995). Crude death rates in Colombia shifted from 16.68 total deaths per $1,000$ inhabitants during the period $1950-1955$ to 6.31 during $1980-1985$ and from 123 infant deaths per $1,000$ live births to 41 for the same period (ECLAC, 1990, p.9-10).

Figure 2.3: Proportion of Deaths by Age Groups in Colombia Between 1946 and 2001. All Causes of Death

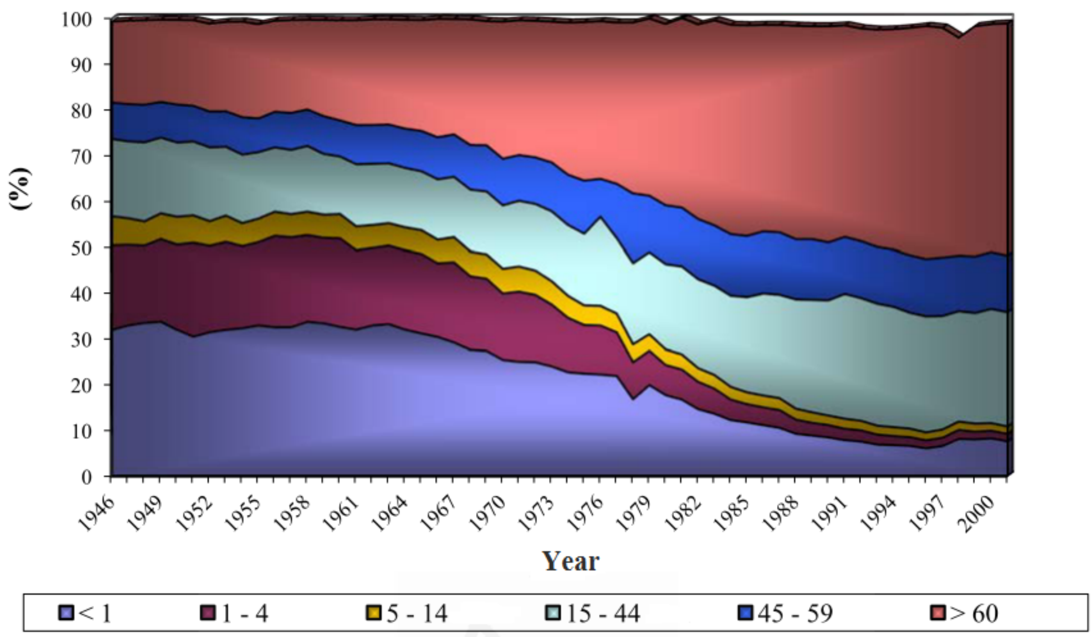
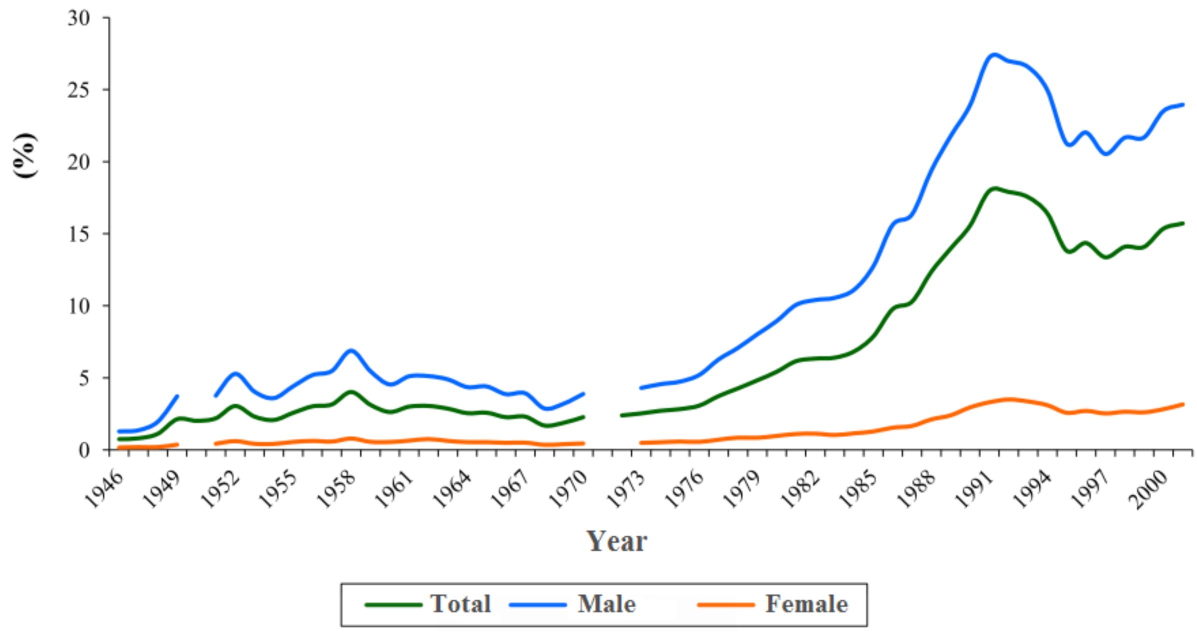


Figure 2.3 presents the changes in the age at death for the Colombian population during the second half of the last century. Child survival started improving from the 60's on while the number of people dying at older ages increased significantly. During the period, deaths from people under age 5 dropped from 50% to around 10% of the total deaths in the country.
Source: (Jiménez-Peña, 2014, p494).

The real effect of the Colombian internal conflict and political violence on the population changes are still under debate. However, the accelerated migration and displacement of ethnic population from rural to urban areas were strongly influenced by the rural violence (Gómez, 2014; Oslender, 2007; Escobar, 2003; Florez et al., 2000). The urbanisation process contributed to reducing mortality, as in cities people could access health care and better public services such as clean water. Additionally, the escalation of urban violence in the main Colombian cities played a fundamental role in terms of shaping the mortality patterns by cause, gender, and age. Gangs and militias groups existed in urban areas since the 60's, but it was only in the 80's when they developed a high degree of specialisation at the level of drug-trafficking gangs (Melguizo and Cronshaw, 2001). Homicide emerged as one of the main causes of death and increased considerably by the end of the 1970's of the last century, as shown in Figure 2.4. At the same time, mortality between ages $15 - 44$ shows an noticeable increase (Figure 2.3), presumably as a consequence of the violence.

The demographic transition, among other factors, has been understood as concomitant to improvements in the general social conditions of the population and the modernization of the health care system. However, in developing countries it does not always follow the same pathway. Signs of a demography transition are not necessarily related with social improvements, and exogenous factors such as contraceptive pills and vaccines can induce the effects of demographic transition keeping the general conditions more or less in the same stage. Poor countries that were relatively untouched by development showed clear signs of rapid population changes, and fertility declined as many developing economies stagnated or lost ground in the 1980s (Robey et al., 1993). In Colombia, it could have been the case in some rural areas; however, in general terms, there was an important improvement in the social conditions and hygiene that help to prevent infections diseases in particular (Sarmiento, 2000; Florez et al., 2000). Additionally, the role of the Pan American Health Organisation (PAHO) was crucial in terms of guidelines and implementation of health policies that helped the country to reduce mortality (Hernández et al., 2002).

Figure 2.4: Percentage of Deaths by Homicide and Injuries of the All Deaths in Colombia Between 1946 and 2001



Contribution of deaths by homicide during the mortality transition in Colombia.
Source: (Jiménez-Peña, 2014, p524).

Colombia has experienced important changes in terms of the mortality transition. The country is ongoing rapid improvements in sanitation and living standards, with medical and public health measures contributing significantly, and at the same time, causes of death in the population such as chronic degenerative diseases, heart disease, and cancer become more frequent, especially in urban areas. These demographic changes, however, are not taking place uniformly across the country. That is, there is little evidence of territorial convergence, particular ly in terms of causes of death. Regions in Colombia like Chocó, a region mainly inhabited by Afro-descendants in the pacific zone, and La Guajira, and Nariño, two departments with a high proportion of Indigenous population, show a high levels of deaths by infectious diseases. In contrast, Bogotá and other cities in the center of the county present a higher prevalence of cancer and degenerative diseases (Spijker et al., 2020; Otero, 2013).

Colombian's mortality transition and causes of death, although with some particular characteristics related to urban violence, show the general patterns of the Latin American structural heterogeneity and epidemiological polarisation (Di Cesare, 2010; Possas, 1991; Frenk et al., 1989). This means different regions or different social groups (such as the White-Mestizos, Indigenous and Afro-descendants groups), reside in the same country but are at different stages of the mortality and epidemiological transition.

2.5 Ethnic Patterns of Mortality

Improvements in life expectancy and mortality reduction have been a constant trend in the world population, as clearly illustrated in the global demographic transition. The "amount of gain", however, has been varied across regions and social groups. Health improvements have not been fairly "distributed" (Nazroo et al., 2006; Hummer et al., 2004; Williams and Collins, 2001). Ethnic groups, for example, continue to lag behind in terms of health outcomes. For instance, the ethnic population is over-represented in COVID-19 related mortality and critically ill COVID-19 patients, due to the higher risk of exposure especially those in the lower-qualified occupations (Aldridge et al., 2020; Gross et al., 2020). This highlights again the historical condition of social disadvantage and health inequalities in ethnic minorities.

Social and health disadvantages have been well-documented in different ethnic groups across the world. A large part of the differences are reflected in the ethnic domain, as a mechanism through which social status and social recognition are granted (Telles et al., 2015; Urrea-Giraldo et al., 2014; Wade, 2010a), principally in societies with a colonial past. Race and ethnicity have therefore acted as a social "filter" to determine the access to social rights and prestige. For example, colonial laws such as the "Compiled Laws of the Indian Kingdoms"(1680) prohibited Black and Indigenous individuals from holding administrative positions in colonial territories, and the Catholic Church forbade interracial marriages between Blacks and Whites (Urrea-Giraldo et al., 2014). Consequently, social exclusion of Indigenous and Afro-descendant populations has been a persistent practice that continues to this day (Wade, 2010b). In Colombia, Blacks have lower incomes compared to Whites; people from ethnic minority groups earn between 30% and 66% of what a White person earns (Flórez et al., 2003). Furthermore, according to four racial categories (Black, Mestizo, Mulato, and White), the darker the individual's skin color, the greater the income disparity compared to their White peers (Viáfara, 2017). A similar pattern was found regarding education (Telles et al., 2015). Likewise, many Indigenous people live in extremely poor conditions in rural areas, under a situation of structural discrimination, especially in health and education where considerable differences exist. (Del-Popolo and Oyarce, 2005; Psacharopoulos and Patrinos, 1994).

A historical analysis of mortality patterns in Colombian ethnic population is not possible, as ethnic information in mortality registers were collected only from 2008 onward. However, recent studies have shown a mortality gap between ethnic and non-ethnic population. The odds of child mortality

Figure 2.5: Distribution of the Afro-descendant Population in Colombian Regions

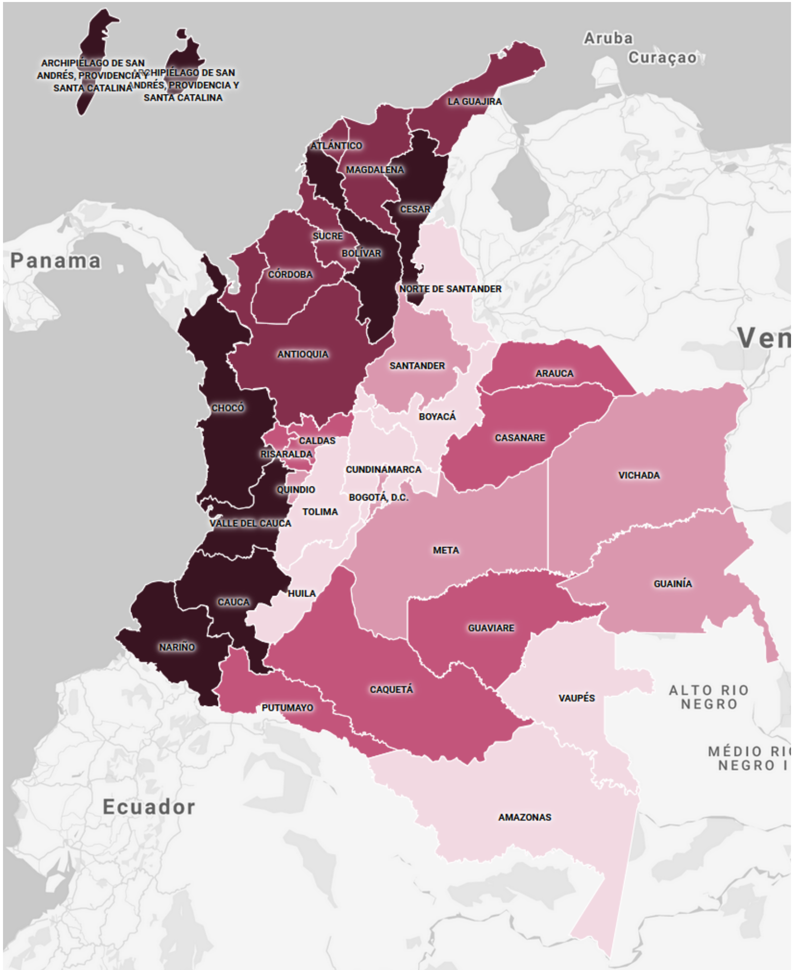


Figure 2.5 shows the regions where Afro-descendants are the largest percentage of the population. Historically, the Pacific and Atlantic coast have been the geographical areas where the most of Afro-descendants are located. The cities of Cali in the department of Valle del Cauca, and Cartagena in the departmento of Bolivar constitute the two cities in Colombia with the highest number of people with African background. Darker colors on the map represent a higher concentration of Afro-descendants.
Source: Own elaboration based on 2018 Census data.

for instance, are 25% higher in Colombian Afro-descendant women than White-Mestizo, and this gap has remained stagnant across different cohorts (Palacios, 2018). These findings are attributed to socioeconomic and educational differences between groups, but also because ethnic minorities in Colombia are less likely to have access to health care than other population groups (Bernal and Cárdenas, 2007). Deaths in ethnic population are usually concentrated in young ages and most of them die due to external causes; about 40% of total deaths of Afro-descendants before age 14 are due to homicide, and more than 50% of total males die before age 45 (Urrea-Giraldo et al., 2015). Self-rated health has proven to be worse in Indigenous and Afro-descendant people. This indicator, although is a subjective measure, has proved to be a reliable proxy of objective health and quality of life, and to be closely related to socioeconomic conditions (Maniscalco et al., 2020; Wu et al., 2013; Yamada and Takahashi, 2012). Indigenous and Afro-descendant people have less chances of having a good self-perceived health than White-Mestizo groups, and the perception could be worse at older ages as a result of a cumulative effect over time, marginalisation and geographical segregation (Mendez et al., 2020; Agudelo-Suárez et al., 2016).

The use of geographical regions as proxy of ethnicity has been a common methodological alternative to overcome problems of ethnic data scarcity in Colombian comparative ethnic studies. The Colombian geography has a had an important association with ethnicity and health. In the first case, there are regions in the country that historically have been populated by certain ethnic groups. Afro-descendants for instance, have been located mainly along the Pacific and Atlantic coast, particularly in the departments of Valle del Cauca, Chocó, Cauca, Nariño, and Bolivar where there are the higher number of persons that identified themselves as Afrodescedants. Indigenous groups on the other hand have been located in the south of the country, the departments with a higher percentage of Indigenous population in Colombia are Vaupes, Guanía, Vichada, Amazonas, Cauca, y La Guajira in the north of the country (Departamento-Administrativo-Nacional-De-Estadística, 2019a,b). See Figures 2.5 and 2.6.

The second aspect in the Colombian context is the particular relationship between geography and health; for instance, malaria and other vector-borne diseases are specific in some Colombian geographies; likewise, some natural phenomena such as “El Niño” can affect the incidence rates of infectious diseases and malnutrition. Findings in this regard have shown that health regional differences in Colombia are not only based on ethnic composition but also on characteristics of the regions themselves such as the altitude, temperate, humidity, and precipitation, that create adverse conditions to develop particular diseases (Paéz et al., 2012; Moreno, 2006). The mortality transition in Colombian regions and departments have been thus quite heterogeneous. The differences are similar to differences between developed and developing countries, and although different factors converge for the explanation of the inequalities, it is also clear that the racial factor have been a constant pattern of regional inequalities. Table 2.2 shows the changes in under-five mortality rates in Colombian departments from 1998 to 2012. The estimations did not consider under-registration of mortality, so it is likely that the reduction of mortality can be over-estimated, principally in departments with a high proportion of ethnic population such as Vichada that appears as the de-

Figure 2.6: Distribution of Indigenous Population in Colombian Regions

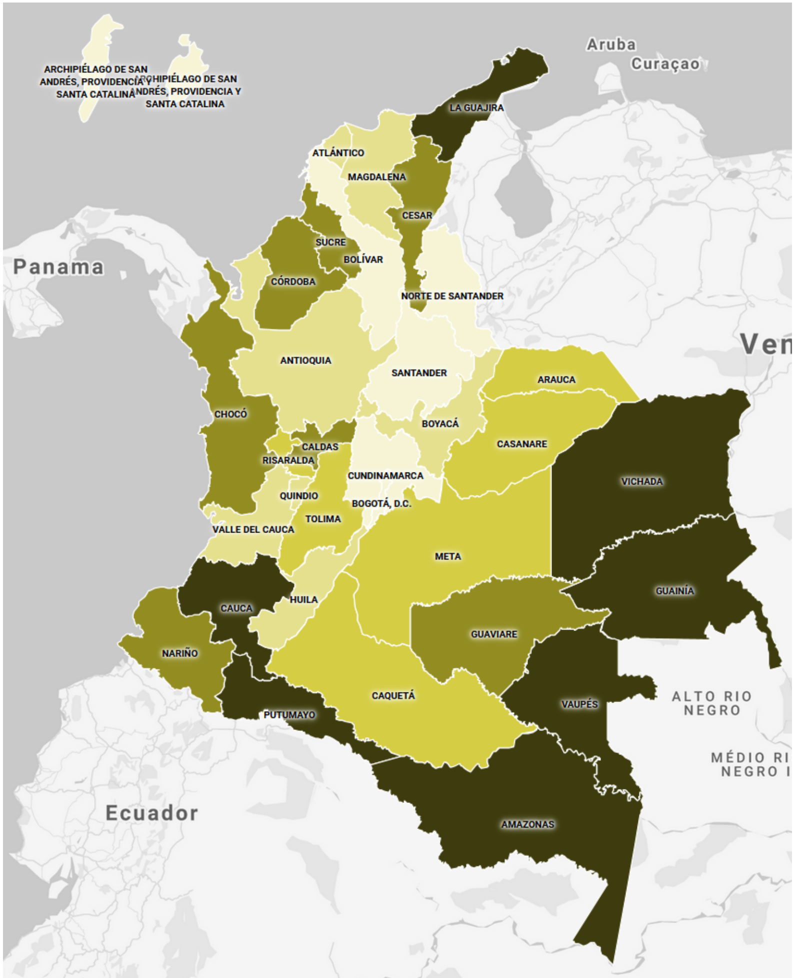


Figure 2.6 The map shows regions with a majority Indigenous population. Indigenous peoples are primarily located in the south of the country and in La Guajira, in the north. Darker colors represent a higher concentration of Indigenous populations. In contrast, the lighter-colored areas in the center of the country are predominantly inhabited by the White-Mestizo population, with a low proportion of Indigenous and Afro-descendant residents.
Source: Own elaboration based on 2018 Census data.

Table 2.2: Changes in Mortality Rate in Children Under Age Five, from 1998 to 2012 by Departments

Department	Rate 1998	Rate 2012	Variation (%)
Vichada	883.0	198.1	-77.6
Quindio	488.5	187.0	-61.7
Caquetá	638.3	249.9	-60.9
Antioquia	391.0	179.3	-54.2
Calda	396.7	187.6	-52.7
Casanare	464.3	221.4	-52.3
Tolima	429.4	206.7	-51.9
Boyacá	398.4	202.5	-49.2
Norte de Santander	412.0	214.1	-48.0
Huila	461.0	247.6	-46.3
Bogotá, D.C	434.3	240.7	-44.6
Risaralda	371.0	207.8	-44.0
Santander	327.5	183.8	-43.9
Guaviare	249.2	141.7	-43.1
Arauca	339.0	193.7	-42.9
Cundinamarca	345.4	206.2	-40.3
Putumayo	279.4	170.2	-39.1
Cauca	413.6	253.6	-38.7
Atlántico	416.6	255.4	-38.7
Valle del Cauca	325.1	200.5	-38.5
Meta	415.5	261.7	-37.0
La Guajira	399.1	251.9	-36.9
Guanía	491.2	314.4	-36.0
Bolivar	398.6	267.4	-32.9
Sucre	293.7	215.2	-26.7
Nariño	229.6	177.9	-22.5
Amazonas	481.9	388.4	-19.4
Magdalena	337.6	274.4	-18.7
Córdoba	352.7	291.8	-17.3
Chocó	379.9	327.7	-13.7
San Andrés y Prov.	242.6	314.8	+29.8
Cesar	251.4	361.3	+43.7
Vaupés	234.9	385.5	+64.1

Table 2.2 presents children mortality rates between 1998 and 2012 and the percentage variation during the period. The Afro-descendant department of Chocó has the lowest reduction in child mortality, while the Caribbean Afro-descendant region of San Andrés y Prov. present an increase of 29.8%. Likewise, the departments of Cesar and Vaupés which report a largest percentage of Indigenous population, increased child mortatlity 43.7% and 64.1% respectively.
Source: (Giraldo et al., 2017, p246)

partment with the greatest mortality reduction. However, in general, the regions where Indigenous and Afro-descendants ethnic groups reside, such as Vaupes, Amazonas, Chocó, and La Guajira, display the lowest levels mortality decline. The department of Cauca for example, in which 24.8% and 32.19% of its population reported themselves as Indigenous and Afro-descendant respectively, shows evidence of mortality improvements; however, they progress at a lower pace than the average department in the country. In some cases, the differences between ethnic regions and the average national indicators is becoming wider; see Table 2.3.

Table 2.3: Demographic Indicators in Colombia and Cauca, 1935-2025

Population/Indicator [a]	1938	1951	1964	1973	1985	1995	2005	2015	2025
CDR (*1000)									
Colombia	22.4	17.2	13.5	9.8	6.8	7.1	7.8	8.9	10.3
Cauca	22.5	17.5	15.0	11.8	10.4	9.2	8.8	9.2	9.8
e0 Male									
Colombia	40.0	45.5	53.3	58.6	63.1	65.3	66.7	67.4	67.9
Cauca	40.9	46.4	50.2	53.4	57.1	59.9	62.2	63.7	64.5
e0 Female									
Colombia	43.8	50.7	56.7	62.8	67.9	70.5	72.2	72.9	73.5
Cauca	44.7	51.6	52.1	56.8	61.9	65.2	68.2	70.3	71.6

Table 2.3 presents comparisons of crude death rates (CDR) and life expectancy at birth (e0) between Colombian national average and the department of Cauca, in which a considerable proportion of Afro-descendant and Indigenous populations are located. The estimations show that differences in CDR and e0 have increased over time.
Source: (Banguero, 2005, p79)

[a] Estimations based on the Censuses from 1938, 1951, 1964, 1973 and 1985. For the years 1995, 2005, 2015, and 2025 estimations are based on projections

Mortality transition is therefore lower in departments with a higher proportion of the population from less advantaged ethnic groups. They present a higher mortality level compared to majority White-Mestizo regions, and additionally, Indigenous and Afro-descendant people are still struggling with deaths attributed to infectious and communicable causes. In 1997, there were $180,910$ people diagnosed with malaria in Colombia; in 1998 it increased to more than $200,000$ cases, the majority of them were from the Pacific Coast, especially from department of Chocó, a typical Afro-descendant region where 200 out of $1,000$ inhabitants acquired the disease (Sarmiento, 2000). In terms of life expectancy, historical estimations show that the Pacific Coast [Litoral Pacífico] has been under the national average. Table 2.4 shows the life expectancy at different ages and by sex; the table compares the trends of the life expectancy in the Colombian Pacific Coast with the life expectancy of the country for the same age group. Although the periods of analysis are slightly different for the Pacific region and the country, it gives however a general idea about the differences in life expectancy between regions during periods of time relatively similar. For instance, in the Pacific region life expectancy for males at age 15 was $47,03$ years during the period $1993-1998$, while at national level almost for the same period: $1995-2000$, it was 53.67 years. The difference was 6.64 years. However, for the period $2008-2013$ for the Pacific region, and 2010 at national level, the difference was 10.5 years. This suggests that the differences in life expectancy between the Litoral Pacífico and the national average have increased over time.

Regions with a higher proportion of Indigenous and Afro-descendants residents in Colombia show

Table 2.4: Life Expectancy in Colombia by Age and Sex in the Litoral Pacífico (L.P) and at National Level (Col)

Age	L.P 1993-1998	Col 1995-2000	Diff.	L.P 1998-2003	Col 2000	Diff.	L.P 2003-2008	Col 2005	Diff.	L.P 2008-2013	Col 2010	Diff.
Male												
15	47.03	53.67	-6.64	46.88	56.99	-10.1	46.90	59.51	-12.6	50.30	60.79	-10.5
20	42.99	49.43	-6.44	42.86	52.76	-9.9	42.89	54.98	-12.1	46.04	56.27	-10.2
25	39.35	46.49	-7.14	39.27	48.93	-9.7	39.30	50.78	-11.5	42.10	52.01	-9.9
30	36.03	42.79	-6.76	36.02	45.07	-9.1	36.06	46.56	-10.5	38.41	47.78	-9.4
Female												
15	53.98	59.11	-5.13	54.54	64,84	-10.3	56.45	65.74	-9.3	58.80	66.69	-7.9
20	49.48	54.59	-5.11	50.00	60.07	-10.1	51.79	60.93	-9.1	54.05	61.84	-7.8
25	45.08	50.09	-5.01	45.53	655.31	-9.8	47.20	56.14	-8.9	49.35	57.02	-7.7
30	40.76	45.49	-4.73	41.16	50.53	-9.4	42.69	51.32	-8.6	44.71	52.19	-7.5

Table 2.4 presents life expectancy estimations for similar periods comparing the Litoral Pacífic (L.P.), which is a region with predominantly Afro-descendant population, with the Colombian national average (Col). The L.P. showed life expectancy for males at age 15 of 47.03 years for the period $1993 - 1998$ while the national average was 53.67 years during $1995-2000$. The difference was 6.64 years of life, but for 2010, this increased to 10.5 years. The difference of lifespan between the average national and the ethnic minority groups has increased over time.
Source: (Carabali et al., 2021, p153)

improvements in life expectancy and mortality, and in general, they present a similar decreasing trend to the national mortality patterns. However, the changes have been in lower proportion, and in some cases, health differences in relation to non-ethnic regions has increased. The present study will analyse the ethnic population at a national level to understand the true state of the mortality differences among ethnic groups in Colombia.

2.5.1 Limitations of Previous Studies

Previous studies on ethnic mortality in Colombia have utilised ecological designs due to the lack of ethnic information at the individual level. For instance, regional data has been employed as a proxy for ethnicity to analyse ethnic disparities in health. However, this methodology has several limitations, including issues with data quality due to under-registration by regions, confounding factors, and challenges in establishing causality at the individual level using group or regional exposure measures (ecological fallacy) (Nazroo and Becares, 2020; Wu et al., 2020; Wakefield, 2004). Moreover, these studies often overlook the consideration of racism as a fundamental determinant of health and mortality inequities.

Using regional data as an indirect approach to examine health ethnic disparities in Colombia provides valuable insights into the social disadvantages faced by ethnic groups in the country. But this methodology presents notable challenges when attempting to distinguish the impact of ethnic discrimination from various other factors, such as geographical areas with higher risk of disease or higher levels of pollution due to economic activities like mining in ethnic regions. Furthermore, the use of geographical regions as a proxy variable for ethnicity assumes ethnic homogeneity within Colombian departments and supposes a similar risk of death for the whole population within a region, although with variations between departments and across the country. These assumptions are quite questionable, especially considering the significant variations of social indicators between the central and peripheral municipalities within departments, as well as the racial heterogeneity of the regional population. Additionally, an ecological perspective on analysing health ethnic disparities may inadvertently neglect ethnic populations residing in major urban centers, particularly Afro-descendant communities, which are primarily concentrated in urban areas. This oversight is notable considering that these urban-based ethnic populations, while constituting a significant portion of the overall ethnic group, may be relatively a small figure in large cities for a regional analyses.

An additional limitation of ecological approaches to analysing racial and ethnic social inequalities is that they usually neglect racism as one of the fundamental causes of social inequities. Colombian studies on health conditions of Indigenous, Afro-descendants, and White-Mestizo groups have normally focused on correlation analysis considering healthcare infrastructure, administrative centralisation, and risk factors linked to the social environment. However, they do not consider the historical conditions of colonisation, slavery, and the "culture" of racial superiority that generates discrimination and exclusion of ethnic minority groups. The consideration of individual-level data in the present document would offer an opportunity to address many of these limitations and provide a more comprehensive understanding of racial and ethnic health disparities in the country.

2.6 Theoretical Perspectives of Ethnic Studies

It has been scientifically proven that racialised groups present poorer health and shorter lifespan than non-racialised groups. However, to what degree ethnic health differences are socially or biologically determined, remains contested. Different approaches have been proposed as theoretical foundation of racial health inequalities. Although the word “race” appears in European languages about the 14^{th} century, it was used to categorise human populations only after the 18^{th} century. After the emancipation of the slaves, the interest in understanding human physical variations was more evident, in order to legitimate a supposed innate superiority in the racial hierarchies in which Europeans were at the top: the so-called scientific racism (Wade, 2008; Braun, 2002; Haller, 1970). More recent studies have analysed the relationship between ethnicity and health from a wide range of perspectives that, in general, can be grouped in five general models: a racial-genetic model, health-behaviour model, socioeconomic status model, psychosocial stress model, and structural-constructivist model (Dressler et al., 2005; Dressler, 1993). A similar analysis focusing on macro-social influences, behavioral risk factors, risk taking and abusive behaviours, adaptive health behaviours, and health care behaviour was made by the National-Research-Council (2004a).

2.6.1 The Racial-Genetic Model

The racial-genetic model considers race differences in health as biologically determined, and race is taken as a biological category through which is possible to elucidate the genetic human diversity (Banda et al., 2015; Jakobsson et al., 2008). Genetic traits of a particular race, therefore, could be considered a risk factor of disease. Sirugo et al. (2019) reports that cystic fibrosis has a high prevalence in Europeans, but it is rare in Africans. This is due to a mutation of a gene named DeltaF508 that causes 70% of the cases in White-Europeans and only 29% in Africans. In contrast, African descendants are almost exclusively affected by cardiac amyloidosis due to a amyloidogenic mutation known as V122I (Buxbaum and Ruberg, 2017; Buxbaum et al., 2006). In Brazil for instance, the Afro population show higher mortality by cardiovascular diseases, primarily by stroke (Lotufo et al., 2007; Kemp et al., 2016). Similar evidence was found among the Afro-population in the United States, where it was suggested that differences in stroke mortality between Black and White people may be a result of biological differences determined by disparities in lipoproteins (Mukaz et al., 2020; National-Research-Council, 2004e). It has been argued that genetic traits play a crucial role in lifespan; so the probability of achieving the oldest age groups in the population is determined, in large part, through a mechanism of genetic selection, in which excessive young-age mortality is a consequence of the frailty or biological pre-disposition to an early death (Vaupel, 2001). In that sense, lifespan largely will depend on genetic factors, and even among people with different social backgrounds, mortality differences will disappear with age if they have resistant genes and survive the mortality selection process of the youth (Beckett, 2000).

The significance of biology in population health studies is unquestionable. However, a racial-genetic characterisation has been questioned for trying to transform race into a biological concept, ignoring that changes in genetic traits are not determined by race, but by geographical regions (Wade, 2012; Sankar et al., 2004). For example, among Black population from West Africa, Black Caribbean, and Blacks from the US, variations in prevalence of heart diseases and hypertension were found. While West-African Blacks show the lowest prevalence with around 15%, the US Blacks displays the highest prevalence with around 35% (Cooper et al., 1997). Despite the poor empirical evidence, and the growing consensus that race is a social construct without biological meaning, the racial-genetic model has remained as it appeared to be consistent with the Western European and American biological construction of race (Sankar et al., 2004).

The racial-genetic model has therefore been thoroughly discredited as a methodology for explaining health and mortality variations among racial and ethnic groups due to its lack of scientific validity. While human genetic variations do exist, race is an inadequate framework for understanding biological differences among humans (Jorde and Wooding, 2004; Goodman, 2000; Ehrlich and Holm, 1964; Wilson and Brown, 1953). Firstly, human variation is continuous across space, with skin color, for instance, gradually changing from one place to another, making it difficult to define the boundaries of racial groups. Additionally, skin color employed as a criterion for racial classification, reflects a non-concordant genetic variation across the human population, indicating that traits like hair, skin color, or height are independently distributed from "resistance genes"; "race" thus fails to predict lifespan or susceptibility to diseases. Moreover, while the understanding of how "race" should be conceptualised remains contested, it is clear that human variations are determined by spatial changes. Individuals of the "same race" but from different geographical locations exhibit more genetic variation than individuals of "different races" located in the same geographic area. These limitations render a questionable approach to explaining social inequalities.

2.6.2 The Health-Behaviour Model

Health-behaviour models, unlike deterministic biological models which argue that health disparities are determined even before people are born, incorporates a stochastic element in health outcomes based on the decisions that the individuals make throughout their lives. In that sense, health inequalities among racial groups are explained by the individual behaviour that, in the case of the black people, is assumed to be less healthy than in the case of a white person (Dubowitz et al., 2011; Winkleby and Cubbin, 2004).

Studies from this perspective focus on smoking, eating habits, physical activity, and alcohol consumption as risk factors of chronic diseases. For instance, some studies have shown that health differences between Whites and Blacks in the U.S are due to the fact that Whites consume on average more fruits and vegetables, have a lower prevalence of smoking, practice more physical activities, and have a lower caloric intake than Blacks (Dubowitz et al., 2011). On the other hand, there is evidence that alcohol intake and cigarette smoking is higher in Whites than in other racial groups (Saint Onge and Krueger, 2017; Krueger et al., 2011). Evidence based on individual behaviour to explain health disparities among social groups is not always consistent; the correlation between racial groups and health behaviours present a cross-over in the long term, as people change their behaviour over time; so although black people have a higher consumption of alcohol and cigarette in young ages, it is not true for older age groups where white population present a higher prevalence (Winkleby and Cubbin, 2004).

This model explains and identifies the determinants of health at an individual level; however, it is unclear why health behaviours change with age, or in other words, it cannot explain why some people cannot change their behaviours. For this reason, the health behaviour model does not explain morbidity and mortality differences *per se*, and it provides a narrow insight into the practice of health in the general society, as they are detached from the social context (Saint Onge and Krueger, 2017). Additionally, health behaviours are shaped by economic and political structures that make it impossible for some individuals to avoid certain risk behaviour, even when they know them; for example, homeless people and sex workers (Frohlich et al., 2001; Link and Phelan, 1995). Therefore, this model could lead to an incorrect identification of the causes of health disparities by focusing on individually-based risk factors, and making individuals responsible for the health outcomes in society. That is, the individual is seen to be ultimately responsible for their behaviour as if there were no systematic influences, sociocultural context, or social meaning ascribed to their behaviour. This has led to an understanding and transformation of the social policy into a "personal policy" based on surveillance and control at an individual level, leaving the social determinants of health inequalities untouched (Frohlich et al., 2001; Link and Phelan, 1995).

2.6.3 Socioeconomic Status Model

Several studies have shown that a large part of the morbidity and mortality differences among racial and ethnic groups are due to socioeconomic status (National-Research-Council, 2004c; Rogers et al., 2000). The measure of social inequalities, however, has not been exempted from debates about what approaches to use when it comes to determine social status, and through which mechanism social advantages are being translated into health advantages (Abel and Frohlich, 2012; Crimmins et al., 2004a; Rogers et al., 2000). It is supposed that health outcomes are determined by social conditions, but there is evidence that the causality could occur in both directions; that is, health may also determine socioeconomic status (Smith, 1999). Nevertheless, there is, in broad terms, a consensus within social epidemiology that macro-social variables determine individual outcomes, and not vice-versa (Link and Phelan, 1995). The influence of socioeconomic status on health is measured through educational levels, income, occupation, and wealth. In the case of education, the mechanism through education influence health conditions is important, as health care behaviour and disease prevention is determined by an individuals education level; higher levels of education are generally associated with better health outcomes (Crimmins et al., 2004a; Lynch, 2003). Likewise, occupation is related to health to the extent that some professions are less exposed to health-threatening conditions than others (Crimmins et al., 2004a). In general, families of higher social status and income have predisposed resources to avoid and buffer health problems (National-Research-Council, 2004c).

The socioeconomic status model shows how changes in living conditions affect health outcomes for social groups, and why racial and ethnic groups show systematically worse health conditions across different societies. The relationship between adverse socioeconomic conditions and higher morbidity and early mortality appear to be linear; however studies have shown that the strength of the correlation varies over time and life stage, and becomes minimal in older-age groups and more intense during the childhood. A large part of the health conditions in the adulthood will be determined by childhood circumstances (Blackwell et al., 2001; Hayward et al., 2000), and part of the mortality gap between black and white population are explained by socioeconomic conditions in early life (Warner and Hayward, 2006). Lower socioeconomic status does not only mean higher mortality in racial groups, but also earlier onset of serious diseases and a longer unhealthy life before death (Crimmins et al., 2004a);

The socioeconomic disadvantages experienced by minoritised ethnic groups result from systematic discriminatory policies that have historically produced and perpetuated ethnic inequalities. Racism should therefore be considered as one of the key factors of social inequities in contemporary societies. Despite often being overlooked in analyses of the causal factors of ethnic health inequalities within the socioeconomic status model, research has demonstrated that racism has long-term effects throughout individuals' lifespans (Stopforth et al., 2022), and its consequences can be transmitted across generations (Hankerson et al., 2022; ONeill et al., 2016). While racism may not always be easily measurable or available for socioeconomic analysis, it profoundly influences the social organisation of the Colombian society, rooted in colonial and slavement historical events.

The socioeconomic status model has some limitations in term of explaining effects on health that are not captured by the social status. For example, the process through which social groups are constituted could potentially generate a selection bias that may improve health outcomes significantly in specific groups (National-Research-Council, 2004f). Biases in selection mechanisms can improve average health outcomes of particular social groups for two reasons. First, the migration process works as a filter in which the stronger and healthier members of each family are normally selected to migrate (because it is believed that they have higher probabilities of success during the stressful experience). This implies that migration improves the average health of receiving countries, and particularly the average health of communities of migrants through a healthy-migrant effect. The so called Hispanic Paradox is one example of group-selection bias in which Hispanic people in the USA, in spite of their lower socioeconomic status, present better health outcomes than the local population (Markides and Coreil, 1986). Second, it is also likely that old and unhealthy migrants may decide to return to their countries of origin to spend the last years of life with their families. This phenomenon is known as salmon-bias effect, as a proportion of migrants return to their homes to die. As a result, mortality measures in migrant population could be underestimated (Palloni and Arias, 2004; Abraido-Lanza et al., 1999). Health differences thus can not be only explained from a socioeconomic status perspective. Additionally, the measurement of health differences based on education could be imprecise, given the existence of considerable differences in quality of education outcomes between Blacks and Whites, even if they have the same years of education (Hanushek and Rivkin, 2006). Similarly, the measurement of income as an indicator of social status may not reflect earnings over ones lifetime (Winkleby and Cubbin, 2004). So income is a good proxy of social status only if the reported income is correct and is measured over a reasonably long period of time. The socioeconomic status model, in general, has problems accounting for health differences when determinants of health fall outside of the market due to *decommodification* (Anderson, 1990). In other words, if relative health advantages are not necessarily "acquired" through the market but are obtained through different mechanisms, such as migration as has been the case of the Latin American Population in the US, then it is probable that we ended up with a healthier population in groups with lower socioeconomic status, which is not necessarily straight forward in terms of explanation.

Health differences are determined by various factors, some of which are intertwined with the ethnic context, such as cultural identity and discrimination in social policies. These factors could significantly contribute to understanding mortality differences between White and non-White groups.

In the Colombian context, it is likely that the disparities in mortality and life expectancy among Indigenous, Afro-descendants, and White-Mestizos groups are associated with socioeconomic factors and geographical segregation. Particularly, racial and ethnic inequalities in the country may be deeply rooted in a colonial discriminatory social structure.

2.6.4 Psychosocial Stress Model

The psychosocial stress model analyses the effect of racism and discrimination on ethnic health differences, and extends the analysis to a new methodological approach of population health inequalities. Racism understood as discriminatory decisions based on racial stereotypes that systematically exclude specific groups in society, has been found at different context including individual, institutional, and structural level (Nazroo et al., 2020; Gee et al., 2009; Jones, 2000; Krieger, 1999). Negative life experiences related to racism, prejudice, and discrimination are considered to play an important role as determinant of health in the psychosocial stress model; a growing number of studies have documented that racial discrimination is positively associated with psychological distress (Myers and Hwang, 2004; National-Research-Council, 2004d; Williams et al., 2003). Life experiences thus play a important role in health inequalities in the psychosocial model, to the extent that negative attitudes towards minorities have an effect on the chances for areas of residence, housing, employment, as well as generate unfair treatment, limited access to health care, and increase exposure to hostile psychosocial environments (National-Research-Council, 2004b; Williams and Collins, 2001; Brown et al., 1999).

Interpersonal racism and discrimination refer to individuals' racist beliefs and behaviors that stem from conscious and unconscious biases, resulting in harm to members of racialised groups (Kornienko et al., 2022; Hall et al., 2015; Krieger, 1999). It has been identified principally in context of low racial density and racial homogeneity where non-White Individual report a high level of discrimination in face-to-face interactions (Bécares et al., 2009). The majority of White population in the US, for example, have beliefs about black inferiority: in particular, they believe that most Blacks are lazy, prone to violence, unintelligent, and that they prefer to live on welfare support (Williams, 1999). In general, Whites are reluctant to endorse positive stereotypes to African descendants, and on the contrary, prejudices against minorities are legitimised. This leads to the phenomenon of residential segregation in areas with poor-quality housing: humidity, inadequate heat, problems with noise, and sometimes with an elevated exposure to noxious pollutants and allergens, and pharmacies with inadequate medication stocks to treat people with health problems. Additionally, the lack of recreational facilities in minority neighbourhoods, as well as concerns about personal safety, discourage residents from physical activities during leisure time (Williams and Collins, 2001; Williams, 1999).

Institutional racism and discrimination, on the other hand, refer to procedures and policies from formal and informal institutions that consistently penalise and disadvantage people from non-White groups (Paradies, 2021; Better, 2008). This phenomenon has been observed in various domains, including the labor market against ethnic minority applicants (Zschirnt and Ruedin, 2016; Nazroo and Williams, 2006; Williams, 1999), and in the justice system (Cunneen, 2006; O'Grady et al., 2005), among others. Black applicants have 50% less chances to be hired than a white person with similar qualification, and even when a picture is not available, employers show a clear preference for white names (Pager and Shepherd, 2008; Bertrand and Mullainathan, 2004); as a result, institutional discrimination restricts the access of ethnic minorities to resources that would support the attainment of better health status (Dressler et al., 2005). Additionally, the experience itself of being discriminated is positively correlated with blood pressure, stress, cigarette smoking, and alcohol use (Johnston and Lordan, 2016; National-Research-Council, 2004b). Structural or systemic racism refers to racism that permeates across institutions and is rooted in a historical perspective of centuries of slavery, segregation, narratives, and ongoing white oppression. These systemic practices limit and restrict the access to good and services of non-White people, including access to adequate health care (Feagin and Bennefield, 2014; Reskin, 2012; Bonilla-Silva, 1997). Systemic racism considers racial discrimination not only as an isolated event in society but as a macro structure in which discrimination in the education system, labor market, justice, segregation, and all forms of discrimination are intertwined with cultural patterns, making it difficult to identify racism using conventional causal models. Figure 2.1 presents the dynamics of a social system of racism.

The psychosocial stress model suggests, in general, a correlation between subjective experience of discrimination and negative health outcomes. The pathways through which experiences of racism are translated into health disparities have been analysed from an ecosocial approach, which considers that the origins of diseases cannot be separated from the social conditions and interactions in which people are born and live. Social experiences become biologically incorporated, leading to a kind of embodiment of the social milieu that determines population patterns of health, disease, and well-being (Krieger, 1999; Krieger and Zierler, 1997; Krieger, 1994). Additionally, alternative approaches to explaining the inter generational effects of racism and historical traumas are based on collective memories (Evans-Campbell, 2008; Brave-Heart and DeBruyn, 1991), and the epigenetic perspectives that suggests the transmission of a "gene for trauma" as a result of highly stress experiences, which could lead to modifications at the molecular level, causing an intergenerational trauma (Yehuda and Bierer, 2009). Stress is therefore one of the main factors through which racism is transformed into health inequalities, as minorities experience more stress than other groups. However, such experiences may be less harmful, as they may have also developed better coping strategies through family support, community, and religious involvement, which mitigate the negative effects (Blaxter, 2005). Likewise, the presence of possible socioeconomic confounding variables needs to be considered, as they may intensify or minimize the effect of racism.

2.6.5 Constructivist-Structuralism Model

Constructivist-structuralism theory is an attempt by Bourdieu (1984) to explain the social dynamic and the construction of the social reality. "By structuralism or structuralist, I mean that there exist, within the social world itself and not only within symbolic systems (language, myths, etc.), objective structures independent of the consciousness and will of agents, which are capable of guiding and constraining their practices or their representations. By constructivism, I mean that there is a twofold social genesis, on the one hand of the schemes of perception, thought, and action which are constitutive of what I call habitus, and on the other hand of social structures, and particularly of what I call fields and of groups, notably those we ordinarily call social classes" (Bourdieu, 1989, p14). According to Bourdieu, the construction of social reality involves two primary components: first, a social structural constraint, or objective reality; and second, the cognitive processes through which individuals internalise the meaning and representation of social structures, or subjective reality. In other words, the constructivist-structuralist model posits that, on one hand, social groups are "located" within a particular social space that encompasses hierarchies and power relations. On the other hand, each social space possesses cultural knowledge that shapes the identity of its members and influences how individuals understand and represent their social reality

Racial ethnic disparities in health can be analysed from a constructivist-structural perspective, as culture (objective reality) plays an important role in terms of the lifespan. Health differences therefore can arise because ethnic groups adopt different social practices in the form of *habitus*. That means, some cultural patterns and living arrangements of ethnic groups that could affect health conditions are socially determined and culturally transmitted. The concept of *habitus* is one of the Bourdieu's central elements in the constructivist-structuralism model, and should be understood as the internalisation of ways of behaving and thinking that people acquire within the family. Unlike "health behaviours" that appear as a voluntary choice devoid of cultural meaning, the *habitus* are social practices that are unconsciously and involuntarily adopted as cultural "inheritance" that has been transmitted from one generation to another; therefore, *habitus* is less sensitive to changes in socieconomic status than "health behaviours". In the case of eating habits of Indigenous or Black people for example, they are not going to change considerably even under an scenario of higher income and wider range of food options, as cultural patterns are embedded in the culture and are less sensitive to economic changes. This behaviour socially structured plays thus a important role in health differences among ethnic groups, and constitutes the cornerstone of the constructivist-structuralism model in the analysis of the ethnic mortality differences. For instance, ethnic groups in central Asia shows different health outcomes based on their cultural practices. Russian ethnic groups, in spite of having a higher socieconomic status than indigenous groups such as the Kazakhs, the Kyrgyz, the Tajiks, the Turkmens, and the Uzbeks, present higher mortality due to alcohol consumption. Indigenous groups are prohibited from drinking alcoholic beverages because of religious reasons, while Russians reported a relatively high intake of alcohol in social interaction (Guillot et al., 2011). The constructivist-structuralism model analyses how cultural practices affect health outcomes, how they account for ethnic health inequalities, and why they should be consider in the design of health policies.

Figure 2.7: Analytical Approaches to Understanding Ethnic Health Inequalities

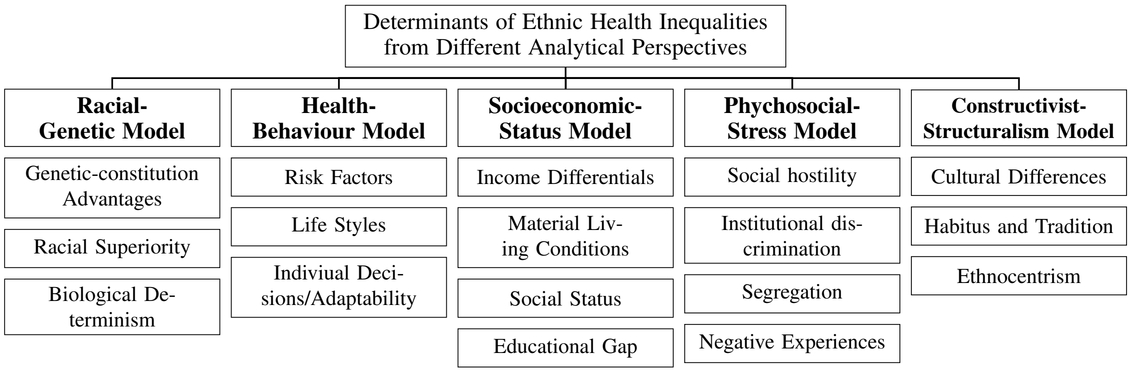


Figure 2.7 summarises the main characteristics of the different approaches to analyse ethnic health inequalities. In practice, it is probably impossible to separate the effects from one model to another; however, in the Colombian context given the visible socioeconomic contrast among ethnic groups, it is believed that the Socioeconomic Status model could explain a large part of the health inequalities among White-Mestizos, Afro-descendants, and Indigenous populations in the country. Source: Own elaboration

The different models we have explored offer various ways to understand racial and ethnic disparities in Colombian society. Each approach has its strengths and limitations. Health determinants are complex, involving biology, behavior, social dynamics, and cultural factors, all of which shed light on the roots of these inequalities and how to address them. Based on our research questions and available data, we will focus on a socioeconomic approach. This means we will consider the demographic and socioeconomic conditions of Indigenous, Afro-descendant, and White-Mestizo populations to better understand mortality risks and underlying causes. However, it is crucial to highlight that the social disadvantages faced by ethnic minorities are not random occurrences; they stem from systematic marginalisation rooted in historical colonial structures.

CHAPTER 3

Ethnic Information in Census and Mortality Registers

The data for the study on ethnic mortality in Colombia are drawn from two different sources: the Census of 2018 and the mortality vital registration records from 2008 to 2018. Censuses have been collecting information on Colombia's Indigenous and Afro-descendant populations since the Spanish colonial period, with the first census held in 1770 (Uribe, 1998). However, the purpose of this census was to count the number of Indigenous that paid taxes and the number of slaves for commercial reasons (Departamento-Administrativo-Nacional-De-Estadística, 2007). After the foundation of the Colombian republic, the new constitution in 1991 proclaimed Colombia as a multi-ethnic and multilingual country, based on self-reported identity.

3.1 Census Data

Ethnicity as self-reported information has been collected in the last three censuses, in 1993, 2005, and 2018 respectively. However, in 1993 census the question was incorrectly formulated, and the information collected could not be used to capture the true ethnic identity of the Colombian society (Rodríguez, 2010; Uribe, 1998). The censuses of 2005 and 2018 formulated the question for ethnic classification in terms of cultural practices, descendant, or physical characteristics that identify a person with a particular ethnic group; see Figure B.

The 2018 Census registered $242,744$ people who died in the year 2017. Figure B shows the census question used to collect mortality information. This information represents the most complete information on ethnic mortality in Colombia in terms of coverage and registration information, although it is available only for the census year. Nevertheless, the Census 2018 data was not exempt from challenges, as it had problems in its coverage: according to post-census correction methods, $4,094,077$ persons were not counted, of which $1,688,936$ were Afro-descendant people. The total population that identified themselves as Afro-descendant in the Census 2018 were $2,982,224$ persons, which represents a decrease of 30% compared to the previous Census 2005, in which $4,311,757$ persons reported themselves as Afro-descendants.

It is important to note that mortality data from the 2018 Census –refer to Appendix B for a description of the census variables- is likely to be biased for the Afro-descendant ethnicity due to significant omissions, particularly affecting individuals living in impoverished areas who are at a higher risk of not being recorded. This potential undercount could lead to misinterpretations regarding the extent of health disparities between Afro-descendants and other ethnic groups. Nevertheless, despite the underestimation in the 2018 Colombian Census, the inequality between groups remains evident. Subsequent chapters will demonstrate that the extent of bias in census data is less pronounced compared to other sources on mortality

3.2 Death Registration Data

National Death record data, published by DANE, collect the information regarding all registered deaths in Colombia each year. From 1980 to 2018, a total of $6,974,345$ deaths were registered: about $183,000$ deaths by year on average (Figure 3.1). The method of collection for mortality information has changed over time: deaths certificates from 1980 included 15 variables, whereas certificates for the year 2018 included 70 variables. A description of the variables included in the mortality registers is shown in Appendix A. Information about causes of death is based on *International Classification of Diseases, Ninth Revision* (ICD-9) until the year 1997; for the following years, death registers were recorded under ICD-10 form. Changes of ICD classification and registration forms could bring difficulties in the analysis of time series data, as the information was codified differently over time (Pechholdová, 2009). In general, death certificates were modify 5 times from 1980 to 2015: in the years 1992, 1997, 1998, 2001, and the last modification was in the year 2008.

The Tenth Revision (ICD-10) differs from the Ninth Revision (ICD-9) in several ways, although the overall content is similar: First, ICD-10 is printed in a three-volume set compared with ICD-9s two-volume set. Second, ICD-10 has alphanumeric categories rather than numeric categories. Third, some chapters have been rearranged, some titles have changed, and conditions have been regrouped. Fourth, ICD-10 has almost twice as many categories as ICD-9. Fifth, some fairly minor changes have been made in the coding rules for mortality (WHO, 2021; CDC, 2020).

Figure 3.1: Total Deaths in Colombia by Year from 1980 to 2018, All Causes

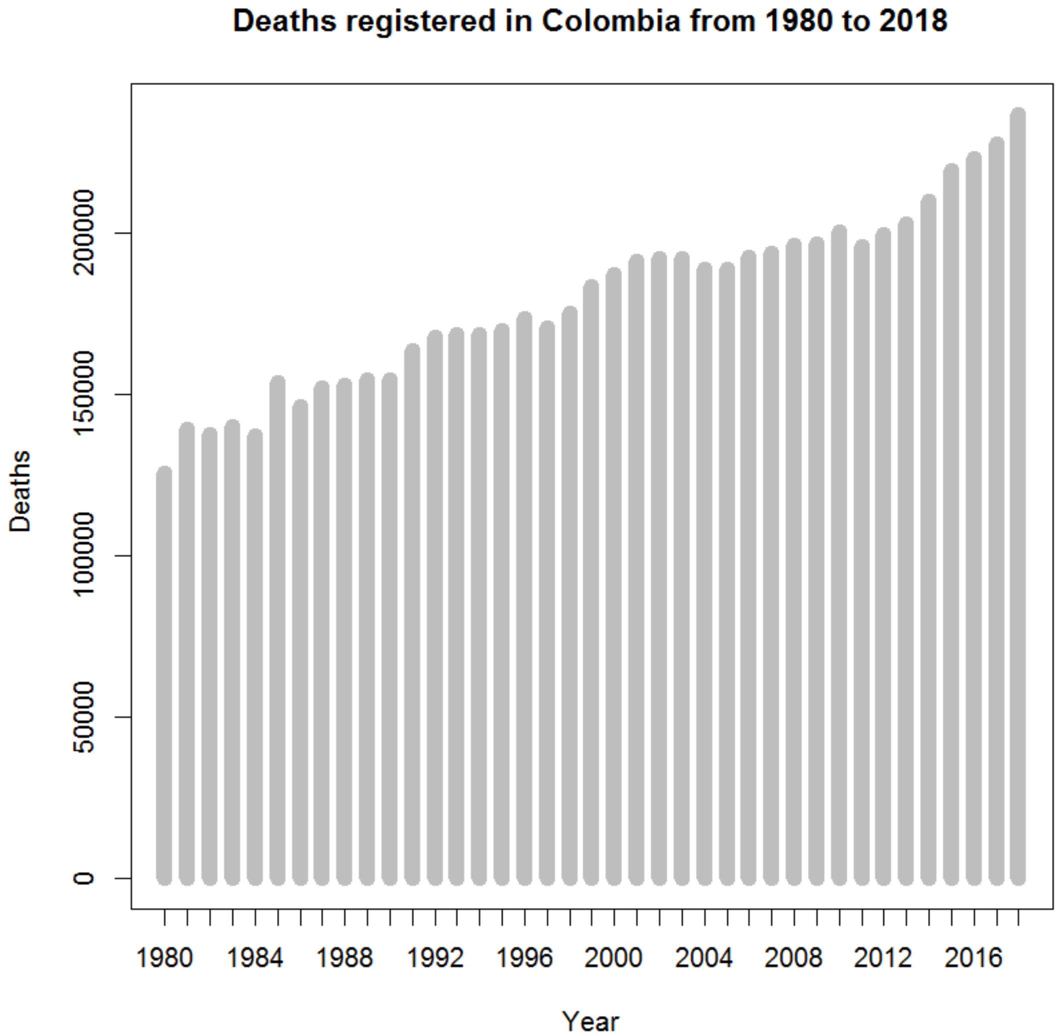


Figure 3.1 shows the total number of registered deaths in Colombia from 1980 to 2019. In 1980 the total number of deaths were around 120,000 and in 2019 it was around 220,000 deaths.
Source: Own elaboration based on mortality registers

3.3 Descriptive Statistics of Demographic Variables

The dataset in its latest version contains seventy variables, many of which will be employed in the later chapters of this thesis. The descriptive characteristics of reported deaths between 1980 and 2018 are briefly discussed. The proportion of male deaths, for example, have been historically higher than those of females; see Figure 3.2.

Figure 3.2: Deceased Population by Ethnicity and Gender in Colombia from 2008 to 2019

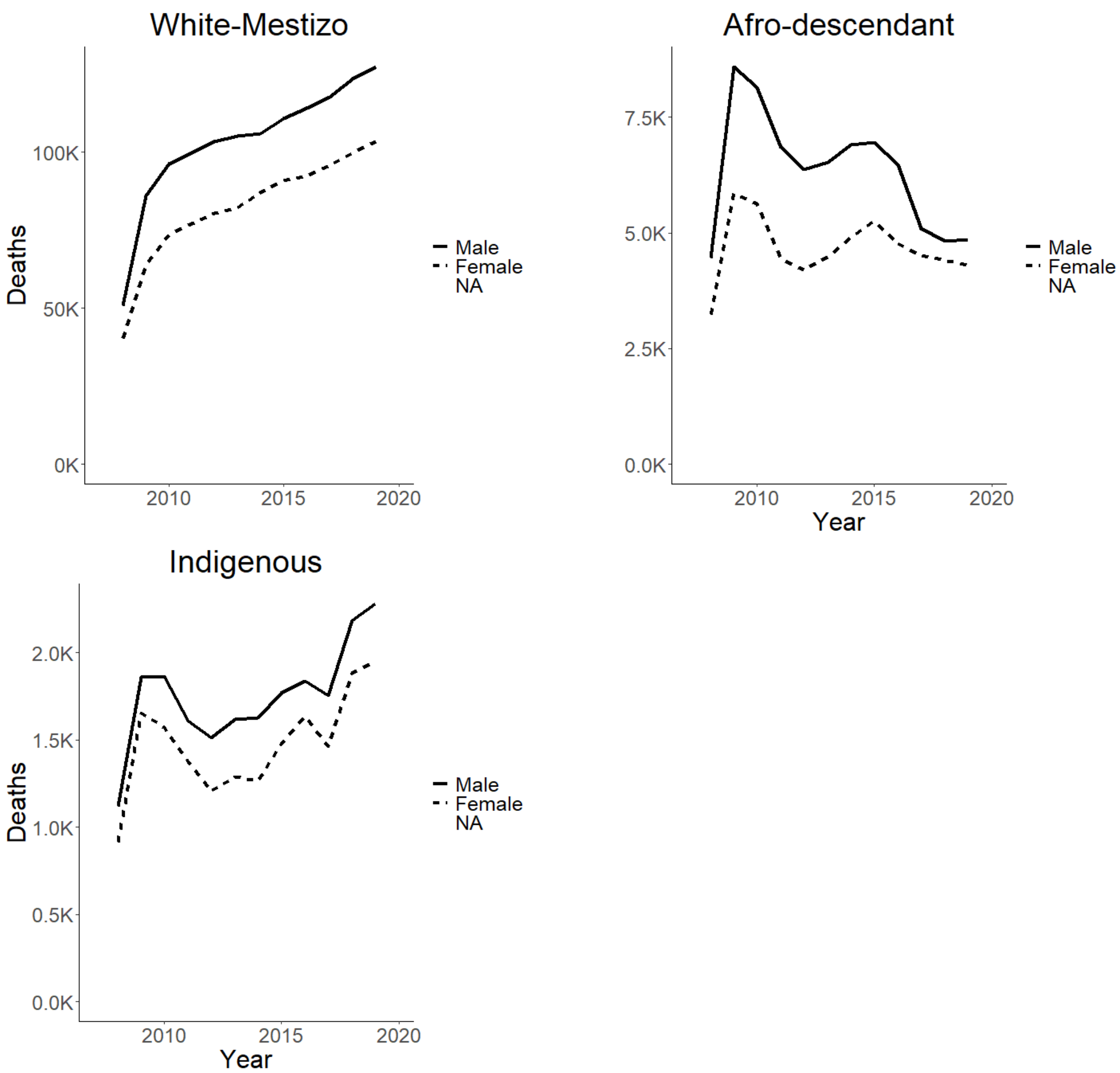


Figure 3.2 shows mortality trends by gender and ethnic group. The ethnic variable was introduced in the mortality records in the year 2008, and for that first year the number of registered people was relatively lower than for the rest of the years. In all ethnic groups, male mortality has been historically higher than female mortality.
Source: Own elaboration based on mortality registers

The ethnic groups present relatively different age mortality patterns that are related to different degrees to the general population structure presented in Figure 1.2, Chapter 1. A description of the age distribution of total deaths by ethnicity from 2008 until 2019 is shown in Figure 3.3.

Figure 3.3: Mortality Distribution by Age and Ethnicity in Colombia from 2008 to 2019

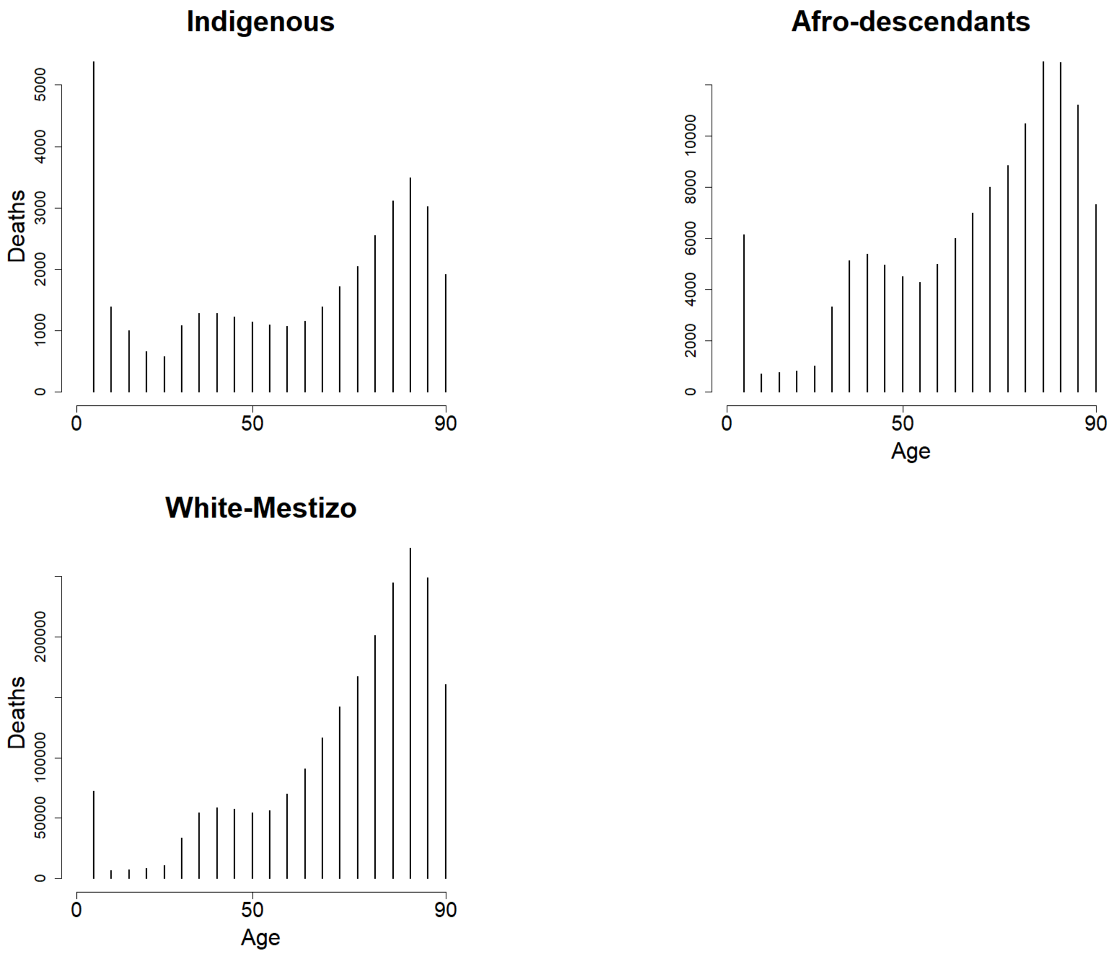


Figure 3.3 shows age mortality patterns by ethnic group. Indigenous population shows a high proportion of deaths in children, while Afro-descendants has a particular high mortality level between ages 25 and 29. White-Mestizos on the other hand, concentrates the largest proportion of deaths at older ages.
Source: Own elaboration based on mortality registers.

Educational attainment of deceased people is available from 1998 onwards, see variable 23 in Appendix A. Figure 3.4 shows the educational level of the deceased population by ethnic group. The level "Low" makes reference to people with primary school or less, "Middle" to secondary school, and "High" to people with more than secondary school. In year 2008 the variable educational attainment changed from 9 to 14 categories that included technical and semi-professional education as well as specialization, master, and higher level of education. The modification of the categories in the variable education in year 2008, is likely the cause of the unusual patterns of that variable; see Figure 3.4.

Figure 3.4: Educational Attainment and Ethnicity of the Colombian Deceased Population Between 2008 and 2019

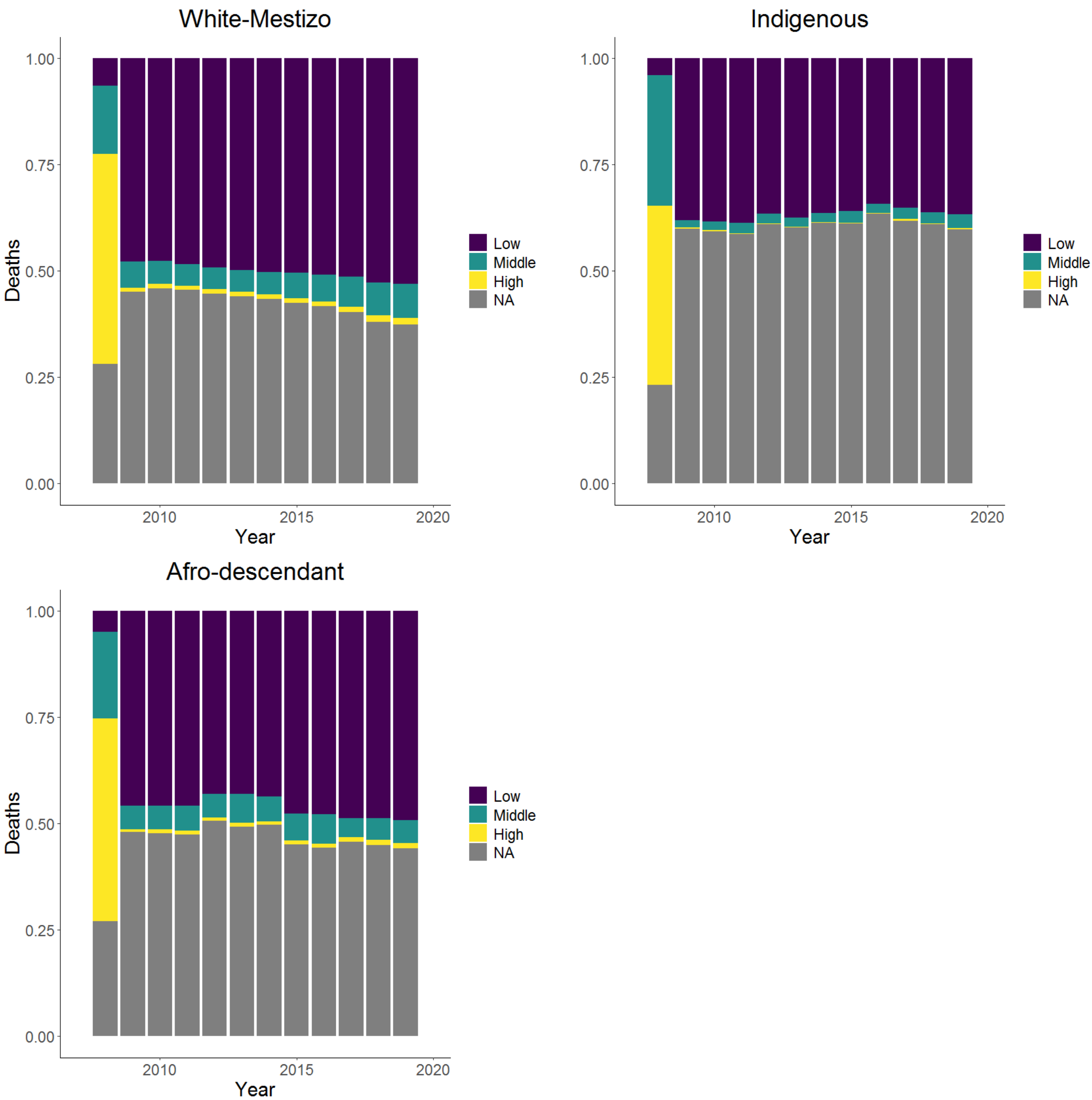


Figure 3.4 shows the educational attainment of deceased by ethnic group between 2008 and 2019. For the first year information about education shows an atypical pattern probably because it was the first time this information was classified by ethnicity. The level of missing information is significantly high, in particular for the Indigenous group.
Source: Own elaboration based on mortality registers.

Information about marital status in the mortality registers has changed over time. The variable was modified in 1997, and it moved from 5 to 9 categories. The current version of this variable includes different forms of cohabitation and the category "divorced" as an independent category. Figure 3.5 presents the marital status of deceased people by ethnicity.

Figure 3.5: Marital Status of the Deceased Population in Colombia by Ethnicity from 2008 to 2019

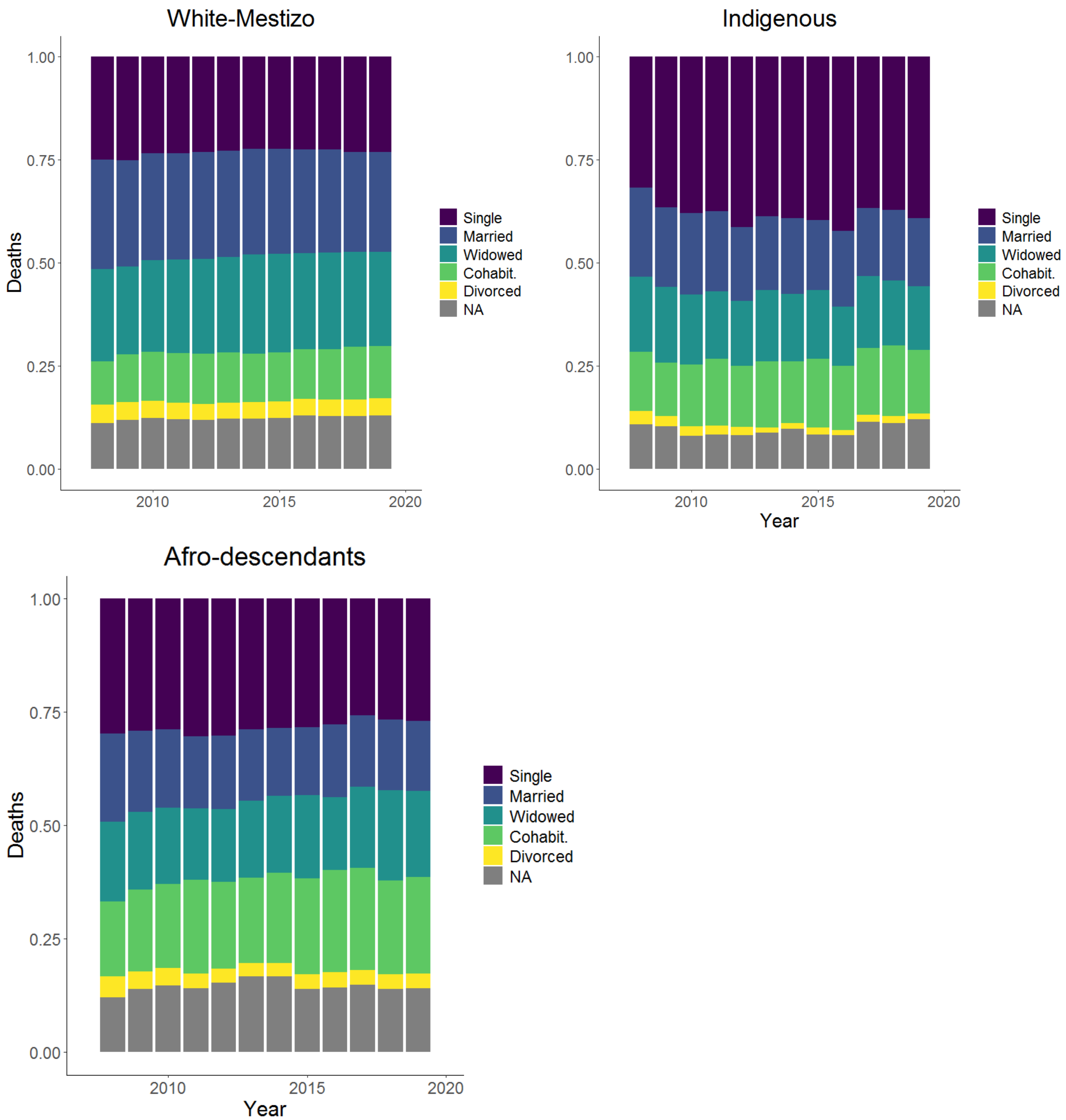


Figure 3.5 presents information regarding marital status in deceased people. Information includes all age groups. The proportion of the different categories have remained relatively constant over time, except for the category "single" in Indigenous which showed a slight increase.
Source: Own elaboration based on mortality registers.

The question about ethnicity was introduced in the mortality records in 2008. Figure 3.6 shows the proportion of people from different ethnic groups that were registered between 2008 and 2018. Ethnic classification in Colombia is self-reported, but in the case of deceased people, family members can report the ethnicity of deceased members, and children are classified with the ethnicity of their mother. According to 2018 Census, 4.4% of the total population identified themselves as Indigenous, while 9.3% reported an Afro-descendant identity.

Figure 3.6: Ethnic Identification of Deceased People in Colombia from 2008 to 2018

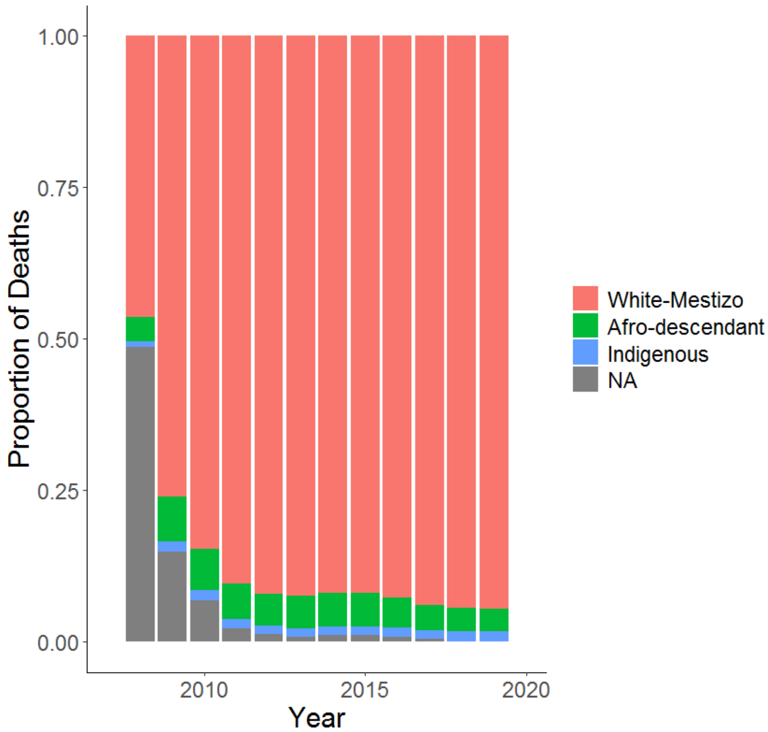


Figure 3.6 presents the proportion of registered deaths by year that belong to each ethnic group. The largest group is the White-Mestizo, follows by Afro-descendants and Indigenous respectively. Source: Own elaboration based on mortality registers.

The limitations of census data and mortality records in capturing information about Indigenous and Afro-descendant populations may be a consequence of power dynamics among ethnic groups in the country. It is crucial to understand that the construction of statistics cannot be separated from the construction of national identity and power structures. The under-registration or "statistical genocide" of ethnic groups may have been a state policy in countries with a colonial past, aimed at socially and politically “minoritise” entire communities. Liberal states in the Latin American region, under the guise of multiculturalism, acknowledge the coexistence of social and cultural diversity only under the condition that it does not challenge historical power dynamics or the dominance of certain cultural and social groups (Azpiroz, 2022). Multiculturalism, therefore, appears to be incompatible with political power and territorial autonomy, transforming social statistics into a contested arena.

Despite these considerations, the foregoing data review aims to present general differences among ethnic groups and identify the main demographic characteristics that are still visible. Additionally, this review identified possible data classification problems and the level of incompleteness for some variables. For example, the high number of NAs in variables such as "Educational Attainment," which has remained almost constant since the ethnic variable was introduced in mortality records, highlights the complexity and challenges of collecting this politically sensitive information in combination with other socioeconomic information. In contrast, NAs regarding ethnic classification have decreased to almost zero. The reduction in this category could be interpreted as an improvement in data quality. However, this could be an effect of the "banalisation" of ethnic identities and a lack of interest by central the government. If ethnic identities are to be self-reported in Colombian society, it is at least suspicious that ethnic information from deceased people may have lower missing values than ethnic information from the census. It may be that in some cases, missing ethnic information from deceased individuals may have been completed by public officials or health personnel. This underscores the need for the evaluation and improvement of ethnic data, which will be considered in more detail in subsequent chapters.

CHAPTER 4

Measuring Ethnic Mortality in Colombia: The Under-registration Gap

4.1 Introduction

Research studies have well acknowledged that ethnic minority groups often face social and health disadvantages across the life course compared to the general White populations (Nazroo et al., 2006; Hummer et al., 2004; Crimmins et al., 2004b; Williams and Collins, 2001). These racial inequalities and discrimination persist even after death, as evidenced by the under-registration of deaths among ethnic and socially disadvantaged groups. Consequently, there is a lack of understanding of mortality patterns among specific ethnic sub-groups of the population, leading to a dearth of coherent social policies aimed at addressing the widening health and healthcare inequities. This is particularly evident in Colombia, which represents a multiethnic, and multicultural society within Latin America.

Studies have shown that under-registration of births and deaths are generally common in low- and middle-income countries, and there is little understanding of the impact of under-registration on mortality levels, especially amongst ethnic and underprivileged social groups. Barriers to birth and death registrations have been often attributed to poor vital registration and monitoring systems, and lack of coverage (Tin Oung et al., 2017; Setel et al., 2005; Tsung-Hsueh et al., 2002); as well as due to socioeconomic and geographic disadvantages including poverty and regional deprivation (Sun et al., 2018; Tin Oung et al., 2017; Guo et al., 2015; Aponte-Gonzáles et al., 2012; Duryea et al., 2006). In Europe and North America, mortality under-registration has been found mainly in maternal mortality due to the high level of misclassification (Deneux-Tharaux et al., 2005; Schuitemaker et al., 1997; Bouvier-Colle et al., 1991), but with no significant impact at the population level.

The lack of studies on ethnic minority groups can be attributed to data scarcity as well the poor quality of available data. The present study contributes to this knowledge gap by analysing the levels of under-registration of mortality in Indigenous, Afro-descendant, and White-Mestizo populations in Colombia. We contrast the number of deaths among different ethnic groups obtained from the civil vital registration records (henceforth mortality or death records) with the number of deaths that were recorded in the 2018 national Census (Censo de Colombia de 2018). A similar study was not possible before because although the previous census held in 2005 included ethnic information, death records by ethnicity was included only in 2008. Therefore, this study is the first analysis to systematically investigate the level of under-estimation of ethnic mortality in Colombia with a focus on the extent of unrecorded deaths in official mortality statistics, and how this varies among Indigenous and Afro-descendants minorities in comparison to the non-ethnic White-Mestizo group. Further, we estimate the probability of a deceased person of ethnic origin being omitted from vital statistics and how sociodemographic characteristics make one more prone to being excluded from registration.

4.2 Literature review

Accurate mortality registration is one of the major challenges for government official statistics in terms of the limitations underpinning data capture of the total number of deaths that occurred in a specific time and geography. Half of the countries worldwide present incomplete mortality data, and little progress has been made so far. For instance, between 2000 and 2012, the percentage of deaths registered globally increased marginally from 36% to 38% (Mikkelsen et al., 2015). Additionally, problems related to age heaping and digit preference (Barrett, 2019; Manish, 2017), misreporting (Agrawal and Kandhuja, 2015; Preston et al., 1996), and under-registration pose considerable drawbacks to understanding the real demographic patterns of sub-populations, particularly of ethnic minority groups.

The level of mortality under-registration is higher in populations under five years old (Campos de Lima and Lanza, 2014; Prasartkul and Vapattanawong, 2006; Yang et al., 2005; Blakely et al., 2000; Lumbiganon et al., 1990), and in rural and sparsely populated areas (Sun et al., 2018; Urdinola et al., 2017; Tin Oung et al., 2017; Guo et al., 2015; Aponte-Gonzáles et al., 2012; Duryea et al., 2006) where often the presence of governmental institutions or local authorities is relatively limited. As a result, facility-based registration systems are frequently biased towards those who rely on the national database systems, and fail to capture information of deaths which occur at home and in remote rural areas. This phenomenon is particularly common in less economically developed countries (Tin Oung et al., 2017; Setel et al., 2005; Tsung-Hsueh et al., 2002), where rural areas often have unreliable sources of data due to limited data collection resources or infrastructure. Nevertheless, under-registration and misclassification are not exclusive problems of poor countries. Maternal and pregnancy-related mortality are often underestimated in the United States, the Netherlands, France, and other European countries (Deneux-Tharaux et al., 2005; Schuitemaker et al., 1997; Bouvier-Colle et al., 1991). The percentage of under-registered maternal mortality varies from 22% in France to 93% in Massachusetts-United States of a sample of 68 and 15 pregnancy-related deaths respectively (Deneux-Tharaux et al., 2005). The reasons for the underestimation are attributed to inaccurate medical diagnosis, errors in data entry by coding clerks, confusions arising from specifications of morbidities with the International Classification of Diseases (ICD) as well as complex pathologies that lead to misclassification or omissions, such as pregnancy related deaths (Bouvier-Colle et al., 1991). In resource-poor settings, the reasons for under-registration of deaths are attributed to economic, legal, administrative, geographic, and cultural barriers (Duryea et al., 2006), sometimes classified into structural and social barriers (Rao et al., 2009). For example, lack of trained and qualified civil servants, insufficient personnel, poor verification mechanisms before funerals, unclear rules and administrative guidelines, costly registration process, and a centralised administration system can all result in inaccurate recording of deaths (Sun et al., 2018; Campos de Lima and Lanza, 2014; Cendales and Pardo, 2011; Rao et al., 2009; Ruiz, 1988; Schuitemaker et al., 1997).

On the other hand, the socioeconomic conditions of communities play a pivotal role in determining the effectiveness of registration systems. Factors such as poor educational attainment, limited understanding of administrative procedures, bureaucratic complexities, poverty, low income levels, geographical remoteness, transportation constraints, absence of personal identification in rural areas, cultural biases regarding specific causes of death, and inadequate incentives for reporting deaths can all exert influence on data recording (Tin Oung et al., 2017; Aponte-Gonzáles et al., 2012; Rao et al., 2009; Duryea et al., 2006). For instance, in Thailand, one significant reason for mortality under-registration is the perception in underserved communities that death registration is only necessary for matters such as financial reimbursement, inheritance claims, or life insurance applications. Consequently, death registration is often deemed unnecessary, especially in cases involving children, if it does not serve a legal purpose (Prohmmo and Guest, 1996; National-Statistics-Office-Thailand, 1969) cited in Prasartkul and Vapattanawong (2006).

Under-registration in Colombia is heterogeneous across provinces and regions; in the Atlantic Coast, only about 40% of total deaths are registered and of which only 5% have complete information, whereas, in comparison, the central part of the country has 93% and 38% of coverage and complete information respectively (Ruiz, 1988). There are no systematic analyses of mortality under-registration among ethnic populations in Colombia. National statistics show the highest coverage of mortality registration in cities with a higher proportion of White-Mestizo people. For example in the capital city of Bogotá, death registration is as high as 95.4%, however in a *department* (upper administrative unit combining municipalities) with a high concentration of Afro-descendant population, Chocó, the coverage is only 64.5% (Departamento-Administrativo-Nacional-De-Estadística, 2006). Additionally, under-registration of deaths by causes or diseases such as cancer present a similar pattern, and the low registration in the departments of Chocó, and Vichada with a higher representation of Indigenous populations (Aponte-Gonzáles et al., 2012). In general, registration of mortality by cancer is higher in areas with predominance of White-Mestizo population, and in the main cities[1] of the country (Hernández et al., 2020).

Studies from other countries have shown that mortality in ethnic minority groups is more likely to be under-estimated and under-registered (Guo et al., 2015; Blakely et al., 2008; Tamargo, 2007; Blakely et al., 2000). Socioeconomic and cultural factors can make the reporting and registration process more difficult. Likewise, the region of residence has been also an important determinant of under-registration (Ribotta et al. 2019, p.16; Campos de Lima and Lanza 2014; Ruiz 1988). Lack of knowledge of the registration process, the difficulties with the official language and the dismissive treatment by civil servants that discourage underprivileged people to participate in the process can increase under-registration of deaths (Tamargo, 2007). On the other hand, incomplete information in the mortality records due to missing age, lack of medical certification, and ill-defined causes of death are also common (Cristancho 2017, ch.2; Cendales and Pardo 2011).

In summary, low coverage and lack of completeness in mortality data can be explained from two different perspectives. The first is to assume that governmental institutions fail to adequately capture or register the total number of deaths. The second is the causes of data incompleteness from the perspective that populations do not report the total number of deaths when they happen. These two different phenomena lead to the same problem of *under-registration* and *under-reporting* indiscriminately. However, by labelling situations in which the information is not captured by the national registration systems as *under-reporting*, the responsibility is assigned to the reporters, even in a scenario that registration was not possible due to coverage limitations or because the registry lacked proper stationery (Duryea et al., 2006, p.11). In the second case of *under-registration*, the responsibility lies principally on the registration system and its ineffectiveness in collecting the information of the whole population. In this paper, the term *under-registration* will be preferred to *under-reporting*.

4.3 Method

Data for the analysis of mortality under-registration analysis was provided by the National Statistics Office (DANE). We used two data sources: the Colombian National Population Census 2018 and the Mortality Records 2017, see Figure 4.1. Since we could not match the individual datasets, we did not consider the capture-recapture method.

These two datasets relating to Colombian mortality were collected using different methodologies. Mortality records were based on the number of deaths registered by national governmental institutions, while the population census recorded the number of deaths by asking respondents in their

[1]It has to be noted that the capital city is independent of the department and is presided by a Mayor as the official head whereas the head of the department is a Governor.

place of residence if any household member had died in the year prior to the census. The Colombian census is conducted approximately every ten years, with the most recent censuses taking place in 2005 and 2018

Ethnic classification in the Colombian national statistics is carried out on the basis of self-reported definitions, which means that a person, regardless of their skin colour, can report the ethnic belonging or cultural identity that they consider appropriate within the six officially recognised ethnic categories: Indigenous, Roma, Black-Caribbean, Black-Maroon, Black-Mulatto, and White-Mestizo. Ethnicity should be understood as a collective identification that goes beyond skin colour and it is related more to the existence of kinship, ancestors, and common social practices (Wade, 2008, 1993a). Ethnicity is, therefore, related to a feeling of belonging to a social group with whom the person in question shares similar characteristics in terms of language, beliefs, customs, and in general, the same cultural background and phenotype, although the latter is not a sine-qua-non condition.

Table 4.1: Number of Deaths by Ethnic Group and Type of Register.

Colombian Ethnic Mortality 2017								
Data source	Indig.	Roma	Black Caribb.	Black Maroon	Black Mulatto	White Mestizo	Un-known	Total deaths
Death Records	3281	82	186	59	9370	213716	990	227624
Census	14548	50	145	65	19972	204883	3081	242744

Source: Own elaboration based on mortality registers and 2018 Census.

Table 4.1 displays the number of deaths by ethnic groups in both the 2018 Census and Mortality records. The groups Black Caribbeans, Black Maroon, and Black Mulatto have been aggregated under the category of Afro-descendants presented in Section 4.4.1. Although there are socio cultural differences within the Afro-descendant groups, a disaggregated analysis considering Black-Caribbean, Black-Maroon, and Black-Mulatto independently was not possible due to the significant variation in sample size among the groups, particularly when compared to White-Mestizos.The use of the category "Afro-descendant" thus captures the general aspects of racial discrimination towards the Afro-Colombian population. Although variations in discriminatory practices towards different Afro-descendant groups may exist, this general category encompasses the broader concept of the construction of "blackness" in the country, encapsulating all forms of discrimination against this ethnic group as a whole. It is also worth noting that although the Roma population may experience ethnic discrimination, this group was excluded from the study for similar reasons. First, it constitutes a relatively small population with a high mortality variance, requiring a more appropriate methodology than the one employed in this document. Second, the Roma population lacks a direct historical link to the colonial past of the country, leading to distinct social dynamics compared to Indigenous and Afro-descendant populations. The analysis of the Roma population is thus beyond the scope of the document.

Ethnic information for mortality records is reported by relatives of the deceased person and in the case of children, ethnicity is reported by the mother. The census, on the other hand, collects data on self-reported ethnicity of all living members of the households, but it does not ask about the ethnicity of the reported deaths (deceased). Thus, for this study it was assumed that the ethnicity of deceased members in households is the same as the ethnicity of the head household. In other words, the ethnicity of the head of household was taken as a proxy variable for the ethnicity of the deceased members. This assumption represents a low error level when considering that in departments with high multiethnic composition, such as Valle del Cauca, only 6% of households reported more than one ethnicity within the family group in the 2018 Census. This implies that only 4% of the population could be misclassified if the ethnicity of the deceased person and the ethnicity of the head of the household differ. These percentages are even lower in more conservative and less mixed regions, such as Antioquia, where only 2.5% of households reported having multiple ethnic identities, indicating that just 1.2% of the population could be ethnically misclassified. Generally, the majority of Colombian households are ethnically homogenous, suggesting that the ethnicity of one family member is a reliable predictor of the ethnicity of the rest of the family members.

Under-registered deaths by ethnic groups were estimated in two parts: first, we compared the number of deaths in the census with those in the mortality records to analyse the level of completeness of mortality registration data. We then produced descriptive statistics, survival curves, and mortality distribution maps. The second part of the analysis used multiple logistic regression models to estimate the relationship between under-registration of mortality and ethnicity. The logistic regression was estimated with a binary outcome: whether a person reported as having died in the preceding year was officially registered or not. We considered a selection of important but available explanatory variables including: ethnicity, gender and age of the deceased as well as the education, gender, and place of residence of the head of household. The head of household is often the person who report deaths in the household, and hence their characteristics are relevant for the regression models.

The existence of the Census microdata including a question on death registration means that we choose not to use established methods for the investigation of death registration completeness, such as the Brass Growth Balance Method (Dorrington, 2013a) or its extension, the Generalised Growth Balance Method (Dorrington, 2013b). This is because we are able to obtain more granular data with fewer assumptions by directly analysing self-reported under-registration [2].

Information about the place of residence included in the regression analysis indicates whether the deceased person resided in a capital municipality (cabecera municipal) or not, which can, in turn, affect the registration of deaths in the mortality records. These capital municipalities coordinate all administrative functions of the departments, thus representing better chances for accurate death registration. In Colombia, the National Administrative Department of Statistics (DANE) classifies the country into 33 political divisions known as departments, which are further subdivided into a total of 1,122 municipalities (municipios).

[2]In this context, we use the term 'self-reported' to refer to reporting on the part of the deceased's household

4.4 Results

4.4.1 Descriptive Statistics

Indigenous, Afro-descendants, and White-Mestizo groups are each spatially distributed across the country; but there are some geographic areas which show a clear racial segregation and have historically led to a geographically racialised country. This intersectionality of race and geography is presented in Figure 4.1, which highlights localities where more than 50% of the population belongs to a particular ethnic group. The Afro-descendants are mainly clustered along the Pacific and Atlantic coast, White-Mestizos populate the area around the centre of the country, and Indigenous groups dominate the south, close to the Colombian Amazon.

Figure 4.1: Colombian Population Distribution by Ethnic Groups

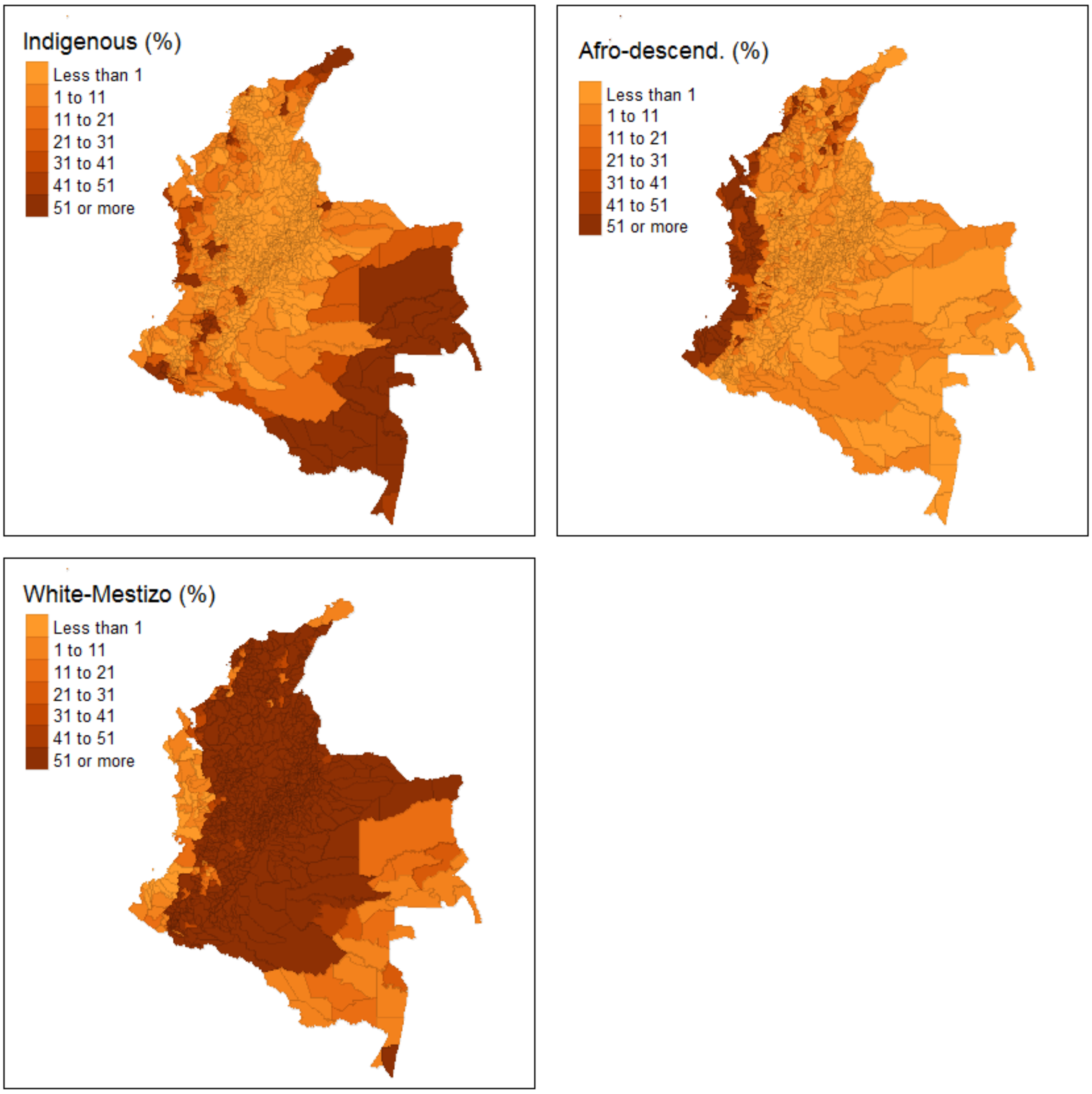


Figure 4.1 shows the Colombian geographical regions where a particular ethnic group is majority with more than 50% of the local population. Darker colours represent a higher concentration of population.
Source: Own elaboration based on 2018 Census.

The maps shown in Figure 4.2 illustrates how mortality registration interacts with race and geography in the Colombian context. The maps on the left present mortality rates estimated using the death records information, while those on the right present mortality rates based on census data. The results show that mortality levels measured by death records (maps A and B) are significantly lower than the mortality levels estimated by the census (maps C and D). Given it is unlikely that the Census is overestimating deaths, this is indicative of considerable under-registration of deaths in the Colombian registration system. In addition, death records captured information predominantly from White-Mestizo-dominated areas, leaving deaths in ethnic regions under-estimated.

Figure 4.2: Comparison of Mortality Rates from Death Records and Census Deaths

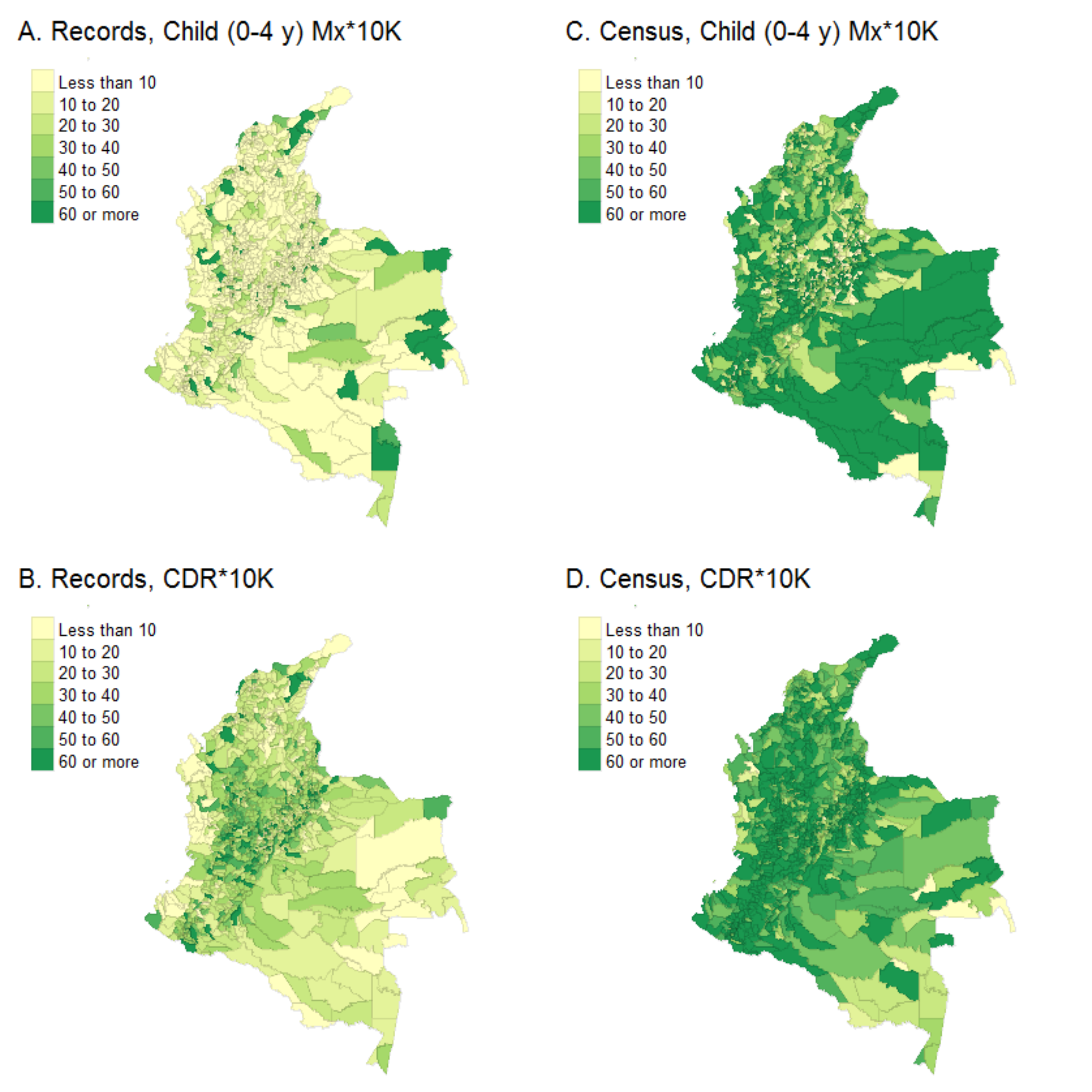
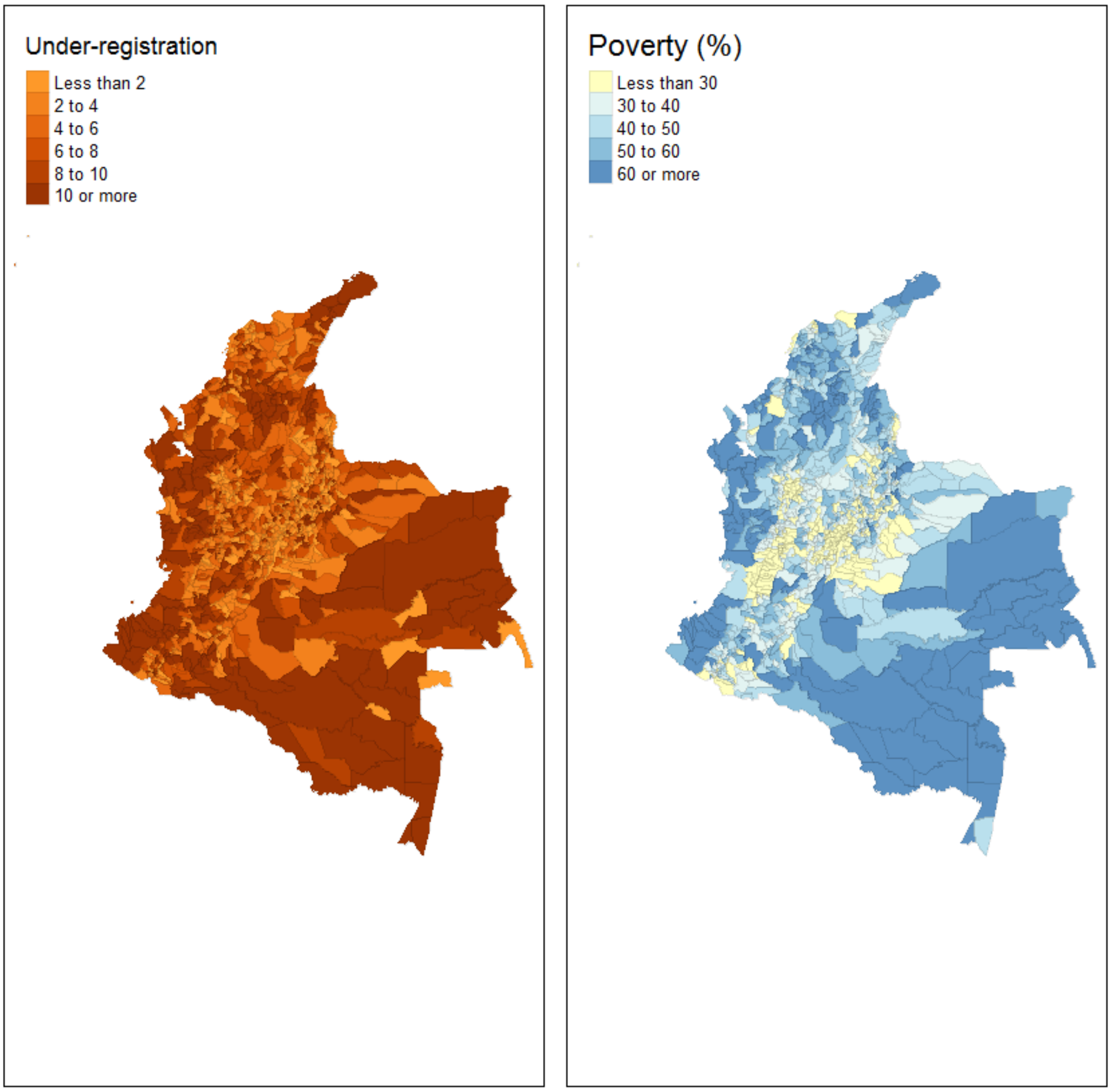


Figure 4.2 show the Colombian regions with the highest mortality under-registration, and how the registered number of deaths differs by source between death records and census information. On the top, age-specific mortality rates by municipalities for the age-group 0-4 are displayed, death records registration on the left, and census on the right. At the bottom, crude death rates (CDR) are displayed , death records on the left side, and census on the right. Data from mortality records and census from 2017.
Source: Own elaboration based on mortality registers and 2018 Census.

Death records in Colombia are biased in Indigenous and Afro-descendants regions, as the number of deaths registered are significantly lower than the number of deaths reported in the census. Municipalities dominated by Indigenous and Black populations that were identified as having high child mortality rates in the census (map C) were reported as areas with very low child mortality levels in the mortality records (map A). This bias is even clearer when the general mortality level is compared between the maps B and D, where mortality in the ethnic regions appear clearly under-estimated. The maps also show concentration of mortality registration in the central part of the country where the White-Mestizo population is located. In contrast, the periphery and the localities of ethnic population remain only partially covered by the registration system. Regions of residence of ethnic groups are also the poorest areas of the country and with the higher number of under-reported deaths (Figure 4.3).

The inconsistency in the number of deaths between death records and the census affects mostly ethnic communities. The census reported $242,744$ deaths and the death records $227,624$ in 2017, which corresponds to a difference of 6.2% at the national level. However, across ethnic groups, the proportion of unregistered deaths is significantly higher for the Indigenous and Black population. Table 4.2 compares the number of deaths from the mortality records versus the number of deaths from the census reports. Two findings draw attention: first, the differences between the two data sources are disproportionately higher in younger ages; second, for some age groups, the number of deaths captured in the census from Indigenous and Black population are four and five times higher respectively than the number that was registered in the death records. For instance, in the age groups < 1 and $1-4$, death records reported respectively 458 and 195 Indigenous deaths in the category Total. However, the census data show $2,052$ and $1,108$ deaths for the same age groups, which represents deaths four times higher at age < 1, and almost 6-fold more at ages $1-4$. Additionally, we must also consider that the census estimates may also underestimate the true number of deaths. First, the census of 2018 suffered from problems of undercount. The 2018 Census yielded biased results principally for Afro-descendant ethnicity due to the exclusion of approximately 1.5 million individuals from this group, particularly those residing in rural and impoverished regions. This omission could led to a minimisation of the observed differences between Afro-descendants and other ethnic groups. However, despite the underestimation, the inequality gap remains unmistakable. Second, the household self reporting measure relies on recall. Third, those living alone will not be captured by the question regarding deaths in the household. The difference between census death estimates and mortality registration data, referred to in the below as under-registration, is likely therefore to be a lower bound on the extent of the undercount.

Figure 4.3: Comparison of Regions with High Mortality Under-registration and High Multidimensional Poverty



The maps in Figure 4.3 illustrate the intersection of mortality under-registration and poverty in regions predominantly inhabited by Afro-descendants and Indigenous people. The map on the left displays the level of mortality under-registration per 10,000 inhabitants in 2017, estimated by comparing the deaths reported in the Census to those in the vital registration records. The map on the right shows the percentage of households experiencing multidimensional poverty (MP) for the same period. MP is an indicator of households facing multiple deprivations in areas such as health, education, and living standards (Alkire et al., 2018; Departamento-Administrativo-Nacional-De-Estadística, 2017). Darker colours indicate a higher concentration of the respective variable.

Source: Own elaboration based on mortality registers and poverty measures from the National Statistics Office, DANE.

Table 4.2: Number of Deaths from Death Records and Census by Ethnicity, Gender and Age, 2017

Age group	Death records									Census deaths								
	Indigenous			Afro-d.			White-M.			Indigenous			Afro-d.			White-M.		
	M	F	Total	M	F	Total	M	F	Total	M	F	Total	M	F	Total	M	F	Total
<1	253	205	**458**	243	178	**421**	3,480	2647	**6,127**	1,181	868	**2,052**	662	539	**1,203**	5,510	4,229	**9,742**
1-4	111	84	**195**	43	39	**82**	708	504	**1,212**	645	463	**1,108**	321	240	**561**	1,864	1,427	**3,291**
5-9	27	13	**40**	24	18	**42**	455	324	**779**	177	136	**313**	133	88	**222**	688	494	**1,148**
10-14	24	26	**50**	27	25	**52**	579	403	**982**	149	108	**257**	122	113	**235**	748	529	**1,278**
15-19	54	52	**106**	106	47	**153**	2,683	716	**3,399**	343	258	**601**	414	335	**751**	2,339	1,671	**4,012**
20-24	79	37	**116**	160	56	**216**	4,217	957	**5,174**	330	295	**628**	631	455	**1,086**	3,479	2,618	**6,100**
25-29	67	43	**110**	155	75	**230**	4,000	1,072	**5,072**	281	231	**512**	535	422	**957**	3,254	2,430	**5,689**
30-34	55	34	**89**	155	91	**246**	3,597	1,195	**4,792**	280	201	**481**	436	361	**797**	3,075	2,313	**5,393**
35-39	55	44	**99**	139	111	**250**	3,295	1,514	**4,809**	289	212	**503**	403	307	**710**	2,956	2,264	**5,226**
40-44	52	35	**87**	161	122	**283**	3,077	1,800	**4,877**	205	170	**375**	383	277	**660**	2,921	2,243	**5,171**
45-49	49	50	**99**	201	167	**368**	3780	2510	**6,290**	301	245	**546**	411	330	**741**	3,615	2,790	**6,410**
50-54	51	43	**94**	226	229	**455**	5,084	3,514	**8,598**	332	213	**545**	579	399	**980**	5,197	3,968	**9,172**
55-59	62	55	**117**	323	275	**598**	6,551	4,726	**11,277**	314	275	**589**	536	445	**981**	6,069	4,560	**10,634**
60-64	80	54	**134**	367	317	**684**	8,350	6,032	**14,382**	438	335	**773**	721	589	**1,312**	7,937	6,188	**14,134**
65-69	96	67	**163**	415	368	**783**	9,752	7,153	**16,905**	354	293	**648**	728	544	**1,272**	8,543	6,519	**15,072**
70-74	124	96	**220**	450	344	**794**	10,968	8,549	**19,517**	483	411	**894**	760	617	**1,378**	10,006	7,657	**17,678**
75-79	143	109	**252**	581	500	**1,081**	13,113	11,227	**24,340**	460	393	**853**	902	709	**1,611**	11,286	8,717	**20,018**
80-84	144	146	**290**	573	575	**1,148**	13,233	13,411	**26,644**	659	450	**1,109**	991	725	**1,716**	13,019	9,977	**23,017**
85+	231	279	**510**	751	977	**1,728**	20,885	27,608	**48,493**	1,007	746	**1,753**	1,679	1,314	**2,994**	23,675	17,781	**41,484**
Total	1,757	1,472	**3,229**	5,100	4,514	**9,614**	11,7807	95,862	**213,669**	8,228	6,303	**14,540**	11,347	8,809	**20,167**	116,181	88,375	**204,705**

The category "Total" represents the number of deaths by age group and includes cases of deceased people whose ethnicity and age was reported, but the gender did not. That means, that the values of the column "Total" are not always the result of the sum of identified males (M) and females (F).
Source: Own elaboration based on mortality registers and 2018 Census.

The estimated proportion of unregistered deaths is calculated as the proportion of deaths from the census that were not registered by the death records. Following the previous example with the Indigenous population, death records reported only 458 from $2,052$ census deaths at age < 1. This means that at least 77.7% of cases were not registered. Table 4.3 shows the proportion of under-registration by age group and ethnicity. In total, 77.8% and 52.3% of deaths in Indigenous, and Afro-descendants respectively are not registered. In the case of the White-Mestizo group, the total number of deaths from death records outnumber the total deaths from the census at 4.5%. As noted above, this may be because the Census 2018 had an omission of $4,094,077$ persons according to official reports (Departamento-Administrativo-Nacional-De-Estadística, 2021; Urrea-Giraldo et al., 2020), which means that deaths from White-Mestizo people in areas and groups that were not covered well by the census may be captured in the mortality registration data. Failure by the census in capturing deaths to single person-households may also be a factor, particularly as this “over-registration” of the death records, however, happens particularly at ages 55 years and older.

The under-registration, in general, decreases with age. In the White-Mestizo group, under-registration decreases after age 30 onwards and becomes positive at age 55 (indicating ’over-registration’ in death records relative to the census). In other ethnic groups, this pattern of decreasing under-registration with age is not as clear. Under-registration of deaths decreases in Afro-descendants at older ages as well, but this trend is more pronounced in White-Mestizos than in Blacks; and in Indigenous communities, the level of under-registration remain more or less at the same level across age groups. In general, the highest proportions of omitted deaths affect populations under 10 years, where under-registration of deaths in Indigenous and Afro-descendant populations is over 80%, (see Table 4.3).

Paradoxically, under-registration at age < 1 is relatively lower than in other age groups. For example, 82.4% of Indigenous, 85.4% of Afro-descendants and 63.2% of White-Mestizo deaths at age $1-4$ remained unregistered in 2017. Meanwhile the percentages of under-registration at age < 1 were 77.6%, 64.9%, and 37.1% respectively. Under-registration of deaths among children who died before reaching the first birthday is lower than for those who died after age 1. Figure 4.4 presents these patterns more clearly.

The under-registration of deaths in ethnic populations present a major hurdle for getting reliable life expectancy estimations for these communities. Figure 4.5 shows survival estimations with the death records as well as with the census mortality information. Life expectancy based on mortality records is significantly longer than those from census data, as the survival curve from the census includes a higher number of deaths which means people are dying in higher proportions in all ages. Consider for example the impact that deaths at age < 1 have on the survivorship of Indigenous people, something which is not captured by the survival curve from death records. On the other hand, the differences in the White-Mestizo groups are minimal and the curves present a cross over, which is the result of a mortality over-registration in the older age groups, as shown in Tables 4.2 and 4.3. In the Afro-descendants, the under-registration gap increases considerably after age 15,

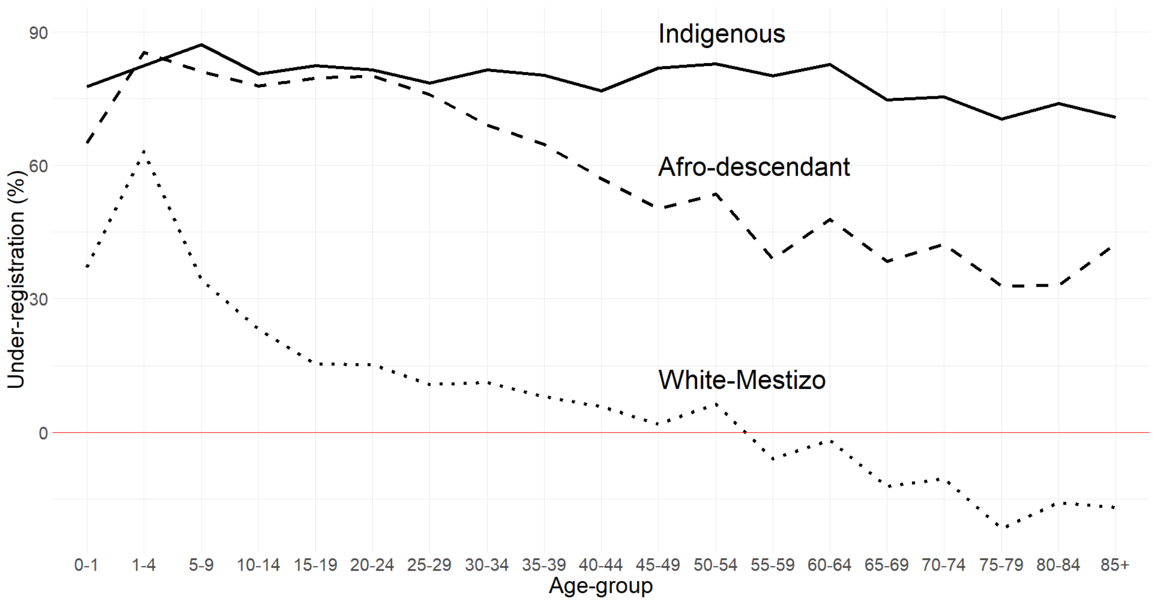
Table 4.3: Percentage of Registration Bias by Age Group and Ethnicity, Colombia 2017. Estimated through Comparison of Census and Registration Data.

	Registration bias								
	Indigenous (%)			Afro-descendants (%)			White-Mestizo (%)		
Age-group	M	F	Total	M	F	Total	M	F	Total
<1	-78.6	-76.4	-77.6	-63.3	-67.0	-64.9	-36.8	-37.4	-37.1
1-4	-82.8	-81.9	-82.4	-86.6	-83.8	-85.4	-62.0	-64.7	-63.2
5-9	-84.7	-90.4	-87.2	-82.0	-79.5	-81.0	-33.9	-34.4	-34.1
10-14	-83.9	-75.9	-80.5	-77.9	-77.9	-77.9	-22.6	-23.8	-23.1
15-19	-84.3	-79.8	-82.4	-74.4	-86.0	-79.6	+14.7	-57.2	-15.2
20-24	-76.1	-87.5	-81.4	-74.6	-87.7	-80.1	+21.2	-63.4	-15.1
25-29	-76.2	-81.4	-78.5	-71.0	-82.2	-76.0	+22.9	-55.9	-10.8
30-34	-80.4	-83.1	-81.5	-64.4	-74.8	-69.1	+17.0	-48.3	-11.1
35-39	-81.0	-79.2	-80.2	-65.5	-63.8	-64.8	+11.5	-33.1	-7.9
40-44	-74.6	-79.4	-76.8	-58.0	-56.0	-57.1	+5.3	-19.8	-5.6
45-49	-83.7	-79.6	-81.9	-51.1	-49.4	-50.3	+4.6	-10.0	-1.8
50-54	-84.6	-79.8	-82.8	-61.0	-42.6	-53.5	-2.2	-11.4	-6.2
55-59	-80.3	-80.0	-80.1	-39.7	-38.2	-39.0	+7.9	+3.6	+6.1
60-64	-81.7	-83.9	-82.7	-49.1	-46.2	-47.8	+5.2	-2.5	+1.8
65-69	-72.9	-77.1	-74.8	-43.0	-32.4	-38.4	+14.2	+9.7	+12.2
70-74	-74.3	-76.6	-75.4	-40.8	-44.2	-42.3	+9.6	+11.6	+10.5
75-79	-68.9	-72.3	-70.5	-35.6	-29.5	-32.9	+16.2	+28.8	+21.7
80-84	-78.1	-67.6	-73.9	-42.2	-20.7	-33.1	+1.6	+34.4	+15.9
85+	-77.1	-62.6	-70.9	-55.3	-25.6	-42.3	-11.8	+55.3	+17.0
Total	-78.6	-76.6	-77.8	-55.1	-48.8	-52.3	+1.4	+8.5	+4.5

Registration bias in mortality records relative to the census is denoted by a minus sign and over-registration with a plus sign.
Source: Own elaboration based on mortality registers and 2018 Census.

and similar to the White-Mestizo, it decreases for the older ages. In general, the under-registration gap is wider in Indigenous and Black population than in White-Mestizos.

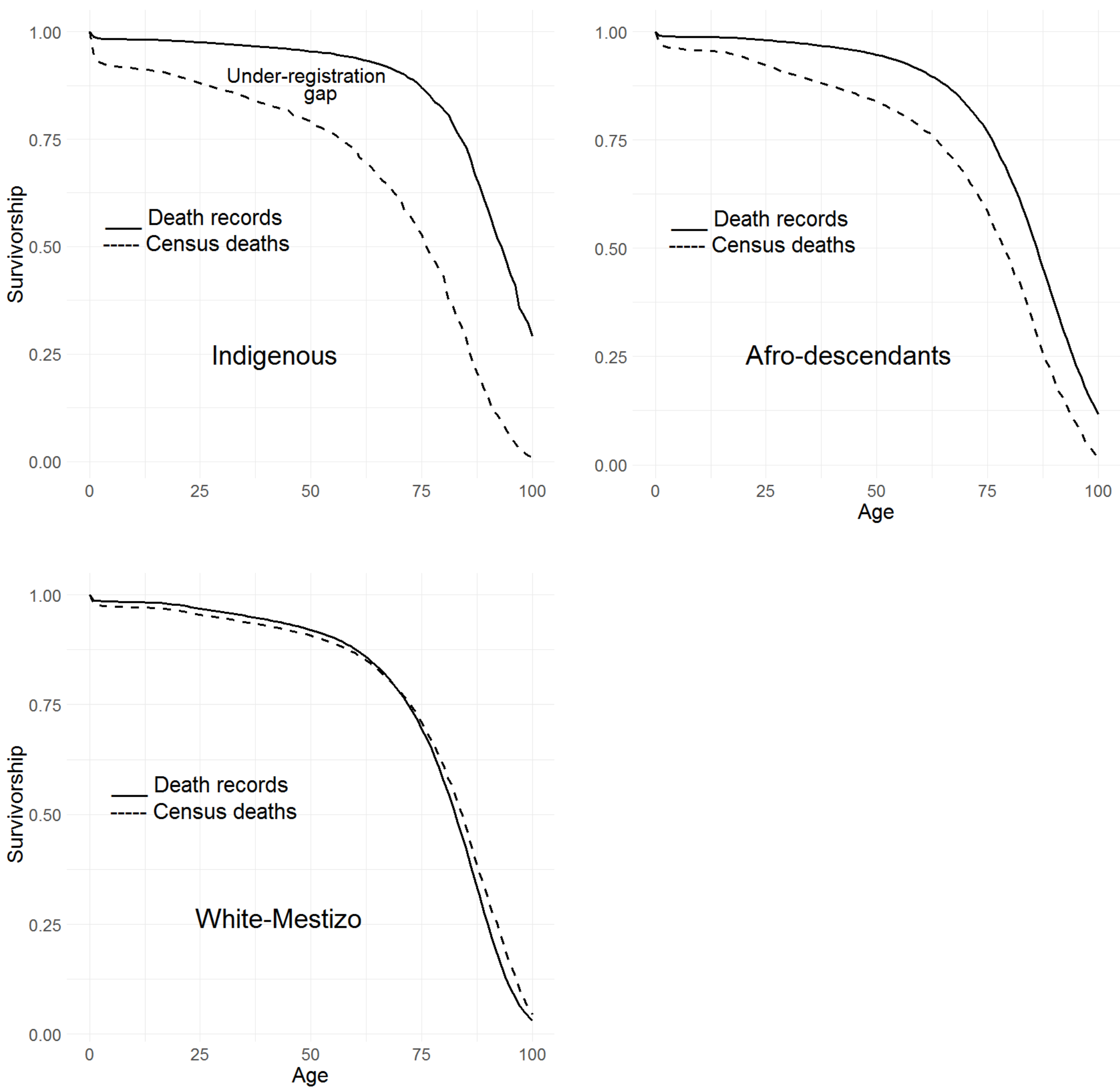
Figure 4.4: Mortality Registration Ratio Between 2018 Census and Mortality Records by Age Group and Ethnicity



Source: Own elaboration based on mortality registers and 2018 Census.

Figure 4.4 presents the ratio of mortality registration between census and mortality registration. It compares registration by age groups and ethnic group. the level of registration between the two data sources is equal when the difference is 0%. A number higher than zero means that the number of registered deaths in the census is higher than those registered in the mortality records, and vice-versa. The results confirm that deaths from census exceed the number of registered deaths at all ages and ethnic groups, except for deaths at older ages in White-Mestizos. The difference in the under-registration ratio tends however to decrease with age. Ratio under-registration numbers lower than zero in the White-Mestizo groups means that the number of deaths registered in the death records were higher than the number of death reported in the Census 2018, signifying possible over representation of deaths in the older age groups. The patterns also reveal a common problem of age heaping and digit preference at older ages in which the slightly increased mortality at ages 50, 60, 70 and 80 is the result of the tendency to round the ages of older people with numbers ending in zero and five when the exact age of death is not available (Barrett, 2019; Manish, 2017; Agrawal and Kandhuja, 2015; Preston et al., 1996).

Figure 4.5: Survival Curves Based on Death Records and Census Deaths Information



Source: Own elaboration based on mortality registers and 2018 Census.

Figure 4.5 presents the survivorship curve based on life tables in which mortality rates (m_x) were estimated with the number of deaths from the census and from the mortality records for each ethnic group. The curves represent thus the survival function (l_x), or the fraction of survivors to exact age x. The population for the denominators of the estimations was derived from the 2018 Census for each ethnic group. The curves show that death records over-estimate the lifespan particularly in Indigenous and Afro-descendants due to unregistered deaths. The difference between the two registration methods is an estimate of the size of the under-registration gap.

To validate the consistency of the data, we analysed the correlation between death records and census deaths at the municipality level. Figure 4.6 shows in the first panel the scatter plot of the logarithm of census deaths reported as non-registered against the logarithm of death records, where each point in the plot corresponds to one municipality. The second panel similarly displays the log of deaths reported as *registered* in the census against the log of death records. One would expect that deaths reported in the census as non-registered would be only weakly correlated with those in the mortality records if such deaths were truly not registered. On the contrary, a strong correlation is expected if the deaths were registered in both systems: death records as well as the census.

The estimations confirm the assumptions: non-registered deaths and mortality records have a weaker correlation compared to registered deaths, This suggests that census methodology captures and measures effectively the under-registration patterns in vital statistics in Colombia. Additionally, the proportion of deaths reported as non-registered in the Census 2018 decreases over ages in all ethnic groups, matching what we find from comparing census deaths to mortality records, see Table 4.4.

Figure 4.6: Scatterplot of Death Records and Census Registered and Unregistered Deaths

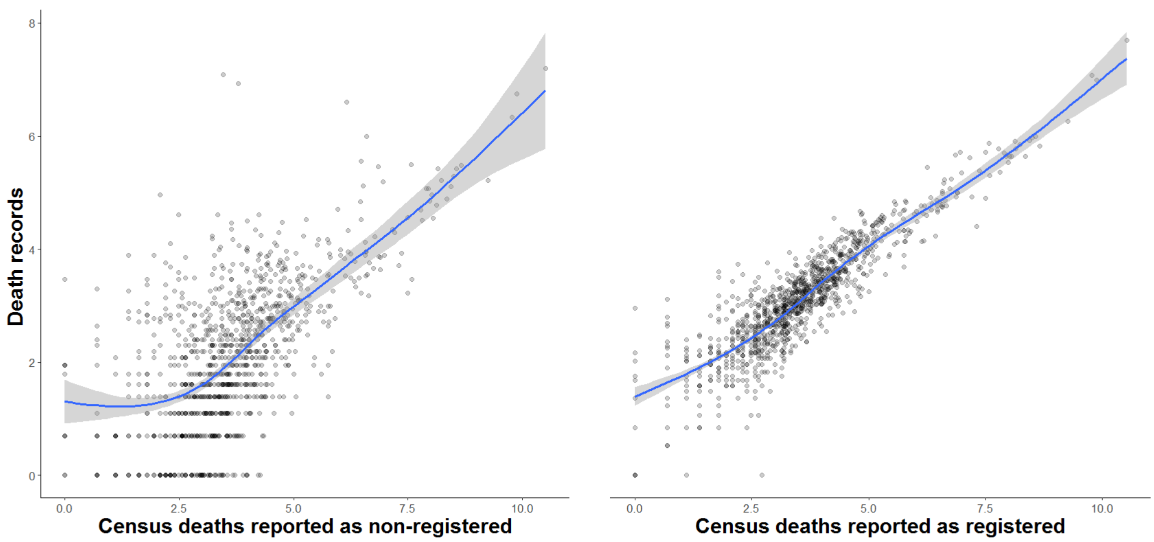


Figure 4.6 presents a local regression (LOESS) analysis. This non-parametric technique highlights the patterns in the relationship between death records and both registered and non-registered deaths from the census. The figure illustrates that data points from death records and registered deaths from the census exhibit lower dispersion and reduced uncertainty when fitting the data in the neighborhood of each point, in comparison to the relationship between death records and non-registered deaths. This suggests that registered deaths and reported deaths from the census are likely derived from the same sources. Conversely, the more dispersed relationship between death records and non-registered deaths from the census indicates a divergence in data sources.
Source: Own elaboration based on mortality registers and 2018 Census.

Figure 4.6 presents two scatter plots displaying, first, the natural logarithm of deaths from the mortality records plotted against the natural logarithm of census deaths reported as non-registered deaths. The right-hand plots logged deaths from mortality records against logged deaths from census reported as registered deaths. In this latter plot, if all deaths reported as registered in the census were indeed registered, and similarly all registered deaths were captured in the census, then the census and registration counts of registered deaths would be equal. Therefore the points on the plot would lie on the diagonal line $x = y$, as the coordinates of each point represent deaths that correspond to the same location and time point.

In the plot itself, a Locally Estimated Scatterplot Smoothing (LOESS) curve is displayed. Although there are fewer deaths reported in the death records, as can be seen from comparing the x- and y-scales, the census and death registration counts are highly correlated. Additionally, the fitted line is approximately linear on the log scale, indicating that if there is some over-reporting of registration in the census, it does not seem to vary by the size of the area (that is, by number of deaths reported).

Similarly, in the left-hand plot, the correlation between non-registered deaths reported in the census and actual death registrations figures should be much less strong, and points should thus be more dispersed, as is the case. In both plots, confidence intervals indicate the uncertainty in the estimation of the LOESS trend line. The figure as a whole illustrates that deaths reported as registered in the census are closely correlated with mortality records, while non-registered deaths are not. This to some extent validates information about mortality registration and under-registration from the census.

4.4.2 Regression Analysis

Comparison between mortality registration records and census mortality is difficult because of under-registration in both sources. To understand better factors that may influence under-registration, and to understand how this differs by ethnic group, we can focus solely on census information about whether household deaths are registered.

Information from the census regarding death registration was used for the logit binary model in order to measure the association of ethnicity with death under-registration. It measures the probability of not reporting deaths given that the deceased person is Indigenous, Afro-descendant, or White-Mestizo (using the ethnicity of the household head as a proxy for that of the deceased person as mentioned previously). The model also considers age and sex of the deceased person as well as age, sex, place of residence, and years of education of the head of household (HH) who is the person responsible for the death registration. Table 4.4 shows the data on percentages of non-registered deaths (N-Re), registered deaths (Re), and deaths with no information about registration (NA) based on the 2018 Census.

The level of mortality under-registration determined by information from reporters who admitted not having registered a deceased person, as shown in Figure 4.4, is lower than the level of

Table 4.4: Percentage of Death Registration and Under-Registration by Ethnicity According to the Census 2018, Colombia

	Indigenous (%)			Afro-descendants (%)			White-Mestizo (%)		
Age-group	N-Re	Re	NA	N-Re	Re	NA	N-Re	Re	NA
< 1	56.8	15.6	27.6	36.2	35.0	28.7	27.1	35.3	37.6
1-4	56.6	19.4	24.0	29.5	49.4	21.1	24.8	50.9	24.2
5-9	51.3	29.2	19.6	19.2	57.2	23.6	16.9	62.8	20.3
10-14	40.7	38.3	20.9	18.3	65.2	16.5	13.1	73.4	13.5
15-19	50.5	36.5	13.0	14.4	74.4	11.2	9.8	80.8	9.4
20-24	38.8	46.1	15.1	14.8	72.8	12.4	9.1	80.9	10.0
25-29	30.8	52.8	16.3	15.3	69.8	14.8	9.1	80.8	10.1
30-34	44.6	41.0	14.4	15.7	70.1	14.2	8.5	81.6	9.9
35-39	41.5	45.2	13.3	12.6	73.9	13.5	9.1	81.1	9.8
40-44	26.7	54.3	19.0	11.3	75.4	13.3	8.0	81.9	10.1
45-49	41.4	45.1	13.5	10.8	75.7	13.5	7.2	83.5	9.3
50-54	34.1	53.2	12.6	11.7	76.6	11.7	7.3	83.4	9.3
55-59	34.1	53.7	12.2	13.2	74.5	12.3	6.6	84.6	8.8
60-64	34.7	51.0	14.3	11.7	75.8	12.5	5.8	85.7	8.5
65-69	26.5	58.9	14.7	9.3	78.8	12.0	5.7	86.5	7.8
70-74	28.9	56.2	14.9	10.0	75.2	14.8	5.4	86.7	7.9
75-79	21.2	65.0	13.8	10.4	78.9	10.7	5.3	87.5	7.2
80-84	26.3	58.9	14.8	11.0	77.4	11.7	5.3	87.3	7.4
85+	20.6	66.9	12.5	9.5	78.4	12.2	5.1	87.4	7.5
Total	37.4	45.6	17.0	13.7	72.4	13.9	7.6	82.5	10.0

The 2018 Census collected information from households about reported and non-reported deaths. The answers were classified by: non registered (N-Re), registered (Re), and unknown-non answer (NA). Table 4.4 shows how census deaths are distributed across these categories by ethnicity.
Source: Own elaboration based on mortality registers and 2018 Census.

mortality under-registration determined by comparing deaths captured in the census with deaths from registers, as presented in Table 4.3. However, the under-registration patterns, specifically the decreasing trend over age and the higher under-estimation of deaths among Indigenous and Afro-descendant populations, are consistent across both data sources. This consistency is clearly illustrated in Figure 4.7.

Table 4.5 presents the estimated coefficients for a series of nested models, each containing a different (but overlapping) set of independent variables. Model 6, the model with the most extensive set of explanatory variables, is the best model presented in this table according to AIC and BIC selection criteria. This model includes a term with an interaction between age and ethnicity, which improves the model performance with respect to these criteria. Coeffients for this model are not included in the table, but illustrative predicted probabilities of non-registration from this model by age and ethnicity are presented in Figure 4.8.

Figure 4.7: Mortality Under-registration in Colombia by Age Group and Ethnicity from the Census 2018

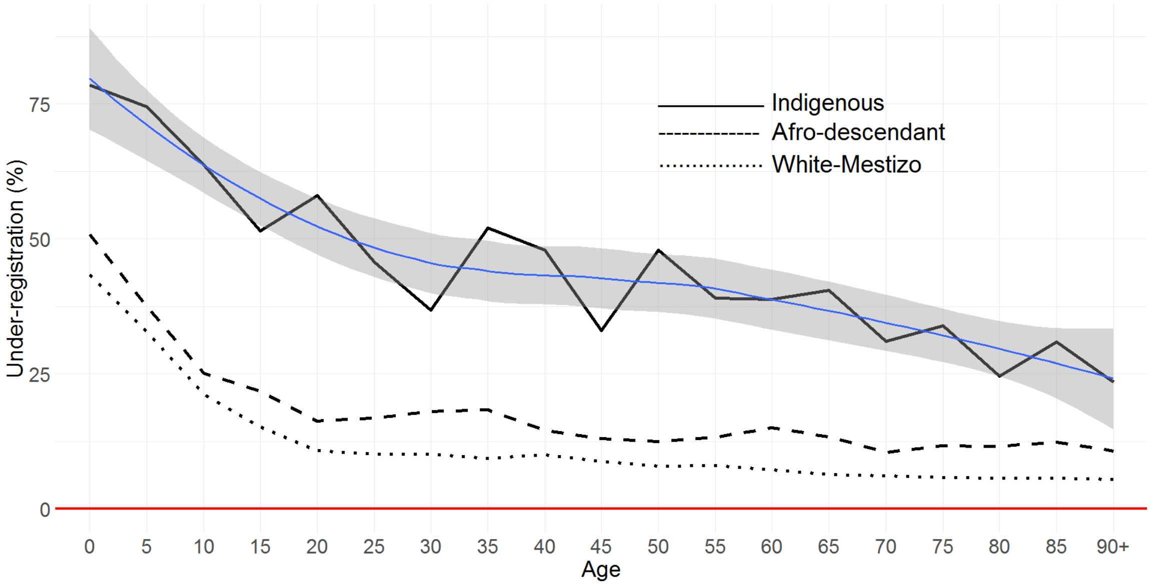


Figure 4.7 shows percentages of registered deaths from the total reported deaths in the census. Mortality under-registration is higher at younger ages and in ethnic groups. Indigenous deaths present the highest level of under-registration at all ages, while the difference between White-Mestizos and Afro-descendant is relatively small, with the lowest level for White-Mestizos. The data from Indigenous groups underwent adjustment through LOESS regression to mitigate age heaping and substantial variance resulting from the relatively limited sample size in this group. The regression plotted line confirms that the trend of under-registration decreases with age among Indigenous populations as well. The confidence interval reflects the variability in the estimation of the LOESS trend line.
Source: Own elaboration based on mortality registers.

The results confirm previous findings: the odds of being unregistered for Afro-descendants and Indigenous groups is higher compared with White-Mestizo people. In the case of age of death, the highest risk of non-registration occurs when the deceased is a newborn or a child under 1 year old. Thereafter, the risk starts decreasing and achieves the lowest risk for people aged 75 years and more, when the odds decreased by 93% for White-Mestizos. This age effect is however different among ethnic groups. Figure 4.4 shows the effect tends to disappear in older ages of White-Mestizos and Afro-descendants, while remain practically unchanged in Indigenous groups.

Similarly, the age of the household head is correlated negatively to under-registration of deaths. This means regardless of the age of the deceased person, under-registration of deaths is higher in younger families, and household heads under the age of 25 present the highest risk of not reporting the death compared with any other age group. In contrast, when the head of household is 70 or older, the odds of being under-registered decreased by 25%. In terms of gender, deceased males have higher risk of being not registered than females: the odds of under-registration increase by 14% in males.

Table 4.5: Model Selection for Logit Models of the Relationship Between Ethnicity and Death Non-registration. Reported Effects Are Estimated Regression Coefficients

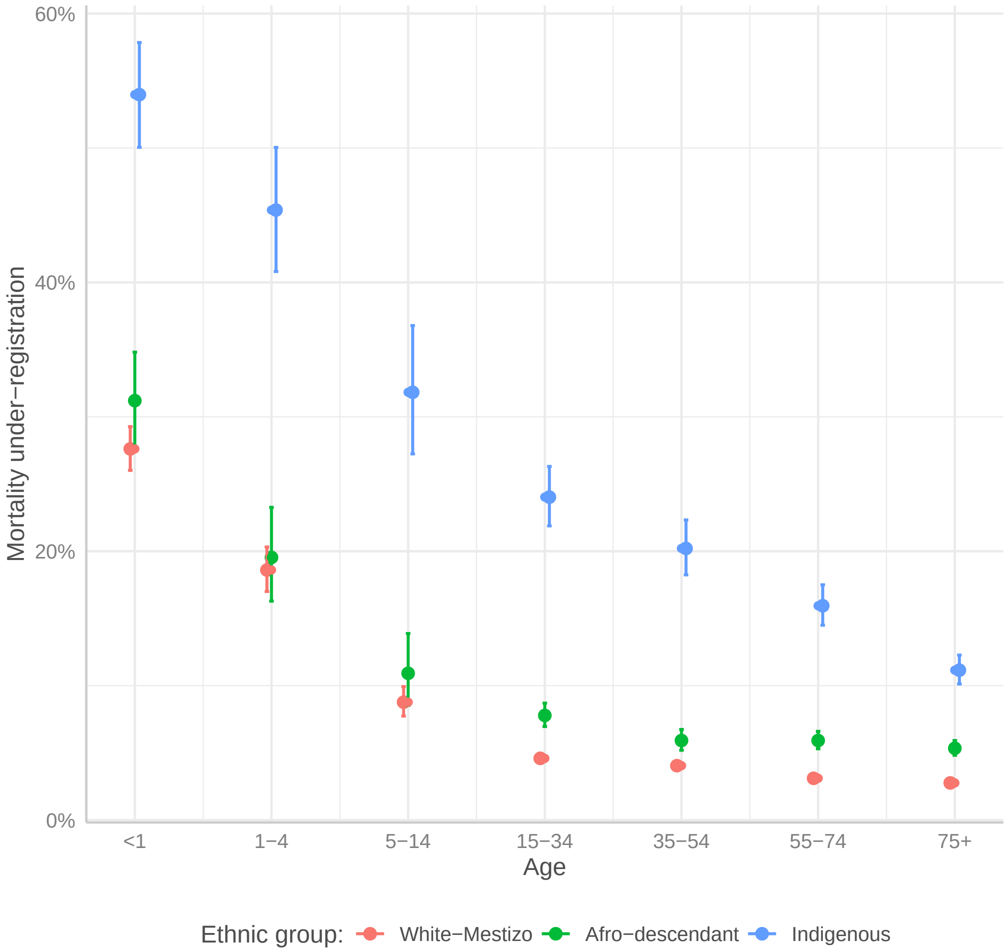
Predictor	Mod 1	Mod 2	Mod 3	Mod 4	Mod 5	Mod 6
Intercept	-2.40 (0.01) [3,4]	-0.32 (0.03)	-0.44 (0.03)	-0.91 (0.04)	-1.05 (0.04)	-0.96 (0.04)
Ethnicity: White-Mestizo = Reference category						
Afro-descendant	0.72 (0.02)	0.60 (0.03)	0.60 (0.03)	0.56 (0.03)	0.52 (0.03)	0.17 (0.08)
Indigenous	2.07 (0.02)	1.85 (0.02)	1.86 (0.02)	1.70 (0.02)	1.63 (0.03)	1.12 (0.08)
Age deceased: <1 year = Reference category						
1-4		-0.49 (0.05)	-0.49 (0.05)	-0.55 (0.05)	-0.54 (0.05)	-0.51 (0.05)
5-14		-1.25 (0.06)	-1.24 (0.06)	-1.33 (0.06)	-1.34 (0.06)	-1.38 (0.07)
15-34		-1.84 (0.03)	-1.86 (0.03)	-1.93 (0.03)	-1.93 (0.03)	-2.07 (0.04)
35-54		-2.05 (0.03)	-2.06 (0.03)	-2.09 (0.03)	-2.10 (0.03)	-2.20 (0.04)
55-74		-2.34 (0.03)	-2.35 (0.03)	-2.35 (0.03)	-2.35 (0.03)	-2.48 (0.04)
75+		-2.52 (0.03)	-2.52 (0.03)	-2.50 (0.03)	-2.51 (0.03)	-2.60 (0.03)
Sex deceased: Female = Reference category						
Male			0.13 (0.02)	0.11 (0.02)	0.11 (0.02)	0.11 (0.02)
Age HH: <25 years = Reference category						
25-49			-0.10 (0.02)	-0.09 (0.02)	-0.09 (0.02)	-0.09 (0.02)
50-69			-0.06 (0.02)	-0.27 (0.02)	-0.26 (0.02)	-0.26 (0.02)
70+			0.06 (0.02)	-0.31 (0.03)	-0.28 (0.03)	-0.28 (0.03)
Sex HH: Female = Reference category						
Male			0.18 (0.02)	0.14 (0.02)	0.13 (0.02)	0.13 (0.02)
Education HH: >12 years						
6 to 12				0.45 (0.03)	0.40 (0.03)	0.40 (0.03)
< 6				1.04 (0.03)	0.95 (0.03)	0.96 (0.03)
Residence HH Capital City = Reference category						
Non-capital city					0.31 (0.02)	0.32 (0.02)
Interactions term Ethnicity * Age [5]						
N	195263	195263	195263	195263	195263	195263
Deviance	123889	114718	114502	112602	112294	112147
AIC	123895	114736	114530	112634	112328	112205
BIC	123925	114827	114672	112797	112501	112500

Table 4.5 presents the model selection. Model 6 (Mod6) is the model with the best criteria in Deviance, AIC, and BIC. The models included an interaction effect between age and ethnicity which is presented in Figure 4.8 for model 6. Source: Own elaboration based on Colombian Census 2018 data.

The results also indicate a clear negative association between education and under-registration. The higher the educational attainment of the household head, the lower the risk that a death remains unregistered. The reference category is household head with more than 12 years of education at technical and university levels, with coefficients for those between 6 and 12 years at secondary school level and those with less than 6 years at primary school. The odds of non-registration increases by 50% and 160% if the head of household had between 6–12 years of schooling and less than 6 years of education respectively, compared with the reference category. The place of residence plays an important role for death registration. Households outside of capital cities had 37% higher odds of non-registration compared with those located in the main city of the department.

In general, the models highlight the specific sociodemographic characteristics of the deceased person, as well as characteristics of households, where death under-registration is more likely. The highest predicted probability of under-registration of deaths are confined to Indigenous newborns who died in families where the head of household is a young male under age 25, less educated, and living outside the main city of the department.

Figure 4.8: Predicted Probability of Mortality Non-registration by Ethnicity and Age - Interaction Model (Mod 6)



Predictions are made at the reference category for those categorical values not included in the interaction. The interaction effect in Figure 4.8 between ethnicity and age reveals under-registration trends consistent with the patterns presented in Figures 4.4 and 4.7, indicating higher under-registration of deaths among younger age groups in Indigenous and Afro-descendant populations. Figure 4.8 demonstrates that differences in under-registration by age are more pronounced in Indigenous groups compared to other ethnic groups, while differences between White-Mestizos and Afro-descendants are relatively minor. In summary, the percentages of the odds in the interaction term indicate that the largest disparities are not only between ethnic groups but also among age groups within the same ethnicity. These differences are particularly significant when comparing under-registration of deaths at young ages across different ethnicities.
Source: Own elaboration based on Colombian Census 2018 data.

4.5 Discussion

The foregoing analyses clearly demonstrate geospatial patterning in under-reporting of death, which correlates with the geographic distribution of ethnic minority populations. Furthermore, Census microdata provides evidence of relatively higher level of under-reporting of deaths amongst Indigenous and Afro-descendants when compared to the White-Mestizo population, even after adjusting for selected household and geographic factors.

In Colombia, ethnic minorities face cultural barriers, social discrimination, geographical segregation and racial discrimination, and these may have contributed to under-reporting of deaths. More explicitly, cultural realities such as the lack of an institutionalised social environment and hence fewer requirements for death certificates; language restrictions for communities for whom Spanish is not the first language; and geographic isolation may partly explain the reasons for non-registration of deaths, as reported elsewhere (Tin Oung et al., 2017; Duryea et al., 2006); Alongside this, the high level of use of traditional and folk medical practices in rural and ethnic groups imply that many deaths occur outside health institutions and are thus unlikely to be registered or reported (Ronan, 2014; Boyd et al., 2000) Meanwhile, under-registration of deaths tend to be more accentuated among the Indigenous, followed by Afro-descendants and the White-Mestizo population, most likely attributed to the higher presence of the latter in urban areas and main cities. According to the 2018 Census, only 21% of the Indigenous population reside in the main city of the departments (cabecera municipal) whereas the corresponding figure for Afro-descendants is 66.7% (Departamento-Administrativo-Nacional-De-Estadística, 2019a,b).

On the other hand, the result showing significant difference in under-registration of deaths between Afro-descendants and Indigenous populations but only modest difference between Afro-descendants and White-Mestizos could be as a consequence of the omission of more than a million Afro-descendants from poor areas in the 2018 census (Departamento-Administrativo-Nacional-De-Estadística, 2021; Urrea-Giraldo et al., 2020). Census data may thus yield biased results, which could possibly underestimate the true level of mortality among Afro-descendants. Censal omission could be related to likely higher representation of deaths in older age groups of White-Mestizos. It is likely that the deaths not captured in the census may have been recorded elsewhere in official registration sources. We find more deaths at ages 55 and above among White-Mestizos, possibly attributed to greater motivation for registering deaths of older people due to inheritance claims and related legal requirements.

A second explanation for the over-representation of White-Mestizo population is the phenomenon of *blanqueamiento* (Whitening), that has been closely related to the construction of a national identity in many Latin American countries, particularly in Brazil and Colombia (Wade, 1993b). The *blanqueamiento* process refers to not only the biological racial aspects, but also cultural homogenization that envisages a future in which blackness and indigenousness are not only absorbed but also erased from the national panorama, giving rise to a *whitened* mestizo nation (Wade, 1993b).

Several studies have suggested that census interviewers tend to whiten the racial classification of higher educated and financially solvent non-white people (Kateri, 2013). In the 1991 Brazilian census, Afro-Brazilians organisations launched a national campaign Not let your color pass as white inviting non-white people to pay attention to racial misclassification (Andrews, 2016, p.19). This phenomenon of social *whitening* could explain the over-representation of white-Mestizo deaths in registration data, when collecting such information from the next of kin of the deceased was not possible as it is with the Census data. Moreover, it explains why over-registration appears in older age groups with better education and accumulated wealth.

Regional segregation of Indigenous and Afro-descendant communities in poor locations plays an overriding role in terms of access to economic resources and registration services. Segregation does not imply only the physical separation of ethnic groups, but other manifestations such as discrimination, social class, and privilege and cultural distinctiveness, and although the separation appears to be voluntary, this is generally not the case (Cell, 1982). On the other hand, there are other individual characteristics of the deceased person, as well as the characteristics of the head of household, that may influence mortality registration. These findings should be interpreted not only indicating under-registration as an individual problem, but the registration patterns should be considered within their wider social and structural context.

Our analysis shows that under-registration is higher at young ages, as highlighted in previous studies (Tin Oung et al., 2017; Guo et al., 2015; Prasartkul and Vapattanawong, 2006; Duryea et al., 2006; Yang et al., 2005; Blakely et al., 2000; Lumbiganon et al., 1990). Deaths of older people are more likely to be registered by family members to fulfil legal compliance related to inheritance of assets such as land, property and money. In contrast, death certificates of newborns and infants are generally less useful for any legal purposes, and hence such deaths are likely to be under-reported or under-registered. Under-registration of deaths is more likely in households when the head of the household is below 50. This might be due to the fact that households with young members generally have less experience or awareness of registration processes. Further investigation is needed to better understand this phenomenon.

We find that male deaths are more prone to under-registration than those of females. However, this finding contradicts other studies (Guo et al., 2015; Prasartkul and Vapattanawong, 2006). In the Colombian context, a higher under-registration of male deaths could be determined by deaths due to internal conflicts and violence, especially in rural and areas with higher proportions of Afro-descendant and Indigeneous citizens. In some cities, for example, male deaths caused by homicides are 18 times higher than those of females (Franco, 1997, p.172). This implies that many of those non-registered male deaths could be the victims of conflict and violence, and whose deaths were sensitive enough not to be reported

Our study has some drawbacks and data limitations. First, the under-estimation of Census deaths in the Afro-descendant population is highly likely to minimise the magnitude of the differences in mortality when compared to other ethnic groups. Second, we do not have information about the causes of death of non-registered people, nor the effect of the conflict on mortality under-registration. Further research is required to understand the reasons for non-registration of deaths. Despite these limitations, this study provides robust evidence of the extent of under-registration gap among ethnic groups in Colombia. The study draws strength by utilising the specific Census question on whether the deaths reported in the household in the preceding year were registered or not. This allowed us to shed light on the individual correlates of mortality under-registration in Colombia. One of the key contributions of this study is the analysis of ethnic differentials in mortality under-registration, which has hitherto not been analysed systematically. The present study yields a critique of how official statistics can be influenced by social disadvantage, marginalisation and institutional discrimination against ethnic minority groups. It is important to note that race is a constructed social category, a politicised and a preconceived notion. The collection and analyses of census and registration data from primarily a White-Mestizo perspective could be affected by perceptions, prejudices and structural discrimination that leads to a bias in favour or detriment of one particular group. While the disparities in under-registration among ethnic groups can be partly elucidated by the socioeconomic status model outlined in Section 2.6, it is crucial to underscore that racism and ethnic discrimination are fundamental factors contributing to racial inequities. The administrative structure of the state, shaped by a system of privileges rooted in colonialism, perpetuates the marginalisation of ethnic groups and fosters disinterest in accurately registering and providing coverage for these populations. The real social conditions of Indigenous and Afro-descendant groups remain thus hidden due to lack of quality data and cross-validation tools. Finally, this study draws attention to the importance of viewing ethnic statistics more critically not only in terms of who is generating the information, but also how data were collected.

CHAPTER 5

Ethnic Mortality in Colombia 2008-2019

5.1 Introduction

The role of ethnicity in morbidity and mortality outcomes has been widely acknowledged in demographic and social epidemiological analysis. Racialised social groups present poorer health and higher mortality than the standard White reference category. Health and mortality inequalities across different ethnic groups have been documented in Europe (Blom et al., 2016), the UK(Chouhan and Nazroo, 2020; Wohland et al., 2015; Nazroo et al., 2006), in the US (Womack et al., 2020; Lariscy et al., 2020; Williams and Collins, 2001), Canada, New Zealand, Australia (Paradies, 2016; King et al., 2009; Bramley et al., 2004; Ajwani et al., 2003; Trovato, 2001), SouthAfrica, and Latin America (United-Nations, 2010; Giufrida et al., 2007; Montenegro and Stephens, 2006) ; and more recently, health inequalities by ethnic groups have been found during the Covid-19 pandemic (Nazroo and Becares, 2020; Gross et al., 2020; Platt and Warwick, 2020; Ogedegbe et al., 2020).

The above findings suggesting higher morbidity and mortality outcomes in ethnic minorities are consistent and universal, with few exceptions. The currently available evidence, while limited, suggests the same may be true for Colombia. Recent studies in specific Colombian cities have shown that Indigenous and Afro-descendants people have a higher risk of death than the White-Mestizo population (Spijker et al., 2020; Palacios, 2018; Agudelo-Suárez et al., 2016; Urrea-Giraldo et al., 2015; Bernal and Cárdenas, 2007). Differences were found in the mortality rates for some specific age groups such as children and cause of death such as homicide in the city of Cali (Urrea-Giraldo et al., 2021; Spijker et al., 2020). In the light of aforementioned research challenges and gaps, this research analyses the trends in mortality and life expectancy differences between ethnic groups at a national level between 2008 and 2019. The analysis focuses particularly on the contribution of demographic rates by age and sex to underlying differences in life expectancy.

5.2 Data and Methods

5.2.1 Data

Data for the analysis are drawn from different datasets: number of deaths reported in the Census 2018, number of deaths registered in mortality records from 2008 to 2019, and information regarding population size from 2005 and 2018 Censuses. Data for the present analysis were provided by the Colombian National Statistics Office (DANE).

Mortality records from 2008 to 2019 contain $2,554,013$ deaths, of which $38,722$ were reported as Indigenous, $132,018$ as Afro-descendant, and $2,229,237$ as White-Mestizo ethnicity. Information on ethnic population from Census 2005 and 2018 is presented in Table 5.1. Population size and number of deaths increased for Indigenous and White-Mestizos groups from 2005 to 2018, but in the case of the Afro-descendant group, the population size decreased during the intercensal period due to census coverage problems that primarily affected Afro-descendants[1] The study did not include the census omission into the correction method, primarily due to the lack of definitive information regarding the age structure of the omitted population, the number of affected households, and the potential number of deaths that could have been reported by the omitted population. The estimation of mortality rates was therefore based on population size information from Census 2005, as it is more reliable when it comes to including ethnic groups and minimises the bias of the population exposure.

Table 5.1: Population Size and Reported Deaths by Ethnicity from Census 2005 and 2018

Ethnicity	Census 2005				Census 2018			
	Population		Deaths		Population		Deaths	
	Male	Female	Male	Female	Male	Female	Male	Female
Indigenous	703,046	689,577	2,061	1,636	951,215	954,402	8,182	6,358
Afro-descendants	2,143,675	2,168,082	10,325	7,165	1,454,135	1,528,089	11,647	8,520
White-Mestizo	17,051,381	17,846,790	88,326	61,627	18,833,585	19,844,756	115,947	88,758

Source: Own elaboration based on 2005 and 2018 Censuses.

5.2.2 Method

The results presented in Chapter 4 above showed that the level of completeness in Colombian mortality records varys across ethnic groups, geographical regions, age, and sex. Different meth-

[1]According to the Colombian National Statistics, $4,094,077$ persons were not counted in the National Population Census 2018, https://www.dane.gov.co/index.php/estadisticas-por-tema/demografia-y-poblacion/censo-nacional-de-poblacion-y-vivenda-2018/cuantos-somos, from which around 1.6 million belong to Afro-descendants communities.

ods have been considered in order to correct underestimation problems (Derrington, 2013a; Dorrington, 2013b; Derrington, 2013b; Bennet and Horiuchi, 1984); however, most of them do not consider the case of estimating completeness for heterogeneous population subgroups (other than the standard demographic groupings of age and sex). Additionally, the assumptions upon which such methods rely may be questionable in this case, and requirements regarding data on migration are difficult in this context, as information on migration is difficult to obtain at the required granularity. Linkage methods needed for capture-recapture methods are also difficult to apply as we have limited variables on which to match. Methods based on model life tables assume particular age-specific patterns of mortality, e.g. Glei et al. (2021) rely on good estimates of ${}_5q_0$ and ${}_{45}q_{15}$, and make assumptions regarding age patterns of mortality that may result in loss of information given in the registration data.

Yet in spite of the above, an approximation to the completeness of Colombian mortality records can be made from the census question about how many members of the household died during the year previous to the census, and whether the death was registered or not. Based on this information, a logistic regression model for the probability of registering deaths p can be estimated, using age x, sex s, and ethnicity e as covariates. The model specification used in this case and the resulting estimates can be found in Chapter 4 Assuming that under-registration comes from a population in which the under-registration rate remains constant each year conditional on ethnicity, sex and age group, then observed deaths D follow a binomial distribution, with mean μ equal to the true number of deaths T multiplied by the probability of reporting p. This is $\mu = pT$. Conversely, for a given number of observed deaths D, the number of unobserved deaths U has a negative binomial distribution with mean $\mu* = D(1-p)/p$. The expected true number of deaths is therefore equal to the number of observed deaths (D) plus the expected number of unobserved deaths (U), and it is given by:

$$
\begin{aligned}
T &= D + U \\
T &= D + D(\frac{1-p}{p}) \\
T &= D\frac{p}{p} + D(\frac{1-p}{p}) \\
T &= D\frac{p + (1-p)}{p} \\
T &= \frac{D}{p}
\end{aligned}
\tag{5.2.1}
$$

Equation 5.2.1 together with 5.2.2 below show that an estimate of the true number of deaths by specific age-group x and ethnicity e, $\hat{T}_{xe}$, can be obtained by dividing the observed deaths $\hat{D}_{xe}$ by the estimated probability of reporting $\hat{p}_{xe}$, where the latter is generated by the regression model presented in the previous chapter.

$$\hat{T}_{xe} = \frac{D_{xe}}{\hat{p}_{xe}} \tag{5.2.2}$$

Equation 5.2.2 thus enables the production of corrected estimates of the death counts in the mortality records for the period $2008 - 2019$ by age, and ethnicity. Population exposures to generate mortality rate estimations were taken from 2005 Census, as this census presents more complete information of ethnic population compared to the 2018 Census, and population size information by ethnicity and year are not available. Therefore, correcting the registered deaths by ethnicity and using the population size from a previous census will give us a baseline estimation of the true mortality rates by ethnicity. In addition to this, we used information from 2018 Census to estimate life tables and life expectancies by ethnic groups for the year 2018. We assume that censal omission in this census affected reported deaths and population size at the same proportion, thus life tables estimation remains still reliable as the numerator and denominator for mortality rates were affected at the same proportion.

In order to understand better patterns of mortality variation, we used a method of decomposition to identify the major contributors to differences in life expectancy (e_0) and lifespan disparity ($e^{\dagger}$) among ethnic groups. The method of decomposition used here is the one proposed by Horiuchi et al. (2008), in which the difference in the dependent variable is expressed as a sum of the effects in its covariates. The explanation below closely follows that given in the above-referenced paper. This means that the total change in life expectancy at birth between two time points or two different populations can be understood as the sum of the changes of life expectancy across all ages during the same time space. For example, the total difference in life expectancy between any two populations, say, population 1 and population 2, will be equal to the sum of the differences at age $0, 1, 2..$ until age n. "The method is based on two assumptions, first: the values of the covariates change continuously along the analysed dimension, and this justify the additivity of covariate effects. Second, it also assumes that changes in the covariates occur gradually and they are proportional to one another" (Horiuchi et al., 2008, 786).

The model of decomposition assumes a numerical population characteristic y which is a differentiable function (f) of n covariates $\mathbf{X} = x_1, x_2, ..., x_n$. In our case, the population characteristic is represented by the life expectancy or life disparity and the covariates are the causes of death and age of death. Let us assume that both y and the x_i depend on time t. Suppose now that information from y and $\mathbf{X}$ is available at two time points t_1 and t_2, and $\mathbf{X}$ is a differentiable vector of covariates between t_1 and t_2. Horiuchi et al. (2008, p787) proposes the Decomposition Model as following:

$$y(t) = f(\mathbf{X}(t)) = f(x_1(t), x_2(t), ..., x_n(t))$$

and the total change between t_1 and t_2 is determined by:

$$y(t_2) - y(t_1) = \int_{t_1}^{t_2} \frac{d}{dt} y(t)dt$$

Applying the chain rule for partial derivatives:

$$y(t_2) - y(t_1) = \int_{t_1}^{t_2} \left\{ \sum_{i=1}^{n} \frac{\partial}{\partial x_i(t)} y(t). \frac{d}{dt} x_i(t) \right\} dt \tag{5.2.3}$$

Exchanging the integration and the summation, and applying the substitution rule of definite integrals:

$$y(t_2) - y(t_1) = \sum_{i=1}^{n} \int_{x_i(t_1)}^{x_i(t_2)} \frac{\partial}{\partial x_i(t)} y(t) dx_i(t)$$

Omitting the temporal dimension of the variables, the difference in y between t_1 and t_2 can be defined as:

$$y_2 - y_1 = \sum_{i=1}^{n} c_i, \text{ where } c_i = \int_{x_{i1}}^{x_{i2}} \frac{\partial y}{\partial x_i} dx_i \tag{5.2.4}$$

Equations 5.2.3 and 5.2.4, proposed by Horiuchi et al. (2008), show that the total change in the dependent variable can be expressed as a sum of effects of the covariates. "This specification is known as *line integral and differs from the notion of the effect of the coefficients in the linear regression model as the later measures the change in y produced by a unit change in x_i, that is $\partial y / \partial x_i$; while the decomposition line integral captures the effect of a change in y produced by a particular change in x_i between t_1 and t_2 measured by $c_i = \int_{x_{i1}}^{x_{i2}} \frac{\partial y}{\partial x_i} dx_i$ (Horiuchi et al., 2008, 788)"*.

Another indicator that we estimated in order to understand ethnic mortality differences is the life disparity, $e^\dagger$, understood as the average years of life expectancy lost due to death (Vaupel et al., 2011; Zhang and Vaupel, 2009; Vaupel and Canudas-Romo, 2003). There are a wide range of similar indices designed to measure the level of dispersion or entropy of the age of death, but they are all closely related (van Raalte and Caswell, 2013; Vaupel et al., 2011; Anson, 2002; Wilmoth and Horiuchi, 1999). Higher values on these indices indicate a great degree of variation in the age of death of the population; lifespans are more dissimilar to each other in such populations than in places where the index gives small values. We use here the life disparity index as it is a life table-based index, defined in Vaupel et al. (2011) as:

$$e^\dagger = \int_0^w e(a,t)d(a,t)da \text{ , with} \tag{5.2.5}$$

$$e(a,t) = \frac{\int_a^w l(x,t)dx}{l(a,t)}$$

$$l(a,t) = \exp\left(-\int_0^a \mu(x,t)dx\right) \text{ , and}$$

$$d(a,t) = l(a,t)\mu(x,t)$$

Where e represents the remaining life expectancy at age a at time t; and $l(a,t)$ is the probability of survival at age a, while $\mu(a,t)$ gives the age-specific hazard of death also known as *the force of mortality*. Therefore, a value of $e^\dagger = 20$ means that on average people in society are dying 20 years earlier than their life expectancy at age of death. Life disparity is defined as the average remaining life expectancy at the ages when death occurs; it measures the life years lost due to death (Vaupel et al., 2011, p.2).

Equations 5.2.3 and 5.2.4 decompose the life expectancy gap between two time points, y_1 and y_2, to identify the age groups contributing most to the changes in life expectancy. However, the methodology proposed here decomposes the differences between two ethnic groups rather than over a period of time, aiming to elucidate the sources of these differences. Similarly, Equation 5.2.5 will estimate the within-group dispersion of age at death. A high dispersion signifies substantial disparity within a specific ethnic group, indicating that while some individuals achieve the expected lifespan, a significant proportion die much earlier.

5.3 Results

5.3.1 Mortality Differences in Census Data

Life tables for Indigenous, Afro-descendant, and white-Mestizo groups are presented in Tables 5.2, 5.3, and 5.4 respectively. They show the differences in terms of survival probabilities and life expectancy at exact ages among ethnic groups. Indigenous and Afro-descendant population show higher probability of death compared to the White-Mestizo. For example, life expectancy at birth for Indigenous people is 66.7 years, for Afro-descendants 71.5, and in comparison White-Mestizo group recorded a life expectancy at birth of 78.9 years. This means that White-Mestizo people live on average 12.2, and 7.4 years longer than Indigenous and Afro-descendants respectively. The column of survivals lx of the life table shows the higher survival probabilities for White-Mestizos. Based on the proportion of survivors by age, about 10% of Indigenous cohort has already died before age 20, Afro-descendants before age 35, and White-Mestizos before age 55.

The speed at which the different cohorts become extinct is much faster for Indigenous and Afro-descendant groups when compared to White-Mestizos. For instance, the probability of a person surviving to age 5 is 0.92; 0.96; and 0.97 for Indigenous, Afro-descendants, and White-Mestizo ethnicity respectively. In the same order, the probability to survive to age 40 is 0.83; 0.87; and 0.93.

Table 5.2: Indigenous Life Table Based on Population and Deaths Information from Census 2018

Age_x	qx			l_x			d_x			L_x			T_x			e_0		
	Male	Female	Both	Male	Female	Both	Male	Female	Both	Male	Female	Both	Male	Female	Both	Male	Female	Both
0	0.063938	0.044567	**0.054512**	100,000	100,000	**100,000**	6,394	4,457	**5,451**	95,048	96,357	**95,661**	6,371,291	6,986,392	**6,668,966**	63.7	69.9	**66.7**
1-4	0.028188	0.023007	**0.025648**	93,606	95,543	**94,549**	2,639	2,198	**2,425**	368,174	376,707	**372,307**	6,276,243	6,890,035	**6,573,305**	67.0	72.1	**69.5**
5-9	0.008450	0.005483	**0.006999**	90,968	93,345	**92,124**	769	512	**645**	452,917	465,446	**459,007**	5,908,069	6,513,328	**6,200,997**	64.9	69.8	**67.3**
10-14	0.006424	0.005452	**0.005946**	90,199	92,833	**91,479**	579	506	**544**	449,546	462,901	**456,036**	5,455,153	6,047,882	**5,741,990**	60.5	65.1	**62.8**
15-19	0.015317	0.012869	**0.014107**	89,619	92,327	**90,935**	1,373	1,188	**1,283**	444,666	458,666	**451,469**	5,005,607	5,584,981	**5,285,954**	55.9	60.5	**58.1**
20-24	0.024705	0.009390	**0.017138**	88,247	91,139	**89,652**	2,180	856	**1,536**	435,783	453,556	**444,421**	4,560,941	5,126,315	**4,834,486**	51.7	56.2	**53.9**
25-29	0.024816	0.009415	**0.017108**	86,067	90,283	**88,116**	2,136	850	**1,508**	424,993	449,291	**436,811**	4,125,158	4,672,759	**4,390,065**	47.9	51.8	**49.8**
30-34	0.024519	0.012761	**0.018522**	83,931	89,433	**86,608**	2,058	1,141	**1,604**	414,509	444,313	**429,031**	3,700,164	4,223,468	**3,953,254**	44.1	47.2	**45.6**
35-39	0.023849	0.019451	**0.021579**	81,873	88,292	**85,004**	1,953	1,717	**1,834**	404,483	437,167	**420,435**	3,285,655	3,779,155	**3,524,223**	40.1	42.8	**41.5**
40-44	0.022370	0.015442	**0.018814**	79,920	86,575	**83,170**	1,788	1,337	**1,565**	395,132	429,531	**411,937**	2,881,172	3,341,988	**3,103,788**	36.1	38.6	**37.3**
45-49	0.039497	0.023431	**0.031285**	78,133	85,238	**81,605**	3,086	1,997	**2,553**	382,948	421,196	**401,643**	2,486,040	2,912,457	**2,691,851**	31.8	34.2	**33.0**
50-54	0.043495	0.029543	**0.036425**	75,047	83,241	**79,052**	3,264	2,459	**2,879**	367,072	410,055	**388,062**	2,103,093	2,491,262	**2,290,208**	28.0	29.9	**29.0**
55-59	0.051175	0.041837	**0.046441**	71,782	80,781	**76,173**	3,673	3,380	**3,538**	349,728	395,458	**372,019**	1,736,021	2,081,207	**1,902,147**	24.2	25.8	**25.0**
60-64	0.085650	0.077533	**0.081615**	68,109	77,402	**72,635**	5,833	6,001	**5,928**	325,960	372,006	**348,355**	1,386,293	1,685,749	**1,530,128**	20.4	21.8	**21.1**
65-69	0.089187	0.075664	**0.082351**	62,275	71,401	**66,707**	5,554	5,402	**5,493**	297,491	343,497	**319,801**	1,060,332	1,313,743	**1,181,773**	17.0	18.4	**17.7**
70-74	0.166113	0.127855	**0.146977**	56,721	65,998	**61,214**	9,422	8,438	**8,997**	260,051	308,895	**283,575**	762,841	970,247	**861,972**	13.4	14.7	**14.1**
75-79	0.202758	0.169863	**0.185762**	47,299	57,560	**52,217**	9,590	9,777	**9,700**	212,520	263,356	**236,833**	502,790	661,352	**578,396**	10.6	11.5	**11.1**
80-84	0.375584	0.329998	**0.351339**	37,709	47,783	**42,517**	14,163	15,768	**14,938**	153,137	199,493	**175,239**	290,270	397,995	**341,563**	7.7	8.3	**8.0**
85+	1.000000	1.000000	**1.000000**	23,546	32,014	**27,579**	23,546	32,014	**27,579**	137,133	198,503	**166,323**	137,133	198,503	**166,323**	5.8	6.2	**6.0**

The category Both represents the number of deaths by age group and includes cases of deceased people whose ethnicity and age was reported, but the gender was not. That means, that the values of the column Both are not always the result of the sum of identified males and females.

Source: Own elaboration based on the 2018 Census.

Table 5.3: Afro-descendants Life Table Based on Population and Deaths Information from Census 2018

Age_x	q_x			l_x			d_x			L_x			T_x			e_0		
	Male	Female	Both	Male	Female	Both	Male	Female	Both	Male	Female	Both	Male	Female	Both	Male	Female	Both
0	0.032496	0.022855	**0.027747**	100,000	100,000	**100,000**	3,250	2,285	**2,775**	97,188	97,985	**97,576**	6,738,217	7,575,129	**7,150,335**	67.4	75.8	**71.5**
1-4	0.013547	0.009587	**0.011614**	96,750	97,715	**97,225**	1,311	937	**1,129**	383,910	388,533	**386,169**	6,641,028	7,477,144	**7,052,759**	68.6	76.5	**72.5**
5-9	0.005135	0.003025	**0.004103**	95,440	96,778	**96,096**	490	293	**394**	475,974	483,157	**479,495**	6,257,118	7,088,610	**6,666,590**	65.6	73.2	**69.4**
10-14	0.004949	0.002892	**0.003946**	94,950	96,485	**95,702**	470	279	**378**	473,574	481,728	**477,565**	5,781,145	6,605,453	**6,187,095**	60.9	68.5	**64.6**
15-19	0.019742	0.005169	**0.012599**	94,480	96,206	**95,324**	1,865	497	**1,201**	467,736	479,787	**473,619**	5,307,571	6,123,726	**5,709,530**	56.2	63.7	**59.9**
20-24	0.033024	0.007090	**0.019974**	92,615	95,709	**94,123**	3,059	679	**1,880**	455,427	476,847	**465,916**	4,839,835	5,643,939	**5,235,911**	52.3	59.0	**55.6**
25-29	0.031151	0.007741	**0.019161**	89,556	95,030	**92,243**	2,790	736	**1,768**	440,806	473,311	**456,797**	4,384,408	5,167,092	**4,769,995**	49.0	54.4	**51.7**
30-34	0.027228	0.008648	**0.017598**	86,766	94,294	**90,476**	2,362	815	**1,592**	427,925	469,434	**448,398**	3,943,602	4,693,781	**4,313,197**	45.5	49.8	**47.7**
35-39	0.024496	0.010101	**0.016921**	84,404	93,479	**88,884**	2,068	944	**1,504**	416,850	465,035	**440,658**	3,515,677	4,224,347	**3,864,799**	41.7	45.2	**43.5**
40-44	0.025324	0.012353	**0.018411**	82,336	92,535	**87,380**	2,085	1,143	**1,609**	406,468	459,817	**432,876**	3,098,827	3,759,312	**3,424,141**	37.6	40.6	**39.2**
45-49	0.027906	0.016950	**0.022057**	80,251	91,392	**85,771**	2,240	1,549	**1,892**	395,657	453,086	**424,124**	2,692,359	3,299,496	**2,991,266**	33.5	36.1	**34.9**
50-54	0.036246	0.027396	**0.031555**	78,012	89,843	**83,879**	2,828	2,461	**2,647**	382,989	443,060	**412,777**	2,296,702	2,846,410	**2,567,142**	29.4	31.7	**30.6**
55-59	0.041679	0.032609	**0.036861**	75,184	87,381	**81,232**	3,134	2,849	**2,994**	368,086	429,783	**398,675**	1,913,713	2,403,350	**2,154,364**	25.5	27.5	**26.5**
60-64	0.072028	0.055896	**0.063518**	72,050	84,532	**78,238**	5,190	4,725	**4,970**	347,278	410,847	**378,765**	1,545,627	1,973,567	**1,755,690**	21.5	23.3	**22.4**
65-69	0.087872	0.078913	**0.083173**	66,861	79,807	**73,268**	5,875	6,298	**6,094**	319,616	383,290	**351,106**	1,198,349	1,562,719	**1,376,924**	17.9	19.6	**18.8**
70-74	0.149066	0.110717	**0.129024**	60,986	73,509	**67,174**	9,091	8,139	**8,667**	282,200	347,199	**314,204**	878,733	1,179,429	**1,025,818**	14.4	16.0	**15.3**
75-79	0.213223	0.174243	**0.192457**	51,895	65,370	**58,507**	11,065	11,390	**11,260**	231,810	298,376	**264,386**	596,533	832,230	**711,614**	11.5	12.7	**12.2**
80-84	0.317926	0.256637	**0.283589**	40,830	53,980	**47,247**	12,981	13,853	**13,399**	171,696	235,267	**202,738**	364,722	533,854	**447,229**	8.9	9.9	**9.5**
85+	1.000000	1.000000	**1.000000**	27,849	40,127	**33,848**	27,849	40,127	**33,848**	193,027	298,587	**244,490**	193,027	298,587	**244,490**	6.9	7.4	**7.2**

The category Both represents the number of deaths by age group and includes cases of deceased people whose ethnicity and age was reported, but the gender was not. That means, that the values of the column Both are not always the result of the sum of identified males and females.

Source: Own elaboration based on the 2018 Census.

Table 5.4: White-Mestizo Life Table Based on Population and Deaths Information from Census 2018

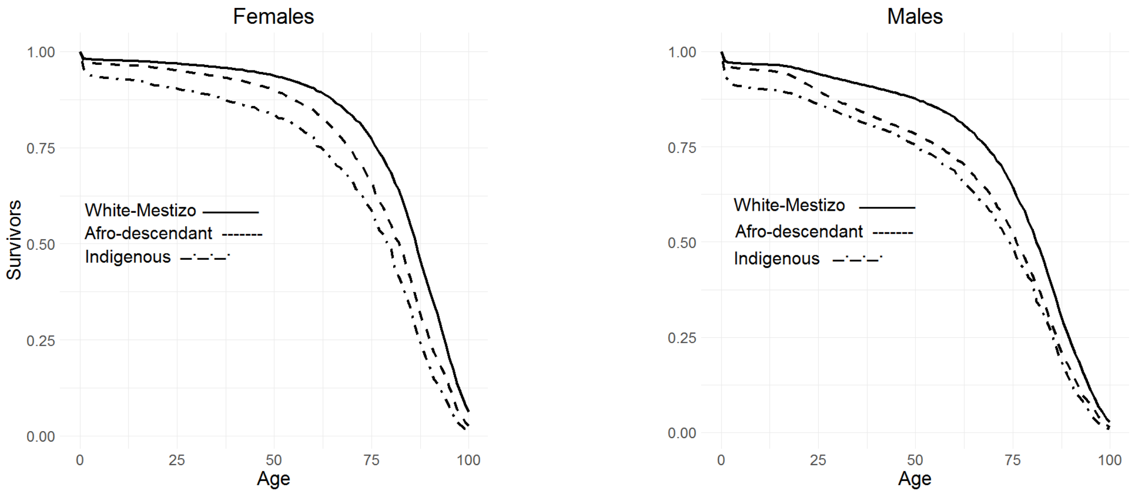
Age_x	qx			l_x			d_x			L_x			T_x			e_0		
	Male	Female	Both	Male	Female	Both	Male	Female	Both	Male	Female	Both	Male	Female	Both	Male	Female	Both
0	0.023992	0.016038	**0.020124**	100,000	100,000	**100,000**	2,399	1,604	**2,012**	97,867	98,554	**98,198**	7,512,169	8,264,301	**7,886,888**	75.1	82.6	**78.9**
1-4	0.007223	0.005302	**0.006287**	97,601	98,396	**97,988**	705	522	**616**	388,744	392,291	**390,461**	7,414,302	8,165,747	**7,788,690**	76.0	83.0	**79.5**
5-9	0.002365	0.001830	**0.002104**	96,896	97,874	**97,372**	229	179	**205**	483,906	488,925	**486,346**	7,025,559	7,773,456	**7,398,229**	72.5	79.4	**76.0**
10-14	0.002432	0.001718	**0.002084**	96,667	97,695	**97,167**	235	168	**202**	482,746	488,057	**485,327**	6,541,652	7,284,531	**6,911,883**	67.7	74.6	**71.1**
15-19	0.009288	0.002825	**0.006112**	96,432	97,528	**96,964**	896	275	**593**	479,919	486,949	**483,339**	6,058,907	6,796,474	**6,426,556**	62.8	69.7	**66.3**
20-24	0.014053	0.003719	**0.008905**	95,536	97,252	**96,372**	1,343	362	**858**	474,323	485,356	**479,712**	5,578,988	6,309,525	**5,943,216**	58.4	64.9	**61.7**
25-29	0.013472	0.004042	**0.008741**	94,193	96,890	**95,513**	1,269	392	**835**	467,794	483,473	**475,479**	5,104,665	5,824,169	**5,463,504**	54.2	60.1	**57.2**
30-34	0.013482	0.004833	**0.009102**	92,924	96,499	**94,678**	1,253	466	**862**	461,490	481,328	**471,238**	4,636,871	5,340,696	**4,988,024**	49.9	55.3	**52.7**
35-39	0.012777	0.005678	**0.009129**	91,672	96,032	**93,817**	1,171	545	**856**	455,429	478,799	**466,942**	4,175,381	4,859,368	**4,516,787**	45.5	50.6	**48.1**
40-44	0.014135	0.007326	**0.010576**	90,500	95,487	**92,960**	1,279	699	**983**	449,303	475,687	**462,343**	3,719,951	4,380,570	**4,049,845**	41.1	45.9	**43.6**
45-49	0.017153	0.010243	**0.013497**	89,221	94,788	**91,977**	1,530	971	**1,241**	442,279	471,511	**456,782**	3,270,648	3,904,883	**3,587,501**	36.7	41.2	**39.0**
50-54	0.024544	0.014967	**0.019445**	87,691	93,817	**90,736**	2,152	1,404	**1,764**	433,073	465,573	**449,267**	2,828,369	3,433,372	**3,130,720**	32.3	36.6	**34.5**
55-59	0.031951	0.019944	**0.025510**	85,538	92,413	**88,971**	2,733	1,843	**2,270**	420,860	457,455	**439,182**	2,395,296	2,967,799	**2,681,453**	28.0	32.1	**30.1**
60-64	0.051303	0.033502	**0.041707**	82,805	90,570	**86,702**	4,248	3,034	**3,616**	403,407	445,262	**424,468**	1,974,437	2,510,344	**2,242,271**	23.8	27.7	**25.9**
65-69	0.071053	0.046432	**0.057801**	78,557	87,535	**83,085**	5,582	4,064	**4,802**	378,832	427,515	**403,421**	1,571,030	2,065,082	**1,817,804**	20.0	23.6	**21.9**
70-74	0.114805	0.072402	**0.091876**	72,976	83,471	**78,283**	8,378	6,043	**7,192**	343,933	402,245	**373,435**	1,192,198	1,637,567	**1,414,382**	16.3	19.6	**18.1**
75-79	0.172224	0.113198	**0.139817**	64,598	77,427	**71,091**	11,125	8,765	**9,940**	295,175	365,225	**330,604**	848,265	1,235,321	**1,040,948**	13.1	16.0	**14.6**
80-84	0.274855	0.195524	**0.229758**	53,472	68,663	**61,151**	14,697	13,425	**14,050**	230,619	309,751	**270,630**	553,091	870,096	**710,343**	10.3	12.7	**11.6**
85+	1.000000	1.000000	**1.000000**	38,775	55,238	**47,101**	38,775	55,238	**47,101**	322,472	560,345	**439,713**	322,472	560,345	**439,713**	8.3	10.1	**9.3**

The category Both represents the number of deaths by age group and includes cases of deceased people whose ethnicity and age was reported, but the gender was not. That means, that the values of the column Both are not always the result of the sum of identified males and females.

Source: Own elaboration based on the 2018 Census.

Figure 5.1 illustrates the differences in lifespan among ethnic groups. The curves represent the *lx* values of the life table. White-Mestizos males and females have the longest lifespan and Indigenous people have the shortest. For females, the differences are mainly determined by the higher mortality of children under 5, particularly amongst Indigenous groups. The differences show an increase until age 50. Overall, White-Mestizo have better survival chances compared to Afro-descendant and Indigenous groups. The gap, however, tends to converge in older ages where the differences are smaller. For males, the differences are determined by the child mortality as well, and additionally, there is evidence of mortality increase in Afro-descendant people at age 15, which in turn generates a big impact on the lifespan differences when compared to White-Mestizos.

Figure 5.1: Survival Curves by Ethnicity and Gender



Survivals curves show that White-Mestizo males and females have a longer lifespan compared to Afro-descendant and Indigenous people.
Source: Own elaboration based on the 2018 Census.

The decomposition of life-expectancy and life-disparity differences show the age at which these differences start building up and the age-groups with the highest contribution. Figures 5.2 and 5.3 show the changes in life expectancy (e_0), and life disparity ($e_0^\dagger$) respectively. The biggest contributors to life-expectancy inequalities are the groups at age 0 and 1. As shown on the left side of Figure 5.2, people at age 0 contribute to more than 0.5 years to the overall 7.4 years life expectancy differences between White-Mestizo and Afro-descendant groups. Deaths at age 25 and 75 contribute likewise with more than 0.1 years to life expectancy inequalities. Comparing Indigenous with White-Mestizos, differences at age 0 constitute more than 2.5 years of the total 12.2 years life expectancy gap between the two groups. Children under age 5 are the largest contributors to the differences, while contributions from the rest of the groups remain more or less constant around 0.2 years. The peaks at ages 40, 50, 60, 70, and 80 are attributed to potential age heaping problems.

Figure 5.2: Decomposition of Life Expectancy (LE) Differences. Age Groups Where White-Mestizo Accumulates Life Expectancy Advantages Relative to Afro-descendants and Indigenous Groups

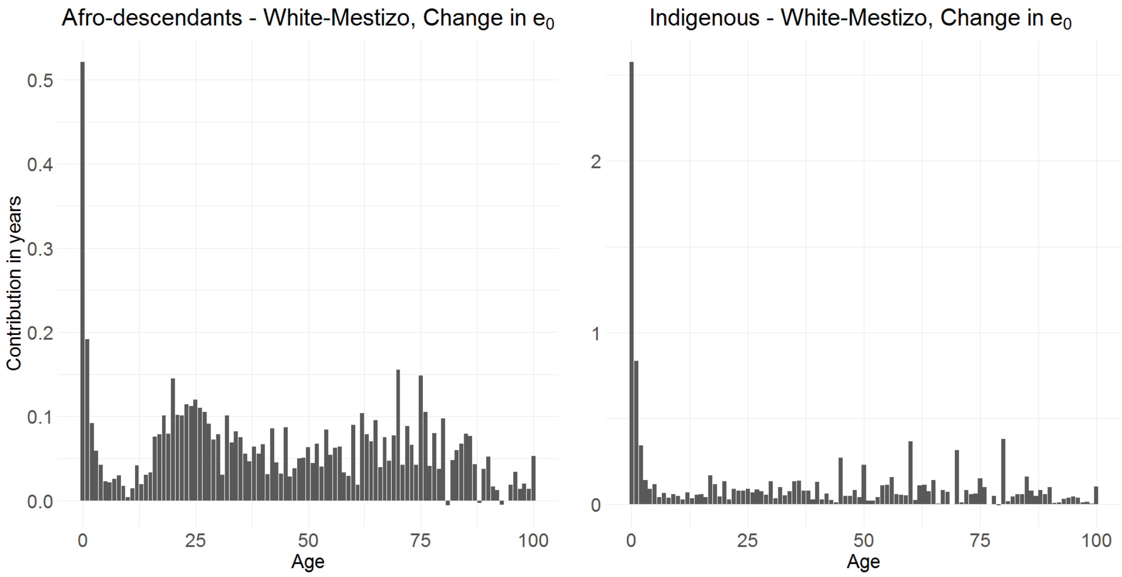


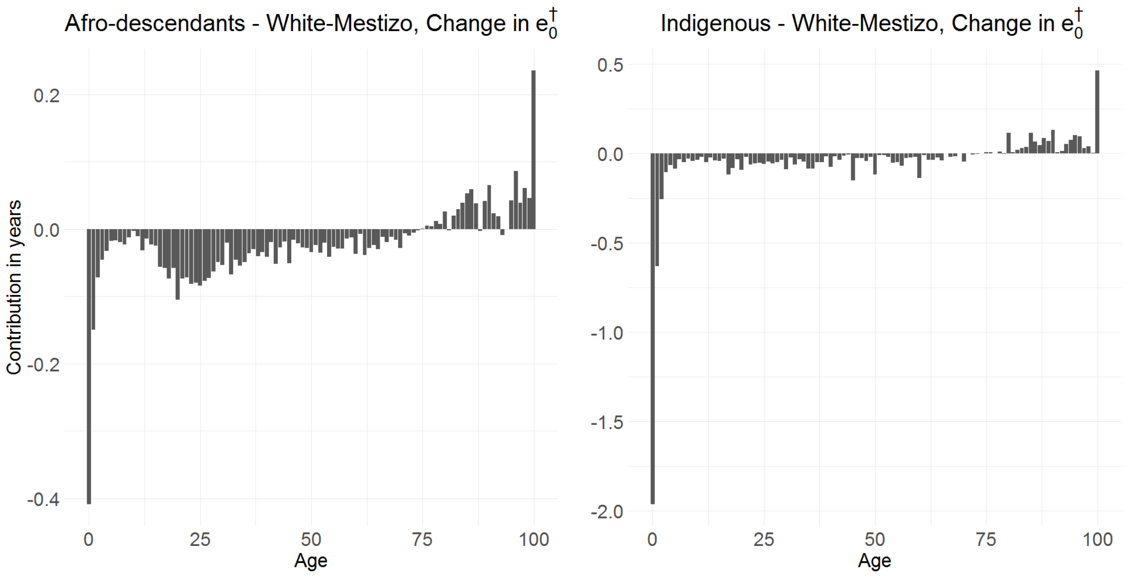
Figure 5.2 shows the age groups which most contribute to life expectancy differences. The figure on the left presents how many years each age group contributes to the life expectancy gap between Afro-descendants and White-Mestizos. The figure on the right shows the age groups' contributions to the life expectancy differences between Indigenous and White-Mestizo groups.

Source: Own elaboration based on the 2018 Census.

Life disparity estimations yield on average 14.8; 17.1; and 19.2 years for White-Mestizos, Afro-descendants, and Indigenous people respectively. This means that on average a White Mestizo person dies 14.8 years earlier than it was expected they should have lived at the age at which they died, while an Indigenous will die 19.2 years earlier or will lose 19.2 years of life on average upon death. For males, the corresponding values are 16; 19.1; and 20.5 years, and for females 13.2, 14.8, and 17.7 years respectively. Looking at how the differences in life disparity across ethnic groups decompose, Figure 5.3 illustrates why White-Mestizos present lower dispersion of the age of death when compared with other ethnic groups. Life disparity index is a measure of entropy and it will be equal to 0 if all deaths were concentrated at a particular age. White-Mestizos present lower disparity compared with Afro-descendants principally due to the contributions of children under 5 and young adults around age 25 (Figure 5.3). This suggests White-Mestizos have a lower proportion of deaths at those ages which decreases life disparity, as more people get closer to their expected age of death. Likewise, they present an increase in disparity at older ages which means more people are outliving the expected population age of death, which represents a higher dispersion of the age of death, therefore any distancing from the expected age of death will increase the disparities among lifespans in the society. There is a threshold age, $a^{\dagger}$, below which mortality reductions decrease lifespan disparity, while reductions above the threshold increase disparity (Aburto et al., 2019, 2018; Zhang and Vaupel, 2009). For this population, the threshold age appears to be approximately 75.

Figure 5.3, right panel, shows that the lower disparity in White Mestizos compared to Indigenous is basically due to the lower mortality in children and newborns. Lower mortality at older ages in White-Mestizos increases lifespan disparity relative to the indigenous group, but it does not counteract the effect of low lifespan disparity at younger ages. Therefore, the final effect is a lower dispersion of deaths in White-Mestizos compared with Indigenous people.

Figure 5.3: Decomposition of Lifespan Disparity (LD). Age Groups Where White-Mestizos Experience a Decrease in LD Compared to Afro-descendants and Indigenous Groups



Differences in life disparity between White-Mestizos and Afro-descendants (left), and between White- Mestizos and Indigenous (right) show that differences are generated fundamentally in childhood ages and young adults .

Source: Own elaboration based on the 2018 Census.

5.3.2 Mortality Differences in Death Registers

Our previous analysis has reconfirmed that mortality from Colombian death registers is underestimated especially for ethnic groups, see Chapter 4. Adjustments based on census data using equation 5.2.2, however, yield consistent results for most of the years from 2008 to 2019. Figure 5.4 shows the mortality trends from 2008 to 2019 by ethnicity and age. The dashed lines represent uncorrected mortality rates, and the solid lines represent corrected mortality rates adjusted for under-registration. Uncorrected data show higher mortality of White-Mestizo people during the whole period, and that for Indigenous from 2008 to 2019, present the lowest level of mortality in the country. However, after considering the under-registration bias, Indigenous and Afro-descendants populations have a higher level of mortality than White-Mestizos. Indigenous groups present high mortality between ages 0 and 10, and Afro-descendants between ages 15 and 30, and in the last three years of the observation period, mortality rates in Afro-descendants decreased considerably.

Corrected mortality rates show that mortality differences between White-Mestizos and Indigenous groups increased during the period 2008-2019 while Afro-descendant people present a significant mortality reduction and the mortality gap in relation to White-Mestizos disappear during the last three years of the analysed period. This result for Afro-descendant group requires however additional investigation due to the underestimation of deaths in both the census from 2018 as well as the mortality records.

Figure 5.4: Corrected and Uncorrected Mortality Rates by Ethnic Groups from 2008 to 2019

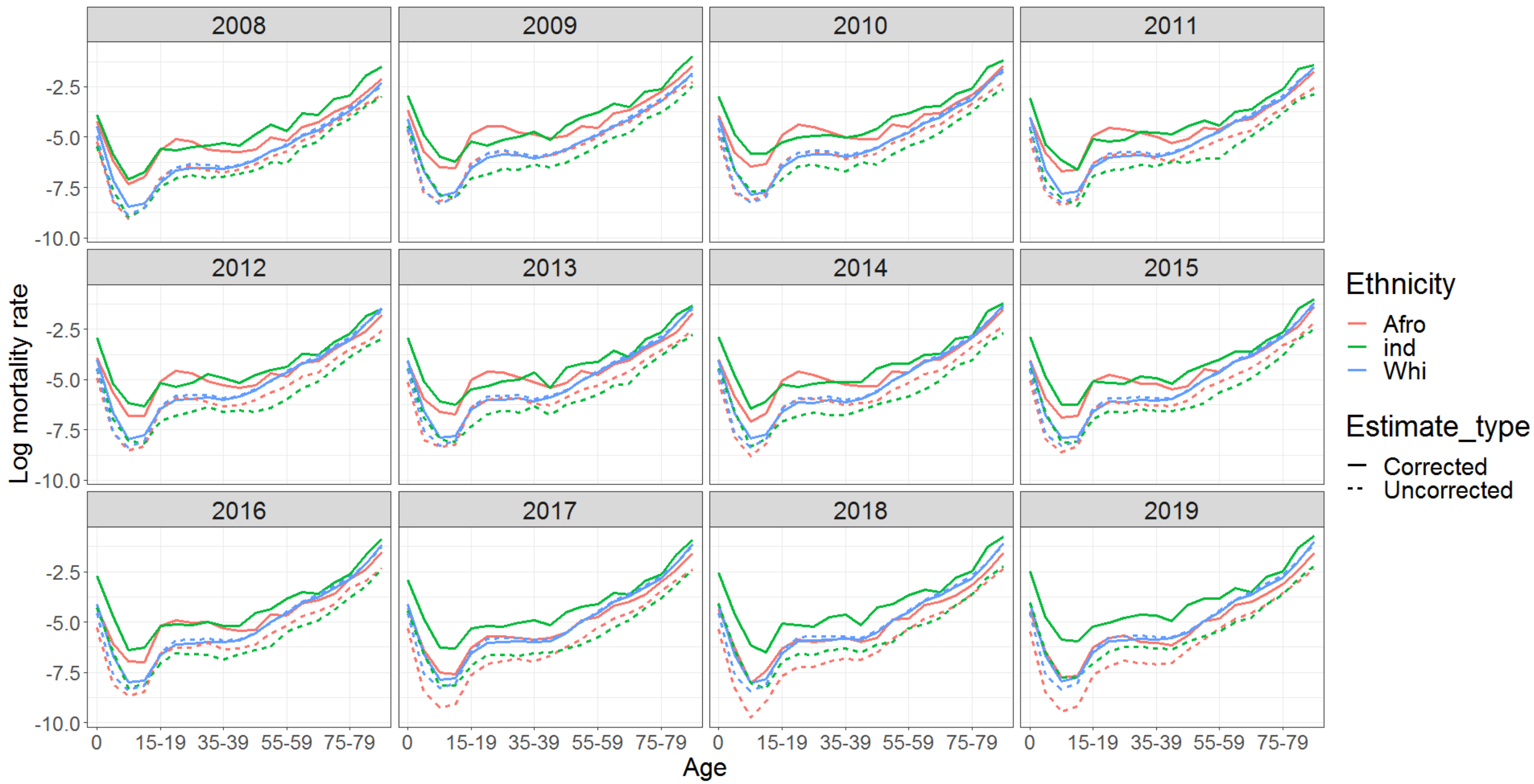


Figure 5.4 presents mortality trends by ethnicity and year for adjusted and unadjusted mortality rates. The unadjusted mortality rates (dotted lines) underestimate the level of mortality in Afro-descendant and Indigenous populations. After corrections, the adjusted mortality rates (solid lines) show that Indigenous and Afro-descendants have a higher level of mortality than White-Mestizos.

Source: Own elaboration based on the 2018 Census.

5.4 Discussion

Mortality and health disadvantages for Indigenous and Afro-descendants people have been apparent in both census and mortality register data. Comparing data from these two different sources yield consistent and complementary results that allow a better understanding of mortality patterns and levels in ethnic population. Census data help identify that Indigenous and Afro-descendant populations display high mortality in children and young populations respectively, which explain lower life expectancy in Indigenous and Afro-descendant people compared with White-Mestizos.

Indigenous population have a lower life expectancy at birth of 66.7. Child mortality, deaths related to pregnancy, childbirth, and nutritional deficiencies play an important role in this group, which suggests potential influence of social and cultural conditions of the ethnic minorities in Colombia. Indigenous groups are the poorest population of the country, and located in remote regions with poor access to health centres. Home birth practices in Indigenous communities and the lack of training, malnutrition and poor hygiene during births could explain part of this finding (Human-Right-Watch and Johns-Hopkins-Center, 2020). An investigation of the causes of death of different ethnic groups in Colombia is presented in the next chapter. Afro-descendant group on the other hand, have relatively better survival chances compared with Indigenous, but the differences were considerable when compared with White-Mestizo people. Life Expectancy at birth in Afro-descendant is 71.5 years.

White-Mestizo group shows a life expectancy of 78.9 years and thus better survival probabilities in the Colombian society compared to other groups. The differences in life expectancy among ethnic groups are surely getting wider over time due to the increase in the mortality gap between White-Mestizos and Indigenous groups. Official data give however opposite results, but the level of bias due to mortality underestimation for Afro-descendant and Indigenous deaths does not allow demographers to obtain reliable results. Estimations with uncorrected Colombian data should be considered thus with caution when comparing ethnic groups.

Despite the correction of bias in the mortality registers for Afro-descendants and Indigenous deaths using weighting methods based on the 2018 Census, the results are not exempt from limitations. The census is the most reliable registry of vital statistics in the country; however, it still suffers from undercoverage errors, particularly with the Afro-descendant population. This suggests that the census may not have captured all deaths among Afro-descendants, partially explaining why the mortality gap between White-Mestizos and Afro-descendants is significantly narrower compared to the gap between White-Mestizos and Indigenous peoples. The apparently better indicators among Afro-descendants may be influenced by this census error. Data quality therefore presents a limitation to the study, as mortality rates for Afro-descendants may be slightly underestimated. However, it is important to note that while the overall mortality levels may be affected, the patterns of mortality by age, and gender are unlikely to be significantly altered, as the level of error is likely proportional within the Afro-descendant group. Furthermore, two aspects should be considered for future studies: First, the occurrence of no deaths or only one death in some regions poses a challenge that cannot be corrected with the employed methods based on weighted factors. Second, the absence of information regarding the population size of ethnic groups for years other than those covered by the census prevents a time-varying analysis of non-reporting probabilities with time-varying exposures.

Another limitation is that uncertainty introduced by the correction process is not quantified in the results, and therefore the reliability of the results is difficult to assess. Such uncertainty may result from the underlying binomial uncertainty relating to the decision to register as well as from the errors in the estimation of the regression model coefficients. Quantification of such uncertainty may be possible. Simulation experiments may aid with examination of the extent of this uncertainty, but a full examination of such uncertainty was beyond the scope of this explanation.

The study however represents an important contribution in terms of the understanding of mortality patterns in the Colombian ethnic population, and allowed us to identify the age groups and gender that determine the ethnic mortality differences in the country.

CHAPTER 6

Causes of Death and Ethnicity in Colombia

6.1 Introduction

Causes of death are probably the best indicator of social inequalities. Unlike indicators of living conditions that show a "picture" of the current situation that could not correspond with the life conditions in which people have lived the most of the time, cause of death captures the cumulative effect of environment and living conditions over long periods of time. The analysis of social conditions and inequalities among social groups requires thus not only the consideration about how people live, but it requires likewise the analysis about how people die.

The transition from leading causes of death such as infections diseases, malnutrition, and maternity complications to causes such as cardiovascular diseases and cancer, were determined by improvements in preventive and curative measures as well as a better hygiene and nutrition of the population. This explains therefore mortality differences among countries with different socioeconomical conditions. But it could be insufficient to explain differences among groups that share the same stage of development and even the same space like the case of multiethnic countries in which particular groups show great differences in terms of age and causes of death.

Causes of death, therefore, represent a significant issue from an ethnic perspective, as they affect White and non-White populations to varying degrees. During the COVID-19 pandemic, for example, non-White populations were affected at a higher rate as a result of underlying social and economic inequalities that led to greater exposure to the virus (Nazroo and Becares, 2020; Gross et al., 2020; Platt and Warwick, 2020; Ogedegbe et al., 2020). Similarly, studies have shown that Indigenous populations in New Zealand, Australia, Canada, and the US exhibit higher mortality rates from diabetes but lower mortality rates from lung cancer. This disparity is attributed to the lower prevalence of habitual cigarette smoking within these populations (Bramley et al., 2004). In the US, Black and Indigenous people face a higher risk of being killed by police than do Whites: for Blacks the risk is 2.5 times higher, and for Indigenous it is between 1.2 and 1.7 times (Edwards et al., 2019). The studies show that racism and ethnic discrimination are a social determinant of health and many of the disadvantages for non-White groups could be explained by poverty, and although in the last decades has been a decreasing trend in causes of death like heart disease, cerebrovascular disease, diabetes mellitus, and influenza and pneumonia for the whole population; disparities among ethnic groups have increased in causes such as heart diseases, respiratory diseases, and nephritis (Chang et al., 2016).

Previous studies have reported health and mortality inequalities among White-Mestizos, Afro-descendants, and Indigenous populations in Colombia. Afro-descendant women exhibit higher child mortality rates compared to their White-Mestizo counterparts, with Afro-descendants having a 25% higher risk of losing a child (Palacios, 2018). Additionally, regions with a higher proportion of Afro-descendant and Indigenous inhabitants show increased mortality from infectious and perinatal diseases, while regions in the center of the country, predominantly inhabited by White-Mestizos, display higher mortality rates from cancer (Spijker et al., 2020). Afro-descendants also experience a higher homicide rate. In Cali, the city with the highest number of Afro-descendants in the country, 80% of deaths among individuals aged 15 to 19 in this group are due to homicide (Urrea-Giraldo et al., 2021). Although these studies focus on specific Colombian cities and specific causes of death, or they use regions as proxies for ethnicity, thereby excluding ethnic populations living outside those regions; these studies do provide essential context for the present work, which aims to examine all causes of death at a national level by ethnicity.

6.2 Data and Methods

As discussed in the previous chapter, the data for this analysis are derived from the 2018 Census and the mortality registration records from 2008 to 2019, with population size information sourced from the 2005 and 2018 Censuses. This information was provided by the Colombian National Statistics Office (DANE) and is detailed in Section 5.2.1.

The analysis of causes of death by ethnic groups was conducted using data from mortality records based on the International Classification of Diseases, version 10 (ICD-10). To examine the primary causes of death by ethnic groups, we addressed the downward bias in death counts resulting from under-registration. This was achieved by assuming that the distribution of causes in observed deaths resembles that in unobserved deaths. Specifically, if the total number of registered deaths from mortality records is underestimated, we presumed that under-registration by causes of death occurs at the same rate as for total deaths, considering ethnicity, age group, and sex. For example, if child mortality in a specific ethnic group was underestimated by 50%, then all causes of death in that group were similarly underestimated by the same proportion. Consequently, the correction factor for a specific cause of death is contingent upon age group, sex, and ethnicity. This correction method ensures a more accurate approximation of the overall level of mortality and facilitates the estimation of the genuine impact of causes of death among Colombian ethnic groups.

Information regarding causes of death was extrapolated to census data from the mortality records as follows: if at time t_i, the total number of registered deaths at age a_i for a specific ethnic group and gender were composed of the causes $x_1 + x_2 + x_3 + ...x_n$ with percentages $p_1 + p_2 + p_3 + ...p_n$; then the total number of deaths reported in the Census 2018 were distributed following the same proportions of the causes by gender and ethnicity. This approach enhances the data quality for two primary reasons. First, the census provides the most reliable estimation of the true number of deaths in the country. Second, while mortality registers may suffer from some level of under-registration (see Chapter 4), the mortality patterns in terms of causes of death by age, gender, and ethnicity are assumed to remain relatively unaffected. Consequently, by unifiying the number of deaths from the census with the patterns of causes of death from mortality records, we obtain a reliable description of the levels of mortality by cause among Colombian ethnic populations.

We applied decomposition methods by age group, cause of death, and ethnicity to identify the major contributors to changes in life expectancy (e^0) and lifespan disparity ($e^\dagger$) among ethnic groups. The analysis of decomposition by single age and causes of death is based on the 2018 Census and life tables for different ethnic groups. Similar to Chapter 5, the decomposition method used here is the one proposed by Horiuchi et al. (2008), wherein the difference in the dependent variable is expressed as a sum of the effects in its covariates, as detailed in Section 5.2.2.

The model of decomposition as presented in Section 5.2.2 assumes a numerical population characteristic y which is a differentiable function (f) of n covariates $\mathbf{X} = x_1, x_2, ..., x_n$.In the results presented in this chapter, the population characteristic of interest is life expectancy or life disparity and the covariates are the causes of death and age of death. Let us assume that both y and the x_i depend on time t. Suppose now that information from y and $\mathbf{X}$ is available at two time points t_1 and t_2, and $\mathbf{X}$ is a differentiable vector of covariates between t_1 and t_2.

Then as before, model of decomposition according to Horiuchi et al. (2008, p787) has the following form:

$$y_2 - y_1 = \sum_{i=1}^{n} c_i, \text{ where } c_i = \int_{x_{i1}}^{x_{i2}} \frac{\partial y}{\partial x_{i1}} dx_i \tag{6.2.1}$$

Equation 6.2.1 shows that the total change in the dependent variable can be expressed as a sum of effects of the covariates. "It is important to note that regression analyses and decomposition analyses are based on different notions of the effect of a covariate on a dependent variable. The linear regression coefcient for x_i shows a change in y produced by a *unit change* in x_i, $\partial y/\partial x_i$. A decomposition analysis estimates a change in y produced by a *particular change* in x_i from t_1 to t_2, which can be expressed as $c_i = \int_{x_{i1}}^{x_{i2}} \frac{\partial y}{\partial x_{i1}} dx_i$" (Horiuchi et al., 2008, p788).

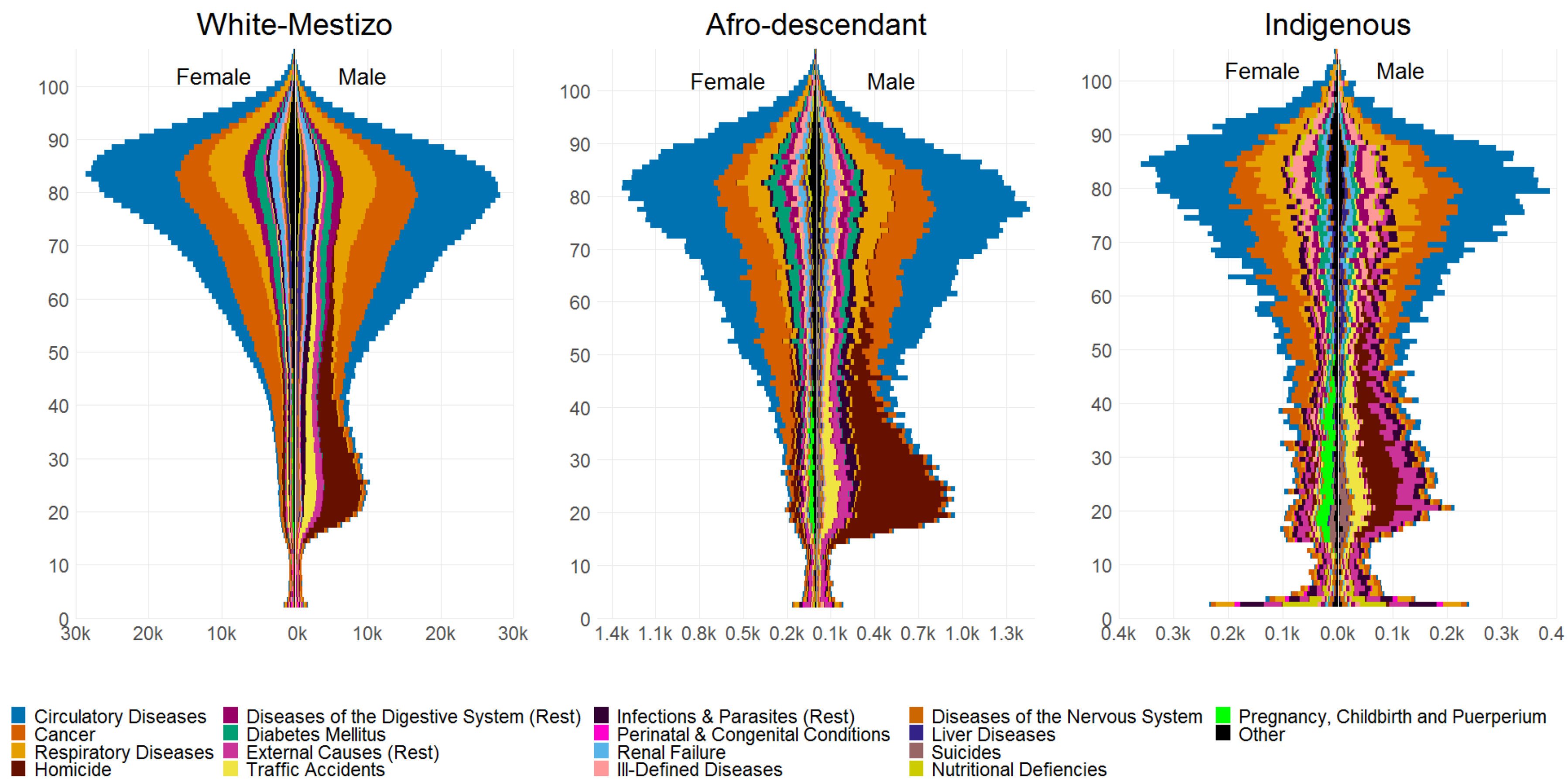
In contrast to the previous Chapter, the vector of covariates **x** consists not of mortality rates by age, but a complete set of *age and cause* specific death rates. These rates are complete in the sense that the sum of all included cause-specific rates for a particular age will give the all-cause mortality rate for that age. The decomposition effects and classification of causes of death were estimated with the function *horiuchi* from *DemoDecomp* package and the algorithm for plotting causes of death from Jonas Schöley respectively.

6.3 Results

6.3.1 Causes of Death

The higher mortality of young males in Afro-descendants and mortality of children in Indigenous are linked to specific causes of death. Figure 6.1 shows the most frequent causes of death in ethnic population over 2 years old. The very young population are excluded from the plot to allow deaths to adult population to be displayed more clearly.

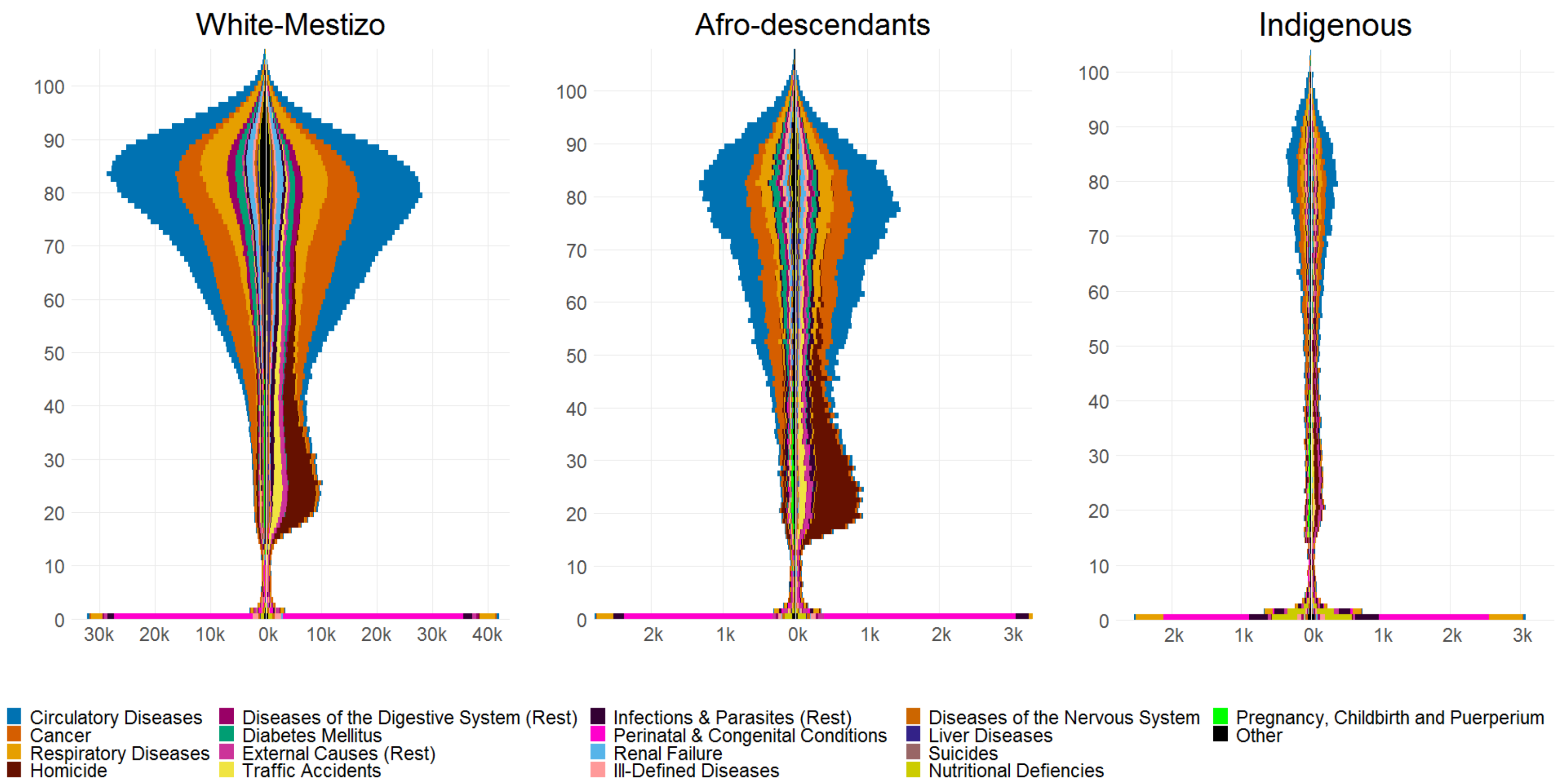
Figure 6.1: Causes of Death in the Ethnic Population Over Two Years Old from 2008 to 2019



Source: Own elaboration based on Colombian mortality records.

The most common causes across all ethnicities are circulatory diseases, cancer, respiratory diseases, homicide, and other external causes, particularly road traffic accidents. The young population in Afro-descendants are particularly affected by homicide, and young Indigenous women experience higher mortality during pregnancy, childbirth and puerperium (represented by a bright green area in the plot). Likewise, the Indigenous groups have a higher child mortality than White-Mestizos and Afro-descendants. This is more clear in Figure 6.2 that includes mortality at ages 0 and 1. Mortality by perinatal and congenital conditions is the main cause of death in children under 1 year old for all groups. But in the case of Indigenous, nutritional deficiencies likewise play an important role in infant mortality; and mortality of Indigenous children is excessively high compared with the mortality at older ages in the same ethnic group.

Figure 6.2: Causes of Death in Ethnic Population in All Ages from 2008 and 2019

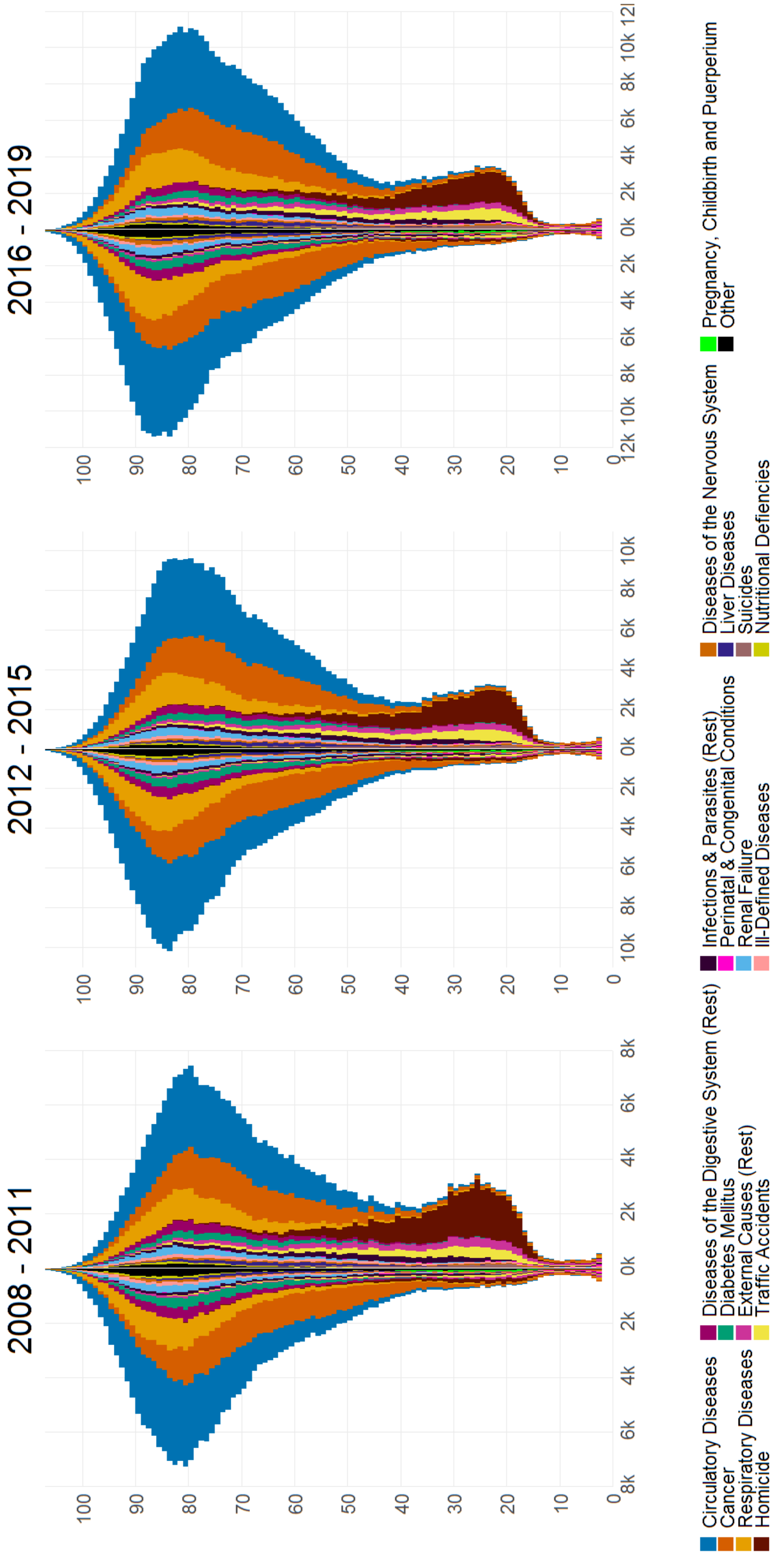


Source: Own elaboration based on Colombian mortality records.

Figures 6.3, 6.4, and 6.5 show the trend of the main causes of death during the periods $2008-2011$, $2012-2015$, and $2016-2019$ for all ethnic groups. Figure 6.3 shows White-Mestizos' causes of death, and how they change with age and gender. The proportion of people dying at the older ages increased during the last two periods, and a higher proportion of people died after age 70 due to circulatory diseases, cancer, and respiratory diseases in the period $2016-2019$. Homicide was the main cause of death in young population amongst the White-Mestizo group, but the proportion of deaths in these groups is relatively low compared with the mortality in older age-groups.

Afro-descendants causes of death are presented in Figure 6.4. The proportion of elderly population dying due to circulatory diseases and cancer is significantly higher in the last part of the period. Deaths at age 0 in Afro-descendants are mainly due to perinatal and congenital conditions. Additionally, Afro-descendants present the highest level of deaths due to homicides in the country. Although homicide is the third highest cause of death after circulatory diseases and cancer, it is the primary cause of death for male Afro-descendant aged between 15 and 40 years old, and they account for a considerably higher level of mortality when compared with Indigenous and White-Mestizo groups. Surprisingly, the trend in homicides shows a substantial decrease among Afro-descendants during the period $2016-2019$.

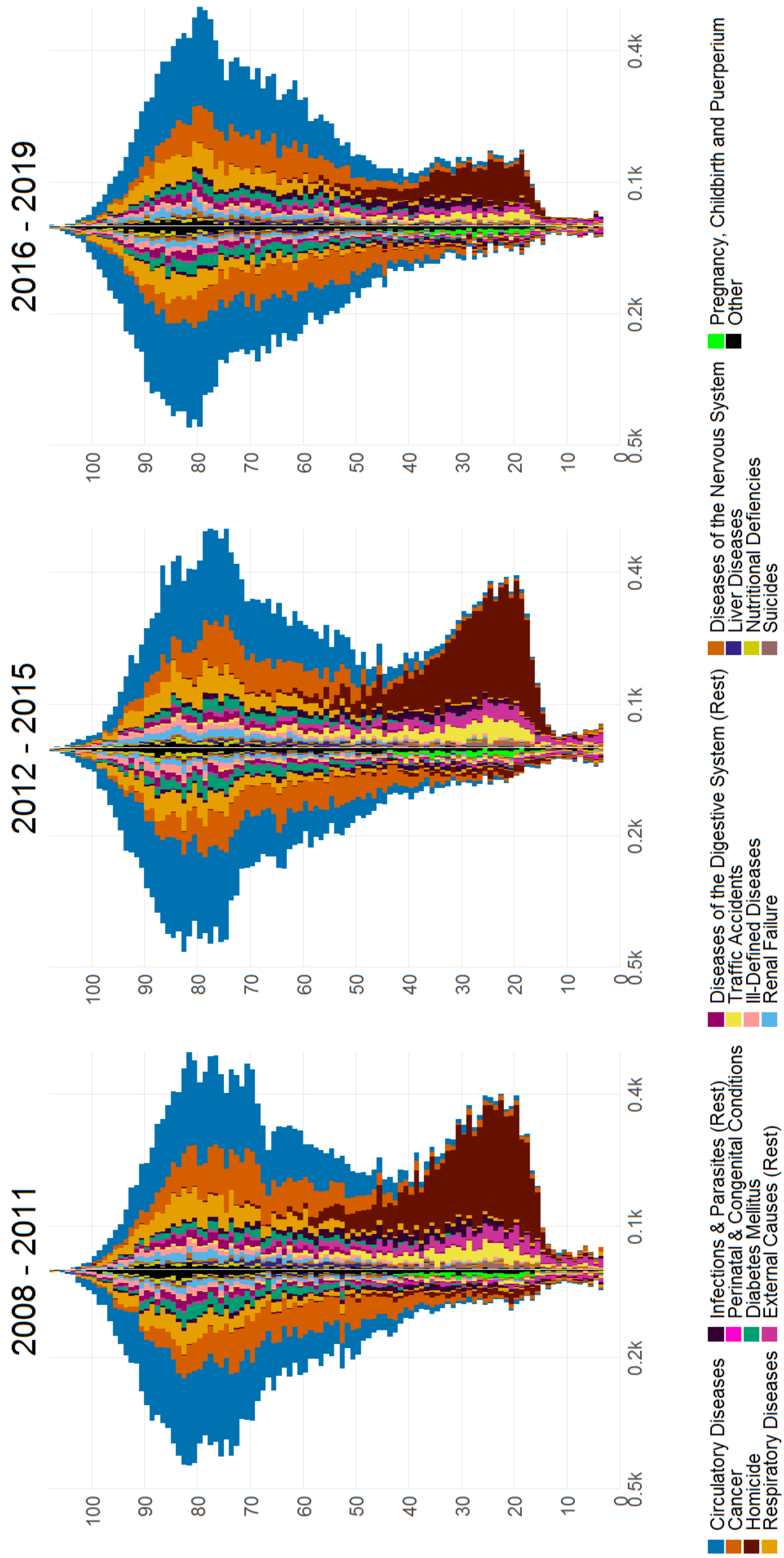
Figure 6.3: Causes of Death in White-Mestizo Population from 2008 to 2019 by Periods



.

Source: Own elaboration based on Colombian mortality records.

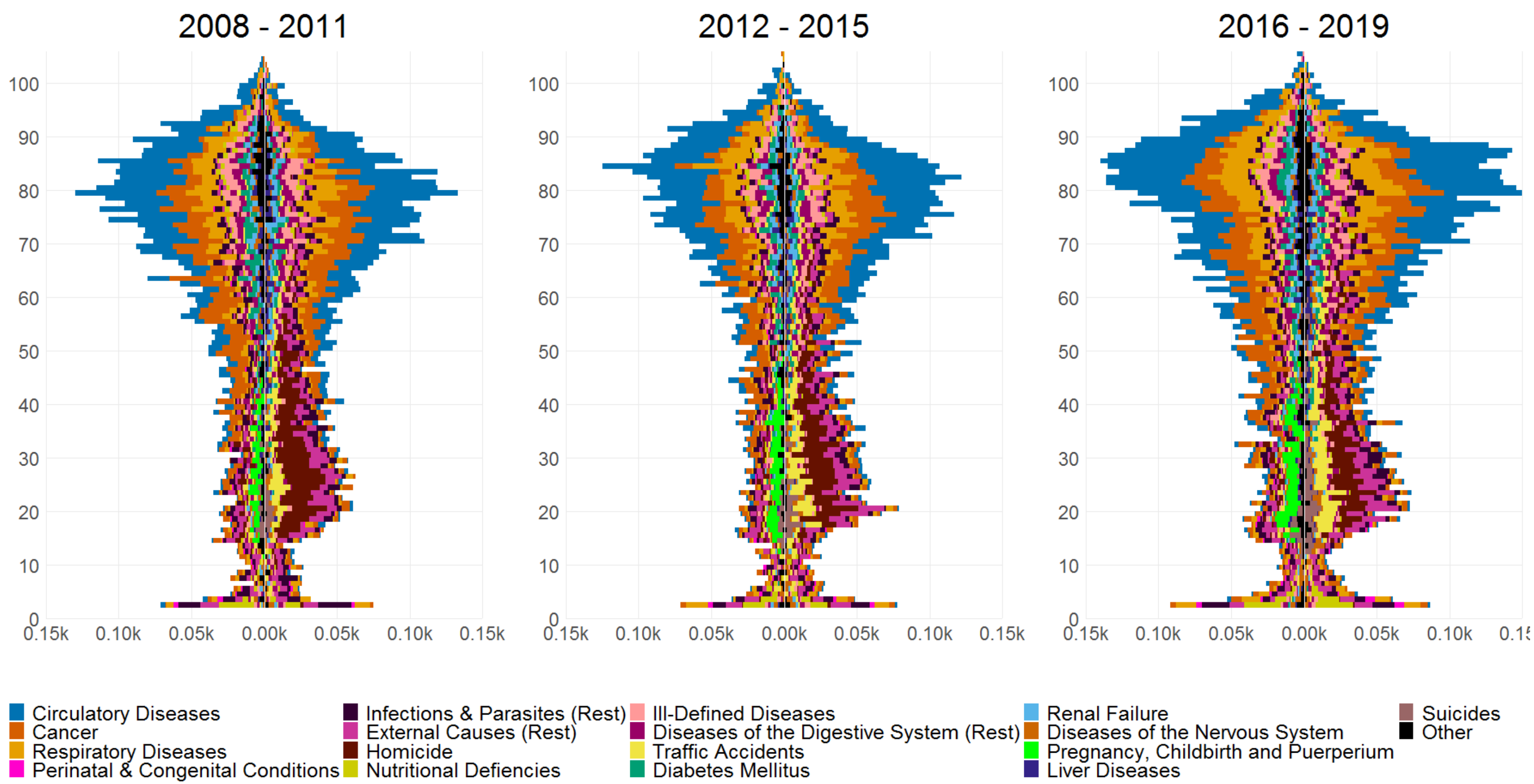
Figure 6.4: Causes of Death in Afro-descendant Population from 2008 to 2019 by Periods



.

Source: Own elaboration based on Colombian mortality records.

Figure 6.5: Causes of Death in Indigenous Population from 2008 to 2019 by Periods



.

Source: Own elaboration based on Colombian mortality records.

Indigenous causes of death in Figure 6.5 show an excessively high child mortality in this group, and the composition of the causes is more heterogeneous compared to those in White-Mestizos and Afro-descendant groups. Perinatal and congenital conditions are the leading cause of death in Indigenous at age 0. This makes Indigenous mortality between ages 0 and 5 disproportionately high compared to other groups, and they are the only ethnic group where mortality by pregnancy, childbirth, and puerperium is significantly high. Other causes such as homicide show a slight decrease while deaths due to traffic accidents and the proportion dying in the older groups increased from 2008 to 2019. The elderly population in Indigenous as well as in White-Mestizos and Afro-descendants people die mostly due to circulatory diseases, cancer, and respiratory diseases.

The contribution of the causes of death to differences in life expectancy and lifespan disparity vary for Afro-descendants and Indigenous groups when compared to White-Mestizos, as shown in Figures 6.6 and 6.7. With respect to contributions of causes of death to differences in life expectancy between White-Mestizos and Afro-descendants, the largest contributions are seen at ages 0 and 5 attributed to perinatal and congenital conditions, respiratory diseases, infectious diseases, and parasitic infections. In young adults, homicide, infectious diseases, pregnancy, childbirth, and puerperium make an important contribution to the life expectancy gradient between White-Mestizos and Afro-descendants. After the age of 40, the largest contributors are circulatory diseases and cancers. The bars under 0 show Afro-descendant mortality to some specific causes reduce the differences versus White-Mestizos slightly, for example at age 1 by congenital conditions.

Comparing Indigenous groups with White-Mestizos show that life expectancy and lifespan differences are determined fundamentally during childhood by perinatal, congenital conditions, respiratory diseases, nutritional deficiencies, and infections and parasites. Likewise, teenagers and adults until age 40 especially females die mostly due to external causes, pregnancy - childbirth and puerperium. Additionally, after age 50, Indigenous population are at a higher risk of dying due to cancer, respiratory diseases, and circulatory diseases when compared with White-Mestizos.

Figure 6.6: Decomposition of Life Expectancy Differences by Causes of Death Comparing Afro-descendants and Indigenous Populations with White-Mestizos in 2018

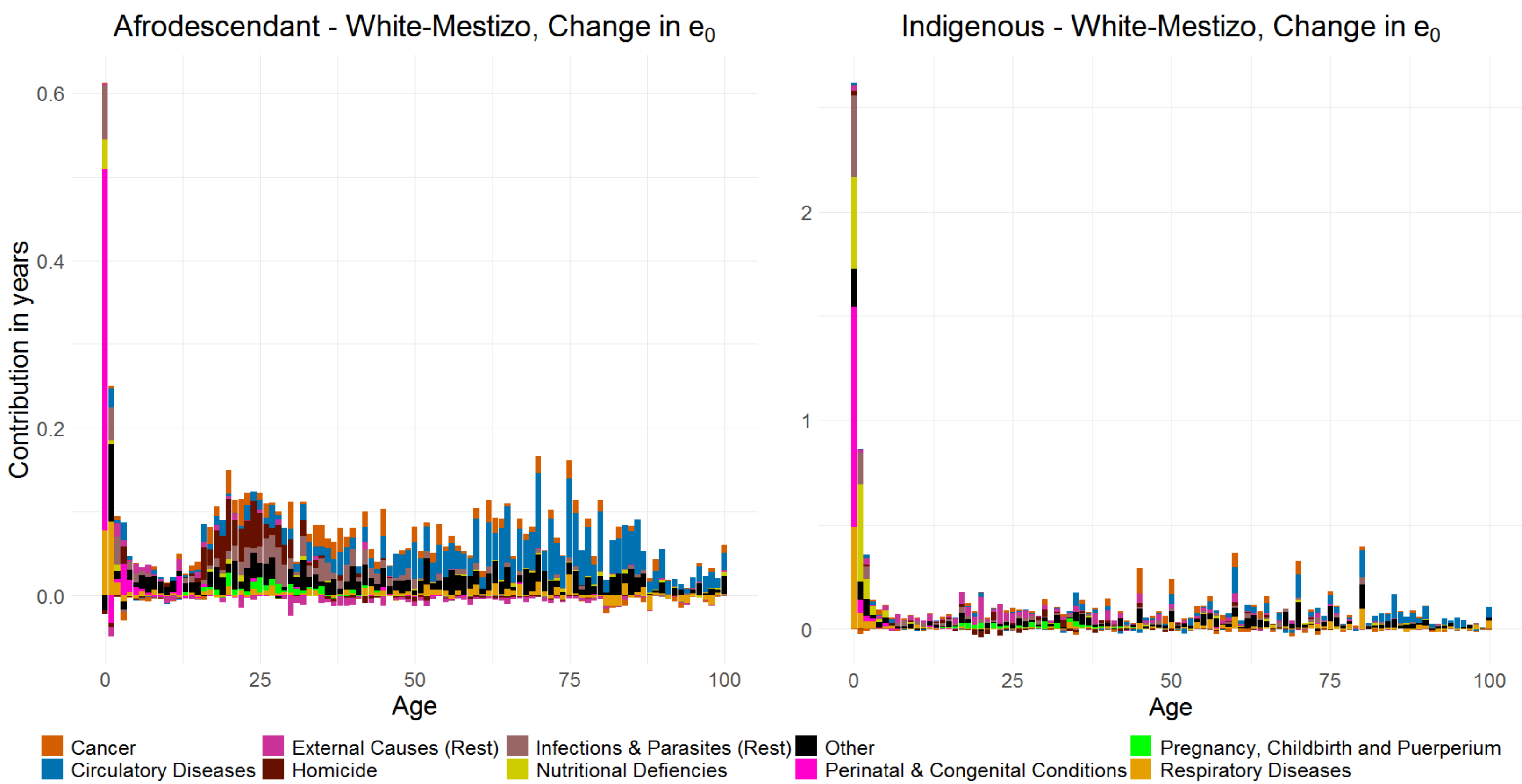


Figure 6.6 illustrates the causes of death and age groups in which the White-Mestizo group has accumulated life expectancy advantages compared to Afro-descendants and Indigenous people. For instance, at age 0, perinatal mortality accounts for the largest contribution to the life expectancy gap between Afro-descendants and White-Mestizo populations, due to the lower mortality rate from this cause in the latter group.

Source: Own elaboration based on the 2018 Colombian Census and mortality records.

Figure 6.7: Decomposition of Lifespan Disparity by Causes of Death Comparing Afro-descendants and Indigenous Populations with White-Mestizos in 2018

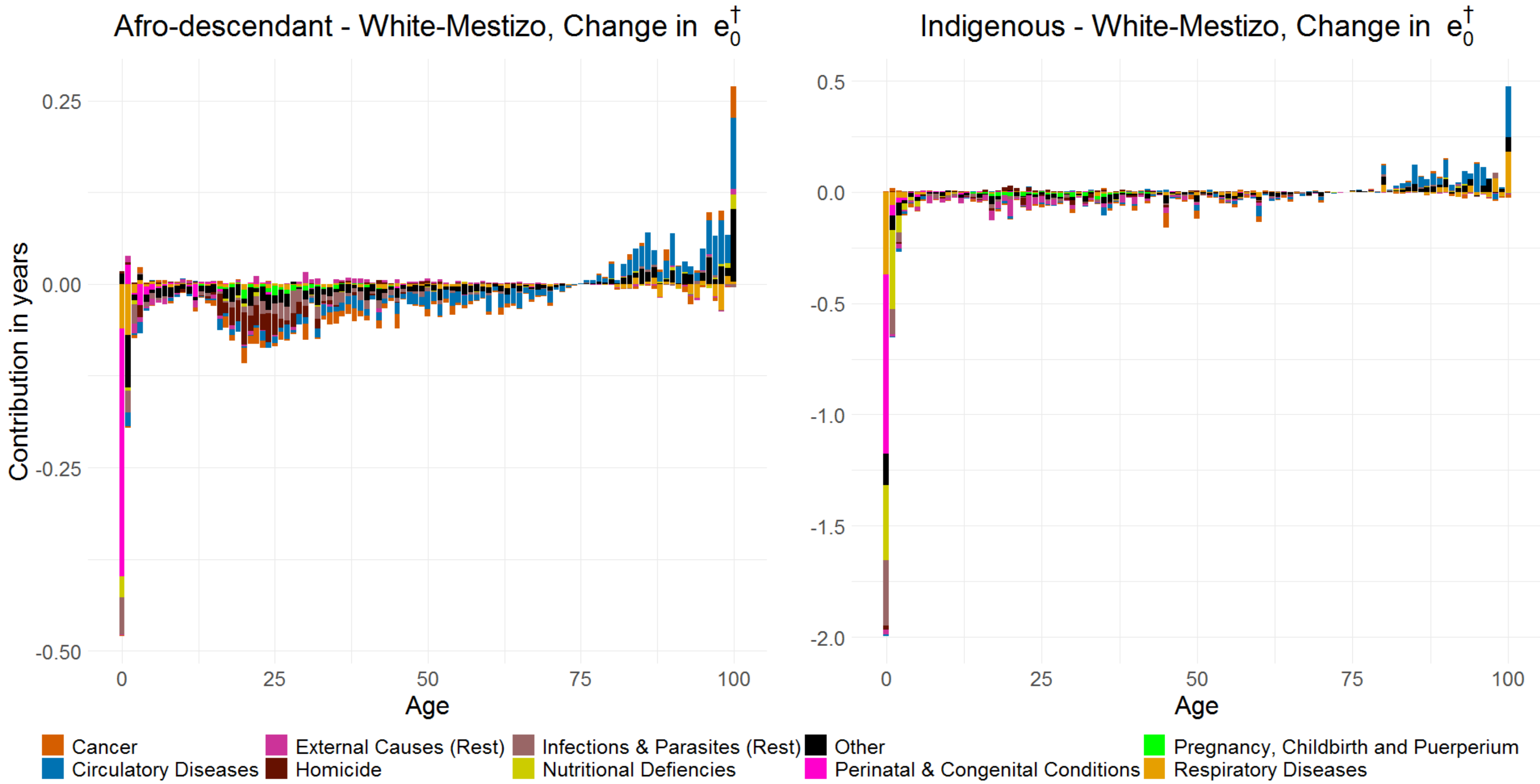


Figure 6.7 presents the causes of death and age groups that most contribute to the lower lifespan disparity among White-Mestizo people. For example, the lower perinatal mortality at age 0 in the White-Mestizo population allows more members of this group to survive into older age groups. Consequently, this cause of death significantly reduces lifespan disparity in White-Mestizo populations when compared with Afro-descendant groups.

Source: Own elaboration based on the Colombian 2018 Census and mortality records.

6.3.2 Adjusted Mortality by Causes of Death

Figures 6.8, 6.9, and 6.10 present the corrected and uncorrected distribution of causes of death by ethnicity and gender. The results show that after the correction, mortality patterns remain almost unchanged, with an exception of slight changes in the distribution of some causes. The adjustments to the causes of death distributions presented below focus on finding variations in age-patterns of deaths instead of changes in the mortality level as was showed in Chapter 5. For this reason, the correction factor weights up to a greater extent the age groups in which mortality under-registration is higher. However, it is possible that changes in proportions by age-groups are not large enough to reshape the curve of mortality. Therefore, mortality under-registration may have a greater impact on the level of mortality, but much less on mortality patterns.

Figure 6.8: White-Mestizos Causes of Death, Corrected and Uncorrected Mortality Distributions

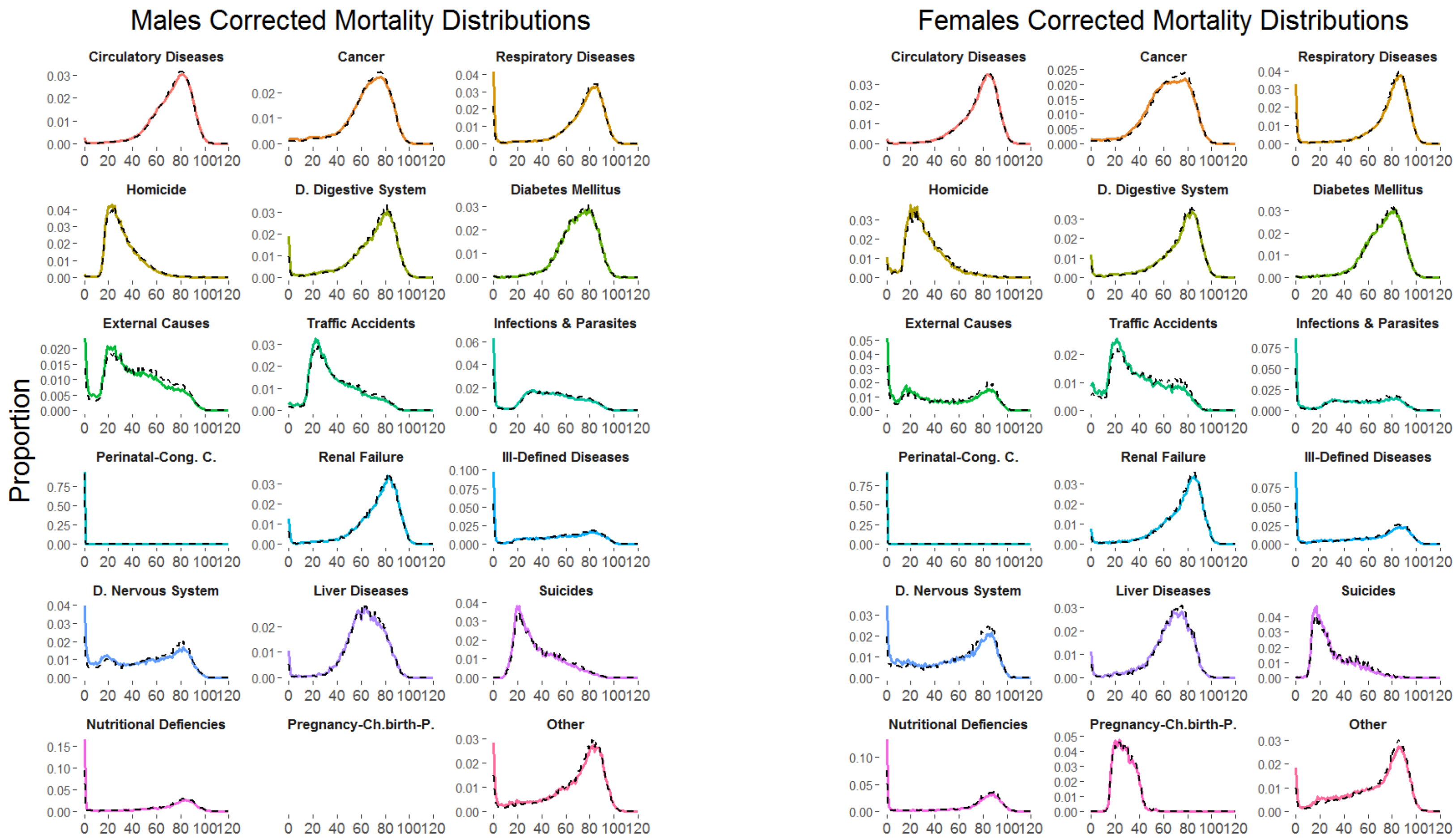


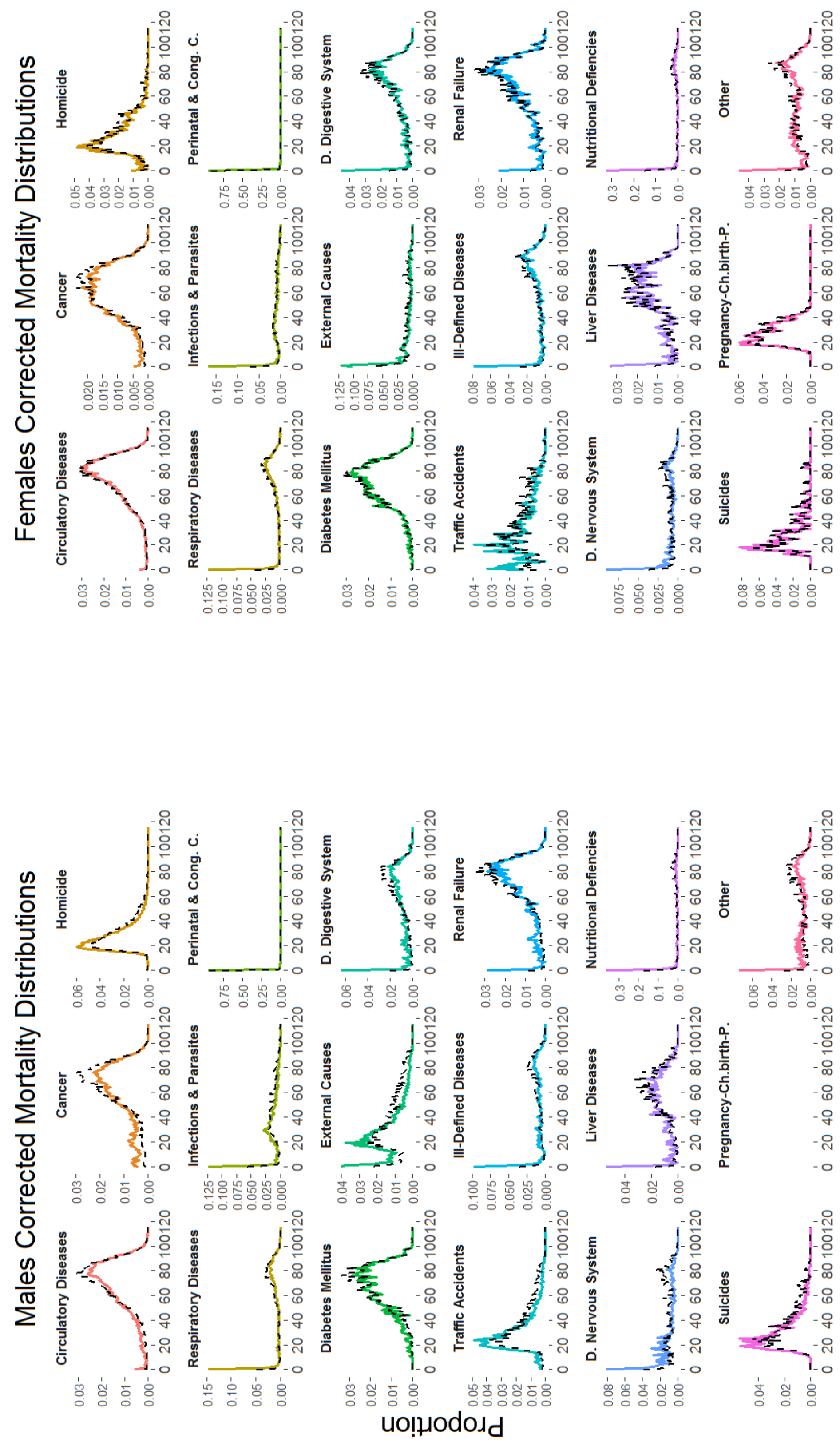
Figure 6.8 shows the comparison between observed deaths (dotted line) and corrected deaths (solid line) by causes and gender. The highest increase in the proportion of corrected deaths is in child deaths, particularly in causes such as infectious diseases, malnutrition, and respiratory diseases. In diseases such as cancer and circulatory diseases, the correction factor redistributed the total deaths, giving a higher weight to age groups with higher under-registration mortality.

Source: Own elaboration based on the Colombian 2018 Census and mortality records.

Corrected causes of death in White-Mestizos present trivial changes in distribution for males and females. The biggest changes in the mortality distribution are at age 0 which presents a clear increase for infectious and Parasites diseases, respiratory diseases, nutritional deficiencies, and external causes. In young adults, much of the increase in mortality are due to traffic accidents, external causes, and homicides, while cancer had a redistribution to younger ages in both males and females. In Afro-descendants, changes in the adjusted distributions are bigger for males than females (Figure 6.9). They present a higher variation in deaths by circulatory diseases, cancer, and homicide than females.

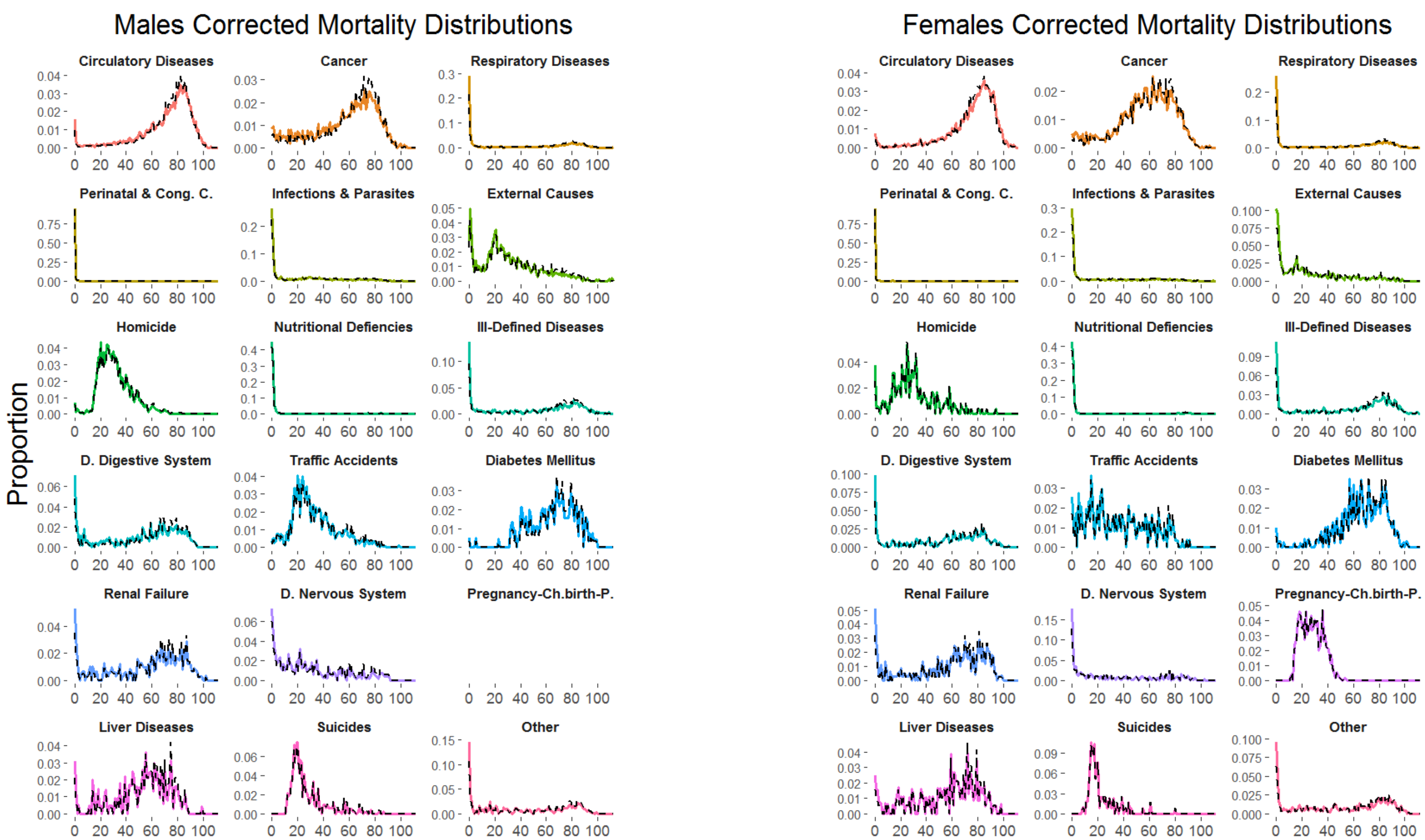
Corrected distribution of deaths in Indigenous, Figure 6.10, do not show any significant changes in the older population, but deaths of children under age 1 show a clear increase in causes of death such as respiratory diseases and ill-defined diseases.

Figure 6.9: Afro-descendants Causes of Death, Corrected and Uncorrected Mortality Distributions



. Source: Own elaboration based on the 2018 Census and mortality records.

Figure 6.10: Indigenous Causes of Death, Corrected and Uncorrected Mortality Distributions



.

Source: Own elaboration based on the 2018 Census and mortality records.

The unusual results showing a decline in homicides in Afro-descendants warrant caution, as it could be a consequence of data quality rather than a reflection of the social reality. Figure 6.11 presents the trend of homicides by ethnicity and departments from 2008 to 2019. The departments of Valle del Cauca, Chocó, and Bolivar host the highest absolute number of Afro-descendants people in Colombia. These regions present a decrease in the number of registered homicides among Afro-descendants. According to 2005 Census more than 1 million of Afro-descendants are located in the department of Valle del Cauca, which reports on average between 500 and 600 Afro-descendants homicides per year, but for 2017, 2018, and 2019, it reported only 25, 2, and 35 homicides respectively. Likewise, the department of Bolivar reported only 2, 1, and 0 homicides for the same years. This phenomenon, however, is systematic across the country. Deaths by homicide decreased abruptly for Afro-descendants across all the departments (Figure 6.11). At the same time, it increased for White-Mestizos, which could be possibly due to an ethnic registration bias. Further analysis is required to understand the potential influence of registration biases on the mortality reduction in Afro-descendant males. In contrast, similar patterns are not seen in other causes; the top row of plots in Figure 6.11 shows trends for deaths Circulatory Disease, for example.

Figure 6.11: Trend of Registered Homicide and Circulatory Deaths by Ethnicity in Selected Departments Between 2008 and 2019

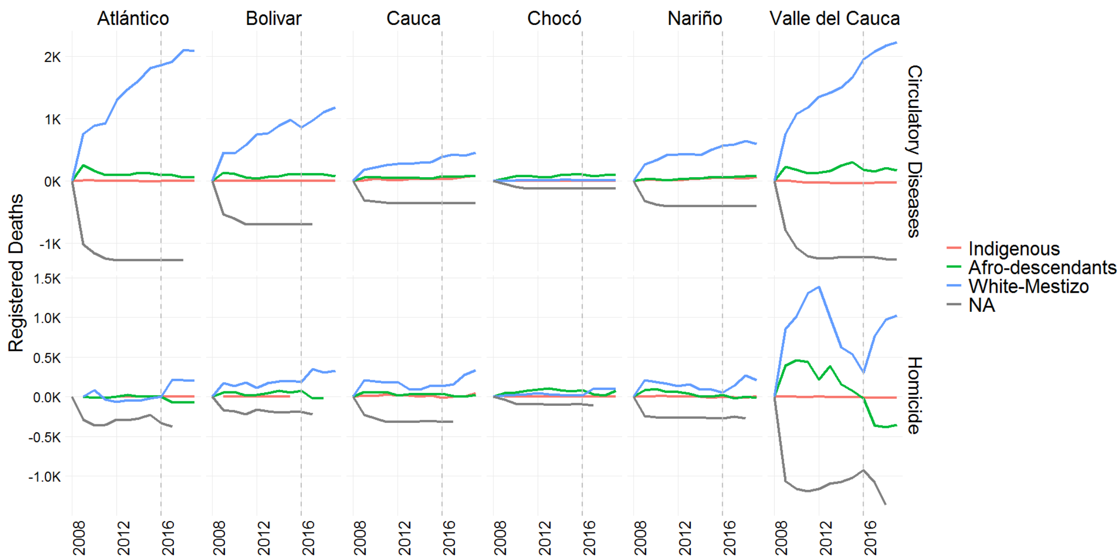


Figure 6.11 shows the number of registered deaths in Colombian department where a significant proportion of Afro- descendants reside. Number of registered deaths are classified by department, cause of death and ethnicity. The plot shows absolute changes over time in the number of registered deaths with respect to the baseline year 2008. Registered deaths in Indigenous remain relatively constant by departments and causes of death. In White-Mestizos and Afro-descendants, the results show an unusual trend in deaths by homicide (bottom row). From 2016 onwards, Afro-descendant deaths show a significant decline while White-Mestizos deaths increased during the same period. The department of Chocó, where around 80% of the population is Afro-descendants, has shown historically a higher number of Afro-descendant deaths by homicide. However, since 2016 the number of registered death by homicide from White-Mestizos is higher than the number of deaths from Afro-descendants. This general pattern in different Colombian departments could be due to a phenomenon of ethnic misclassification. The top row of the plot shows trends in deaths by circulatory disease by way of comparison. Source: Own elaboration based on Colombian mortality records.

6.4 Discussion

The most significant finding of this paper is that the Indigenous population, in comparison to other ethnic groups, has higher child mortality, and causes of death by infectious diseases are more common in this group. Indigenous populations have been historically located in rural areas with difficult access to health facilities and clean water. Home birth is likewise a regular practice in indigenous communities, and probably the lack of training in the personal and poor sterilisation of material and equipment employed during this procedure could explain not only the disproportionate infant mortality compared with other ethnic groups, but also the high mortality of Indigenous women during pregnancy, childbirth and puerperium. Additionally, environmental conditions in Indigenous remote areas could make this population more prone to deaths by infections, parasites, and nutritional deficiencies.

The Afro-descendant population on the other hand, shows higher mortality in teenagers and young adults, and the external causes of death, such as homicide, occur more often in Afro-descendant than in persons with another ethnic background. Afro-descendants are located along the Pacific and Atlantic coast and over time an important part of their population has migrated to the main cities of the country where they reside in the poorest areas. The city of Cali is probably the more clear example of this phenomenon. The excess mortality of young Afro-descendant males by homicide could be explained by the urban violence and criminality in depressed areas where they are located. Infant mortality is proportionally higher in Afro-descendants compared with With-Mestizo group, but it is much lower compared with Indigenous. This could be the result of having a greater percentage of the population in urban areas (Departamento-Administrativo-Nacional-De-Estadística, 2019b).

Mortality in young Afro-descendants decreased sharply in the last period 2016-2019, mainly due to the drop in the proportion of homicides. However, this finding based on data from death registers is not consistent with the information from 2018 Census which shows that Afro-descendant males were affected by relatively high mortality rates at young ages. Mortality by homicide in Colombia has been declining over the last few years, possibly attributed to the initiation of peace agreement with the guerrillas group FARC-EP in 2016. However, the observed decrease in homicides were more general across Colombian population and not exclusive to a particular ethnic group. The case of the department of Chocó for example is quite striking. In this department more than 80% of the population identified themselves as Afro- descendant which makes it the department with the highest concentration of Afro-descendants in the country. Between 2017 and 2019, the number of homicides from Afro-descendants dropped by a third, but the number of homicides among White-Mestizos were multiplied by five. As a result, Chocó, the typical Afro-descendant region in Colombia, has four times more homicides among White-Mestizos than among Afro-descendants.

The results in the white-Mestizo population are more consistent with the global mortality transition. It concentrates the majority of deaths at older ages and shows a higher prevalence of degenerative diseases such as cancer, circulatory, and respiratory diseases. This could be due to the higher proportion of White-mestizos surviving until older age groups when these causes of death are more common. Although homicide affects young White-Mestizo males and constitutes the main cause of death for males between the age of 15 and 40, it is not as high as in Afro-descendants, where the proportion of male deaths due to this cause at the age of 20 is similar to the proportion of male deaths at age 80. Causes of death by ethnicity present important differences in terms of what causes are more frequent in Indigenous, Afro-descendants, and White-mestizos as well as what age groups and gender are more affected. From Omran's perpective (Omran, 1971) these results suggest profound social inequalities among ethnic groups similar to differences between a poor and a rich country.

Data quality represents an important limitation of the present study. In some municipalities where deaths were not registered or reported, adjustments could not be made. Additionally, while the Horiuchi decomposition method offers valuable insights into mortality differentials among ethnic groups, it may oversimplify the complex mortality patterns in Colombian society as the method may not adequately capture the underlying factors related to racism and social inequities that contribute to mortality differentials. Furthermore, the method's reliance on continuous changes over time in the causes of death may be inappropriate if some causes change in discrete manners. Although changes in causes of death could be perceived as discrete shifts between individuals, when aggregated across a population, these changes tend to occur gradually over time. As such, this could be considered a minor limitation.

Despite the mentioned limitations, the study makes a significant contribution by enhancing our understanding of mortality patterns within Colombian ethnic populations. It enables the differentiation of causes of death by age, gender, and ethnicity, which will enhance the efficacy of interventions aimed at improving health outcomes in marginalised ethnic groups.

CHAPTER 7

Ethnic Identification in Mortality Records Using a Random Forest Model

7.1 Introduction

The study of health inequalities in ethnic minorities is often challenging due to lack of quality and reliable information. Difficulties with data quality can arise due to digit-preference (Myers, 1954), misreporting (Preston et al., 1999), overestimation (Scholz and Jdanov, 2008), underreporting (Gakidou et al., 2004), or simply missing information. In Colombia, the mortality registration form, for example, did not collect ethnic information before 2008 which makes any retrospective analysis of ethnic mortality impossible. Identifying the ethnicity of those who died before 2008 will enable better understanding and monitoring of long-term trends in ethnicity mortality.

However, ethnic classification is an ethically sensitive topic, and the “criteria” used to ascribe people or classify groups sometimes involve considerable political sensitivities and interests. In the United States, for example, the hypodescent method of ethnic classification determines that children from mixed couples are automatically assigned to the ethnic group with lower status. In this context, the "one-drop rule", which remained legal in the US until 1967, stipulated that having one drop of blood from a non-White person was enough to classify an individual as non-White. This biological classification of race stemmed from political notions of racial superiority and "racial purity" propagated by a slavery system. In Latin America, ethnic ascription has often followed a hyperdescent ethnic classification model in contrast to the hypodescent model observed in the US. This means that individuals of mixed heritage, such as Mestizos, are more commonly identified as White rather than Indigenous and are classified within the more socially dominant category of the parents' ethnicity or ethnic group (Iverson, 2022). The main ethnic groups in Colombia are White-Mestizos, Afro-descendants, and Indigenous, with ethnicity being self-reported. For mortality registration, the ethnic affiliation of the deceased is provided by family members. Individuals with multi-ethnic ancestry can identify with an ethnic group based on their social practices, traditions, and culture. Although ethnicity has a physiological component, it is fundamentally a socially constructed category rather than a biological determination.

Ethnic groups in Colombia exhibit particular characteristics in terms of geographical location, age at death, life expectancy, causes of death, and other socioeconomic conditions. Consequently, it is feasible to ascribe ethnicity to individual death records missing this information using machine learning applications. Machine learning provides a set of tools that generally perform better at predictive tasks than traditional statistical models, although they are more complex and less capable of making inferences about the underlying data-generating process (James et al., 2014; Hastie et al., 2009). These methods can classify ethnicity by training an algorithm to estimate the probability that a person with a specific vector of characteristics $\mathbf{X}$ belongs to a particular group. This approach is particularly relevant when ethnic information cannot be collected, as is the case with the Colombian deceased population before 2008. The identities estimated through this ethnic recognition methodology should be understood as the probable ethnic identity that a person would have reported if they were alive.

New methodologies of ethnic recognition based on demographic characteristics, epidemiological profiles, and spatial information have demonstrated positive results in ascribing ethnic origin to individuals (Kim et al., 2018; Elliott et al., 2008). Software algorithm methodologies for identifying ethnic identity have been successfully employed in various studies conducted in the Netherlands (Bouwhuis and Moll, 2003), the UK (Nanchahal et al., 2001), Germany (Razum O, 2001), the USA (Lauderdale and Kestenbaum, 2000), and Canada (Sheth et al., 1999). More recently, studies utilising names and surnames (name network clustering techniques) for ethnicity classification have reported positive predictive values (PPV) around 98% (Voicu, 2018; Imai and Khanna, 2016; Mateos et al., 2011). PPV is defined as $PPV =$ Number of True Positive/Number of True Positive + Number of False Positive, measuring the proportion of cases that are truly positive out of the total cases classified as positive (Akobeng, 2006). Therefore, PPV should be understood as the probability that a person identified as positive for an attribute or condition indeed possesses that characteristic.

The aim of the present analysis is to apply machine learning methods to investigate the feasibility of estimating the ethnic background of deceased populations for which ethnic information is not available, and to evaluate the biases and other problems that may be encountered when doing so.

7.2 Literature review

Software algorithms are playing an increasingly important role to inform decision making and to determine human decisions in different aspects of the social life (Burrel and Fourcade, 2021; Gillespie, 2013). Decisions on bank credits, job candidates selection, personalised advertisement, and potential persistent offenders are determined by software algorithms in modern societies (O'Neil, 2016). At the same time, there is a growing concern regarding the use and the "understanding" of ethnic and racial categories from algorithms and the introduction of racial bias into the algorithm process. Algorithms tend to incorporate in their estimations the beliefs and bias of the society in which they were created (Silva and Kenney, 2018; Winner, 1980) and in consequence task performed by machine learning algorithms could be racially biased, in the sense that they can lead to discrimination and over-estimation of negative behaviours in non-White population (Obermeyer et al., 2019; Chouldechova, 2017). It is possible that racial bias resulting from algorithms can not be prevented by omitting racial categories in input data; as race is a social constructed variable, it is related to the social environment in which the information is collected. Thus social segregation, spatial segregation, and social classes are already racially biased. This means racial biases are not introduced by the algorithm but probably result from the data itself: the algorithms are trained with biased information. For instance, members of minority groups might be less likely to be approved for credit even if they have the same characteristics as members of the majority group, simply because on average members of the minority group are less likely to be able to afford repayments, because of underlying economic disadvantage. This reasoning has led to suggestions that unsupervised machine learning models could give better results in terms of algorithm fairness (Benthall and Haynes, 2019).

Ethnic recognition through machine learning methods does not aim to assign ethnic identities based on arbitrary categories or biological characteristics. It seeks to use known information on epidemiological, demographic, and geographical locations of ethnic populations to estimate probabilities of group membership for individuals with those specific characteristics but for whom ethnic information was not registered. The model, therefore, does not aim to "predict" identities or future behavior. Instead, it seeks to identify the probable ethnicity of deceased individuals based on their specific social backgrounds and epidemiological profiles. From this perspective, the moral issue of potential racial bias may be considered to have fewer social implications compared to algorithms that attempt to predict, for instance, potential criminal behavior based on racially biased data.

Methodologies for ethnic recognition using machine learning algorithms are increasingly prevalent in various social contexts. The combination of forenames and surnames with geographical and demographic information has shown promising results in ascribing ethnic origin (Voicu, 2018; Imai and Khanna, 2016; Komahan and Reidpath, 2014; Mateos et al., 2011; Lakha et al., 2011; Macfarlane et al., 2007). Name-based ethnicity classification methods operate on the premise that the practice of assigning forenames and surnames is not random but is instead socially and geographically determined, reflecting an individual's ethno-cultural and geographical background. Consequently, it is possible to identify ethno-cultural patterns within a population based on names. For instance, individuals with British, Indian, or Turkish backgrounds can be grouped into ethno-cultural clusters, where names play a crucial role in their identification and classification.

Medical registers from general practices in England have been employed to identify the ethnic backgrounds of individuals using data related to religious affiliation, spoken languages, and names. The results demonstrate that more than 90% of individuals could be accurately classified, indicating levels of sensitivity and specificity[1] around 92% and 96% respectively (Macfarlane et al., 2007). Likewise, algorithms have been employed to predict morbidity patterns and classify phenotypic characteristics such as systolic and diastolic blood pressure, heart rate, and high- and low-density lipoproteins, achieving 100% accuracy for the prediction of certain diseases (Seffens et al., 2015). In these particular cases, race and ethnic information served as crucial predictors, as ethnic minorities are often linked to specific social and cultural environments, exposing them to particular disease risks. Similarly, models and algorithms using clinical records of patients have been used to predict mortality with notable success (Xia et al., 2012). Machine learning models have demonstrated satisfactory results in prediction and classification tasks, particularly when the outcome is a categorical variable. For mortality and morbidity prediction, these models generally outperform logistic regression models (Rahman et al., 2013). However, it has also been argued that the advantage of machine learning models is not significantly superior compared to more classical approaches for classification. Furthermore, the higher complexity and lower interpretability of these models make logistic regression models a viable alternative for better understanding the relationships among variables (Xue et al., 2019).

In racial and ethnicity recognition tasks, machine learning models have been used to impute information for individuals who did not report the ethnic background in their medical records. Racial information is considered a significant predictor of health conditions in epidemiological studies because these categories are assumed to be strongly correlated with socioeconomic status, access to quality education, and healthcare. Consequently, the relationship between morbidity and ethnicity has been analysed from a machine learning perspective using demographic and anthropometric data. For instance, Kim et al. (2018) used medical histories and clinical information, such as age, gender, and codes from the International Classification of Diseases, Ninth Revision (ICD-9), to ascribe ethnicity to patients.

Machine learning models in these contexts can handle large and complex datasets, identifying

[1] Assuming that a classification task involves identifying points that possess a determined characteristic (positive cases) from those that do not (negative cases), sensitivity and specificity are critical metrics. Sensitivity, defined as $\text{Sensitivity} = \text{Number of True Positive}/(\text{Number of True Positive} + \text{Number of False Negative})$, refers to the proportion of positive cases that were correctly classified (true positive cases) out of the total positive cases in the sample. Specificity, defined as $\text{Specificity} = \text{Number of True Negative}/(\text{Number of True Negative} + \text{Number of False Positive})$, measures the proportion of negative cases correctly classified (true negative cases) out of the total negative cases in the sample. Sensitivity and specificity, therefore, assess the model's ability to correctly identify positive and negative cases, respectively.

patterns and making predictions that may not be evident through traditional statistical methods. Results of the supervised machine learning model yielded on average accuracy of 0.668, and cross-entropy loss of 0.857 on test data which outperformed the alternative logistic regression model[2]. Supervised learning models are predictive models that have been trained with data for which the correct answers are provided. Through this training process, the models improve their predictive power and learn to make more accurate predictions for similar tasks.

Machine learning procedures show a high level of accuracy in classification tasks. However, problems of misclassification could arise if the algorithm has not been correctly trained and does not consider all the possible variations such as in the case of names variations (Brant and Boxall, 2009), or if geographical location as a predictor is not clearly defined among the analysed units. In that case geographical location of individuals could make little contribution to the model while age and gender could be more relevant to identify the ethnicity of specific groups (Adjaye et al., 2014).

7.3 Data and methods

7.3.1 Data

We analysed $22,064$ deaths from the year 2016. The dataset was divided into training and testing datasets, see Figure 7.1. The training data contains $15,446$ deaths from which $14,420$ were White-Mestizos, 783 Afro-descendants, and 243 were Indigenous. The testing dataset contains $6,618$ deaths, $6,180$; 335; and 103 were White-Mestizos, Afro-descendants, and Indigenous respectively. The available information of the deceased person included the age at death, municipality and department of residence, place of death such as hospitals, at home, at work, on the street, etc.; the marital status, cause of death based on *International Classification of Diseases, Tenth Revision* (ICD-10), educational level such as primary or secondary school, years of education, and access to health insurance. Information regarding year of education allows us to differentiate between persons with the same educational level but with a different number of years of education completed.

Mortality records began to include ethnic information in the year 2008; however, as shown in Section 3.3, ethnic information in the initial years presents considerable variability, for which reason the training data for the model is sourced from the year 2016. This is crucial because the training dataset, which includes ethnic information, enables the algorithm to "learn" the specific

[2]The prediction error in machine learning models is commonly quantified using a loss function, with cross-entropy being one of the most popular options. This function measures the level of "disorder" or uncertainty in predictions comparing the "loss" associated with the predicted and the estimated distribution. Formally defined as $D = -\sum_{k=1}^{K} \hat{P}_{mk} \log \hat{P}_{mk}$, where $\hat{P}_{mk}$ represents the proportion of observations from the kth class (James et al., 2014, p.312). Cross-entropy loss values range between 0 and 1 and the lower the loss function, the more accurate the model. If all the observations are from the same class, low cross-entropy measures will indicate low uncertainty and higher accuracy in predictions. In this context, accuracy is understood as the number of correct predictions divided by the total number of predictions. Further details about accuracy in machine learning models can be found in Section 7.3.2.

characteristics of each ethnic group. Subsequently, the algorithm will be able to "recognise" the ethnicity of an individual when this information is unavailable. Figure 7.1 illustrates the ethnic identification process with the machine learning algorithm. First, the algorithm uses a training dataset to "familiarise" itself with ethnic characteristics. Second, the trained algorithm is applied to a test dataset without ethnic information to assess the model performance in terms of its capacity to correctly identify the ethnic identity of a deceased person.

Figure 7.1: Machine Learning for Ethnic Classification based on Mortality Registers

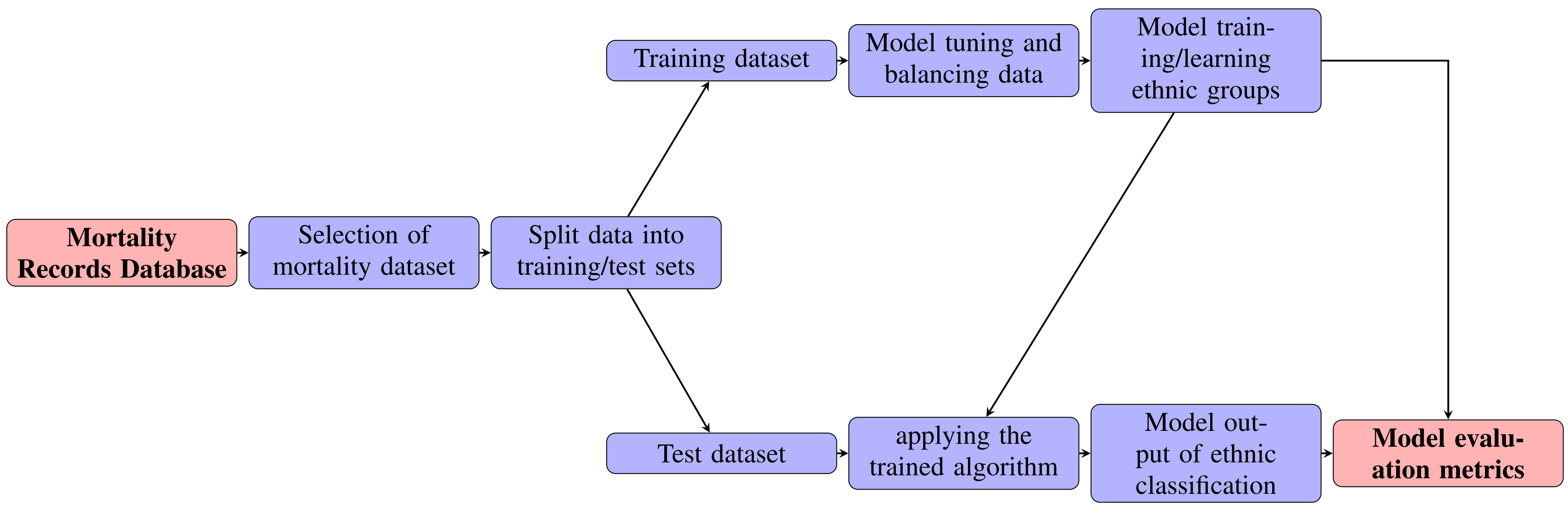


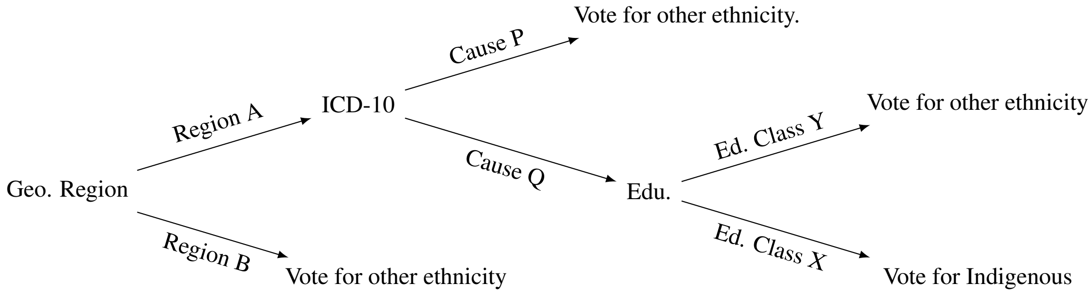
Figure 7.1 shows the summary of the machine learning estimation process and the relevance of the training and test data sets.
Source: Own elaboration.

7.3.2 Method

In this analysis, we propose the use of a random forest (RF) model to classify the ethnicity of deceased people based on information related to the cause of death, geographical location, and demographic and socioeconomic characteristics of the person. The proposed classification model is a supervised learning model in which the aim is to devise a method or construct a rule for assigning characteristics to one of a set of racial categories on the basis of a vector of variables. The information on which this rule is to be based is a training dataset with known variables and classifications through which it estimates the probability that a point or characteristic belonging to a particular class indicates strength of class membership (Hand and Till, 2001). The Random Forest model proposed by Breiman (2001) belongs to the class of ensemble machine learning algorithms (Hastie et al., 2009). Such methods rely on the observation that by taking an average of a set of weak predictors, we can often outperform other methods based on a single stronger individual predictor. In the case of a random forest, classification or regression trees are the weak predictors that make up the ensemble. Such trees work by recursively splitting the training sample based on covariates, so that each split results in the maximally homogeneous groupings of the sample. These splits occur in a graph or tree-like structure, with each node in the graph representing a split. To predict, at each split point it is determined which 'branch' an observation falls within. At the terminal point of the tree (the 'leaf node'), regression predictions can be determined by average value of training observations in the leaf. In the classification tasks, the most frequent class is taken as the prediction for that observation.

Training a RF involves training a large number of regression tree predictors such that each tree depends on a random resampling of the data with replacement. The prediction of the RF is simply the average or modal prediction across the trees in the forest. Figure 7.2 illustrates the decision-making process of a tree at each node or variable until it reaches a final decision -vote- for one of the categories under consideration. This method has demonstrated strong performance across various prediction tasks. For instance, in classifying ethnicity, the RF employs classification trees that iteratively split the dataset and subsequently predict that each observation belongs to the most commonly occuring ethnic class among the training observations (James et al., 2013, p.311). In other words, if Indigenous ethnicity is the most frequent class within a particular leaf node of the training dataset, then the tree will assign a high probability of belonging to the same ethnic class to observations in the same node.

Figure 7.2: Decision Tree Method to Determine Indigenous Ethnicity



Figures 7.2 shows the data splits at various nodes: Geo. Region (geographical region), ICD-10 (cause of death from International Classification of Diseases version 10), and Edu (Educational level). The tree of Figure 7.2 estimates probabilities of a person with specific information regarding region, cause of death, and educational level being Indigenous or not. The estimation of these probabilities is based on the training dataset. In a RF model, for each node with a tree, the algorithm selects at random $m = \sqrt{p}$ variables, where p represents the total number of variables in the model. A splitting variable is then chosen from among these m variables only, with the criterion for choice being the homogeneity of the resulting sample groupings created by the split. The RF grows thousands of trees, with the final ethnicity classification being determined by the average votes across all trees.
Source: Own elaboration.

Formally, the tree-growing process through random selection of input variables can be presented as follows: assuming that a trained RF model grows a number of trees $1, 2, ..., k$, with each tree characterised by independent and identically distributed (i.i.d.) random vectors $\Theta_1, \Theta_2, ..., \Theta_k$. Each tree is grown using the training set and the random vector Θ_k, resulting in a classifier $h(\mathbf{X}, \Theta_k)$, where $\mathbf{X}$ represents an input vector (Breiman, 2001). Here, Θ_k defines the kth random forest tree in terms of split variables, cutpoints at each node, and terminal-node values (Hastie et al., 2009, p.598). The mechanism through which all the generated trees collectively vote for the most popular class is known as the random forest approach.

Breiman (2001, p.6) defines the RF model as a classifier consisting of a collection of tree-structured classifiers $\{h(\mathbf{X}, \Theta_k), k = 1...\}$ where the $\{\Theta_k\}$ are i.i.d. random vectors and each tree cast a unit vote to determine the most popular class at input $\mathbf{X}$. Similarly, Hastie et al. (2009, p.587) defines the RF process as a substantial modification of bagging[3] that builds a large collection of *de-correlated* trees which are averaged.

The presentation of the RF algorithm in Table 7.1 follows the estimation process outlined by (Hastie et al., 2009, p.588). Similarly, Figure 7.1 illustrates the general process of machine learn-

[3]If a single training data set is repeatedly sampled to generate B different bootstrapped training data sets, and an algorithm is trained on each of the obtained subsets from $b = 1$ to B to produce different predictions. Then the process of averaging all the predictions to obtain a more reliable estimate is called bagging (James et al., 2013, p317)

Table 7.1: Estimation of the Classification Random Forest

1. The process below is repeated for each of the bootstrap samples from $b = 1$ to B. In our case, the training datasets were drawn under stratified sampling as well as weighting sampling
   (a) A bootstrap sample $\mathbf{Z}^*$ of size N is taken from the training data
   (b) Grow a random forest tree T_b using this sample until the minimum node size n_{min} is reached. Trees are grown by repeating the steps below, with each iteration creating a 'node' in the tree.
      i. Select m variables at random from the p total variables of the model, see Figure 7.3. The default value for m in a classification model is $m = \sqrt{p}$.
      ii. Determine the best variable split-point among the m variables. Best in this context refers means the split of the dataset that results in more homogeneous (or pure) mixtures of classes in the resultant data subsets. For example, it might be determined that by splitting the dataset into two sections using a specific cutpoint x_1^* for a continuous variable x_1, the half of the dataset defined by values of $x_1 < x_1^*$ consists entirely of observations of indigenous deaths, while the other contained no indigenous values - in which case this split would be optimal for predicting indigenous deaths.
      iii. Split the node into two daughter nodes and repeat the two steps above for each of these daughter nodes until the minimum node size is reached.
2. Output the ensemble of trees $\{T_b\}_1^B$. From $b = 1$ to B.
   The prediction in the classification model at a new point x will be determined as follow: Let $\hat{C_b(x)}$ be the class prediction of the bth random-forest tree. Then the class prediction of the RF model will be $\hat{C}_{rf}^B = \ majority\ vote\ \{\hat{C}_b(x)\}_1^B$.

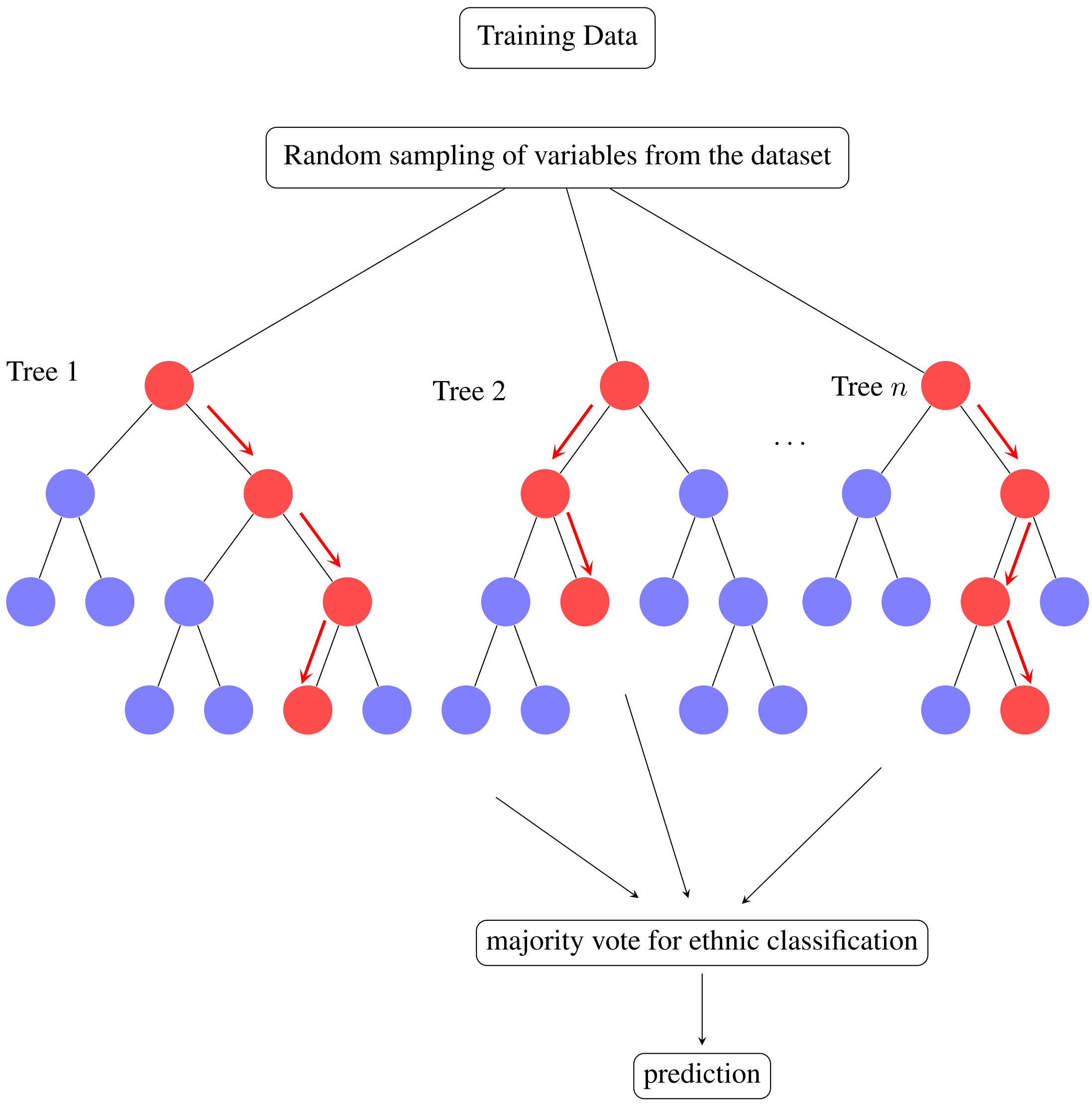
ing classification. The RF model, as shown in Figure 7.3, is an effective tree-based classification model. Unlike a single decision tree, RF enhances prediction accuracy by employing multiple decision trees, averaging their decisions, and making final decisions based on the majority of votes for a predicted class, see Figure 7.3.

The accuracy of the model is given by the margin function (Breiman, 2001, p7):

$$mg(\mathbf{X,Y}) = average_k\, I(h_k(\mathbf{X}) = \mathbf{Y}) - \max_{j \neq Y} average_k\, I(h_k(\mathbf{X}) = j) \tag{7.3.1}$$

Equation 7.3.1 represents the margin function which evaluates a single point $(\mathbf{X,Y})$ in the dataset distribution by computing the difference between the proportion of votes for the correct class $\mathbf{Y}$ and the maximum proportion of votes for any incorrect class j. The margin function can be conceptualised as a collection of classifiers denoted by $h_k(\mathbf{X})$, with $I(\cdot)$ representing the indicator function that determines the membership of an element to one of the predefined categories. *Therefore, the margin function measures the extent to which the average number of votes at $(\boldsymbol{X}, \boldsymbol{Y})$ for the correct class surpasses the average vote for any other class. Specifically, a positive margin under majority voting indicates correct classification ($mg(\boldsymbol{X,Y}) > 0$), while a negative margin*

Figure 7.3: Structure of a Random Forest Model



Figures 7.3 presents the structure of the RF model. Decision trees are built upside down, and they are made up of branches (connectors) and leaves (nodes). For every observation that falls into a specific region or leaf, the algorithm will make a decision based on the mean response values obtained from the training process from observations within the same leaf node. The red nodes represent the path how decisions were made until arriving at a terminal node.
Source: Own elaboration.

signifies incorrect classification ($mg(\boldsymbol{X},\boldsymbol{Y}) < 0$) (Breiman and Cutler, 2022, p.9).

The supervised classification is based on a set of objects with known classifications, and estimates the probability a point belongs to an specific class or group. There are many indicators to compare the performance of a supervised classification. Two of the most popular approaches are, first, to measure the misclassification or error rate; second, to estimate the area under the Receiver Operating Characteristic (ROC) curve. Misclassification rate refers to the proportion of cases that were classified incorrectly. However, as the cost of misclassification is rarely known (Bradley, 1997, p1146), "an alternative classification rule is to compare the overall distribution of $\hat{p}$ over different values of threshold t for objects in classes 0 and 1, and the area under the ROC curve (AUC) will measure the difference between these two distributions" (Hand and Till, 2001, p172). The threshold here refers to the value at which the predicted probabilities of belonging to a particular class are converted to an actual assignment to that class, so that a threshold of 0.3 would mean that a predicted probability of an observation belonging to class 1 of greater than 0.3 would result in the model assigning that observation to 1. "Let $\hat{p}(x)$ be the estimated of the probability that an object with measurement vector x belongs to class 0, let $f(\hat{p}) = f(\hat{p}(x)|0)$ be the probability function of the estimated probability of belonging to class 0 for class 0 points, and let $g(\hat{p}) = g(\hat{p}(x)|1)$ be the probability function of the estimated probability of belonging to class 0 for class 1 points. Let $F(\hat{p}) = F(\hat{p}(x)|i = 0)$ and $G(\hat{p}) = G(\hat{p}(x)|i = 1)$ be the corresponding cumulative distribution functions. Then the ROC curve is defined as a plot of $G(\hat{p})$ on the vertical axis, against $F(\hat{p})$ on the horizontal axis. Clearly this plot lies in a unit square. A goodclassification rule is reflected by a ROC curve which lies in the upper left triangle of the square. This follows since any point above the diagonalcorresponds to a situation in which $G(\hat{p}) > F(\hat{p})$, so that the class 1 points have *lower* estimated probability of belonging to class 0 than do the class 0 points" (Hand and Till, 2001, p173). The ROC curve can also be understood more intuitively based on considering that the true positive (TP) values are displayed on the Y axis, and the false positive (FP) values on the X axis. See Figure 7.4.

Figure 7.4: Assesment of ROC Performance

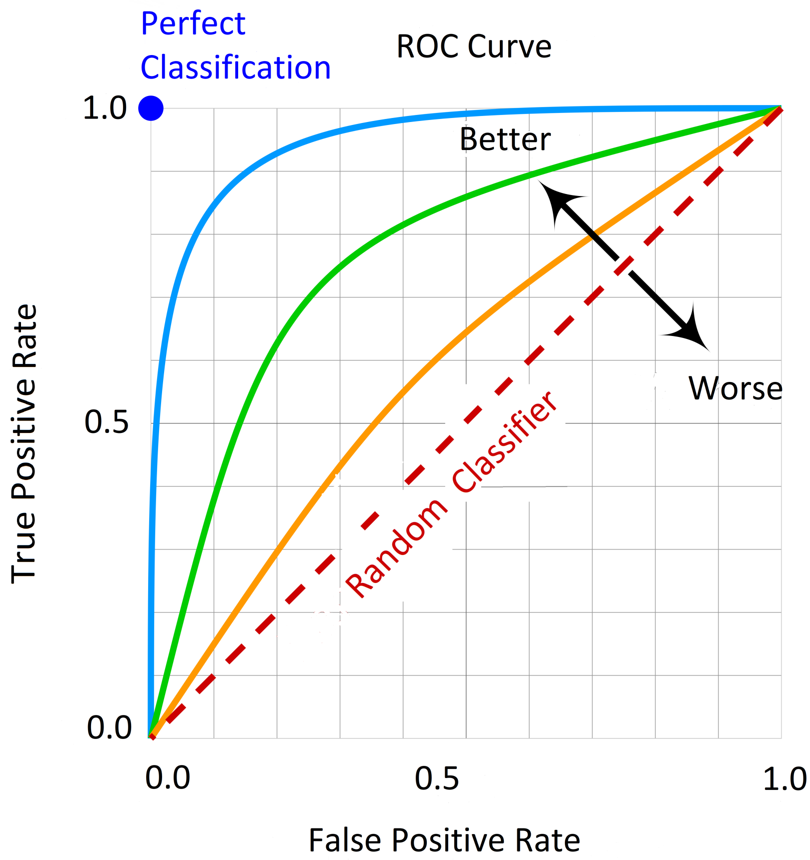


Figure 7.4 shows the true positive and the false positive rates. The diagonal line represents a random classifier while the point (x= 0, y = 1) represents the perfect classifier.
Source: (Thoma, Martin, 2018).

Equivalently, the y and x values can also be understood as sensitivity and specificity, where the latter values are reversed so that the specificity decreases as we move from left to right. Each classifier defined by a different threshold value t is represented by the point in ROC space corresponding to its (FP,TP) coordinates, in which the point $(0, 1)$ correspond to the perfect classification and the diagonal line $y = x$ represents the strategy of randomly guessing the class (Provost and Fawcett, 2001, 207).

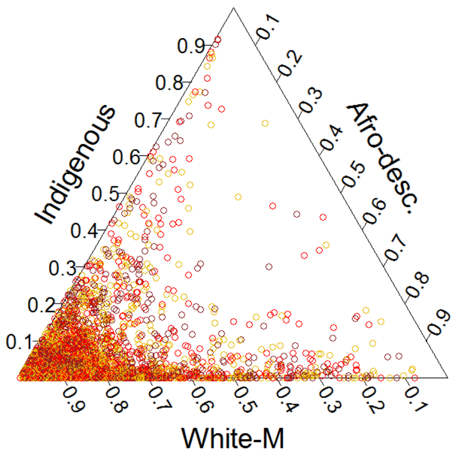
The ethnic classification task can be understood as a supervised machine learning problem that intends to predict a missing variable based on a training dataset in which the missing variable distribution is well-known. In other words, the objective is to determine the mutually-exclusive ethnic identity of deceased individuals using a trained algorithm that incorporates information on epidemiological, geographical, and socio-demographic characteristics, including age, gender, municipality of residence, cause of death, educational attainment, and others. Crucially, the random forest is able to learn complex non-linear combinations of these variables, and thus is able to take account of known features of the structure of deceased sample. As demonstrated in Chapters 5 and 6, attributes like age and gender, particularly when analysed alongside causes of death and other covariates, emerge as potent predictors of ethnicity. For instance, the higher prevalence of homicides among Afro-descendant men aged between 15 and 35 highlights this correlation. But because of the ensemble structure, the algorithm does not rely only on these relationships.

For developing an algorithm to identify the ethnicity of Colombian deceased population, we used ethnic mortality information from 2016 in order to build an algorithm able to learn a representation of an individual with a specific ethnic identity. This will allow the algorithm to estimate probabilities that an individual truly belongs to a particular ethnic group given a vector of characteristics. This algorithm however, could be negatively affected by the fact that 90% of the training data belong to a particular ethnic class, which generates an unbalanced problem (Blagus and Lusa, 2013; Chawla et al., 2002). For this reason, we applied extensions of the RF classification model in which we solve the unbalance problem in two ways. First, we use a weight factor in order to over-sample the information from Indigenous and Afro descendant groups. Second, we propose the use of stratified sampling in order to randomly select the same number of observations from all ethnic groups. This means that information from the White-Mestizo group was under-sampled to the level of Indigenous and Afro-descendants. These procedures will equalise the proportion of individuals from the different ethnic groups in the training dataset, in such a way that the training data will have a similar level of information from all ethnic classes.

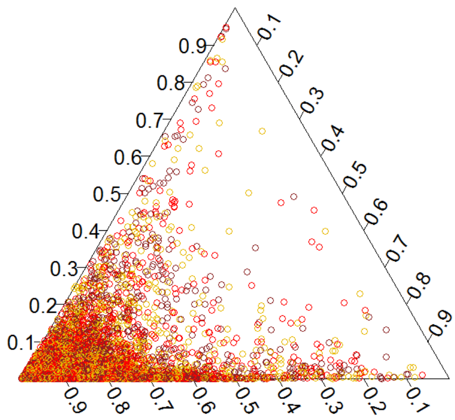
7.4 Results

We run a RF model with two different correction methods for the unbalanced number of individuals from White-Mestizo, Afro-descendant, and Indigenous ethnicity in the sample. Figures 7.5 and 7.6 show the effect of correction through sample stratification and weighting methods respectively. These methods improve the effective ratio of the minority ethnic classes in the training dataset and avoid possible bias problems due to the disproportionate numbers of one particular ethnic category in the training data. The results show a slight difference between the RF with stratified sampling and the RF with weighted sample. In both versions of the RF, the relative proportion of ethnic minority groups in relation to the White-Mestizos is increased until a more equal representation of all ethnic groups in the sample is achieved. The triangular graphs are ternary plots that allow the plotting of observations for which sets of values much sum to one such as class probabilities predicted by a random forest (see Schöley and Willekens (2017) for other uses in demographic analysis). Each point in the plot represents a random forest prediction of one observation of the training data. In these plots, the closer the point to top corner of the triangle, the higher the predicted probability that the corresponding observation belongs to the Indigenous group. Similarly, those points close to the left corner are predicted to be more likely to belong to the White-Mestizo group, and the right corner indicates high predicted probabilities for the Afro-descendant group. The centre point with coordinates $[1/3, 1/3, 1/3]$ represents equal probability of being assigned to any class.

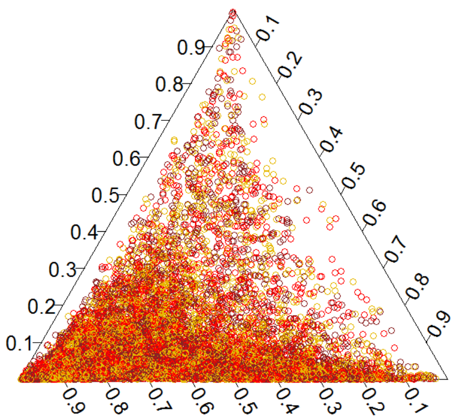
Figure 7.5: Sample Stratification for the Random Forest Model



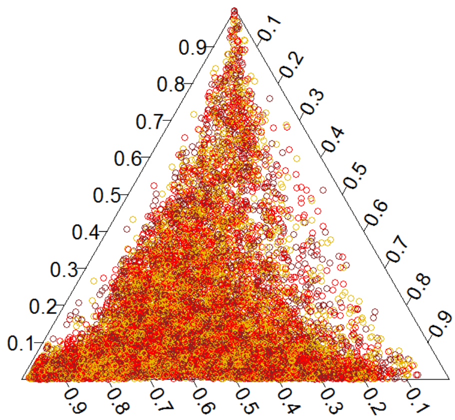
(a) Without Stratification



(b) 30: 3: 1



(c) 3: 3: 1



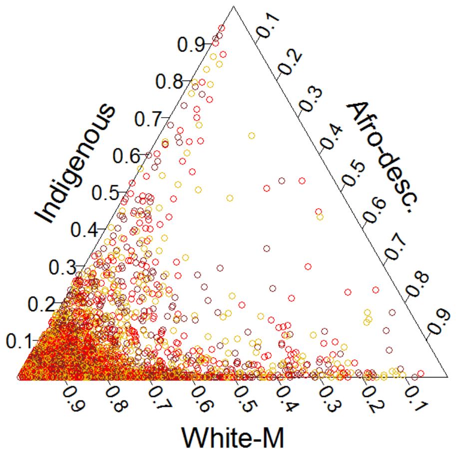
(d) 1: 1: 1

Figures 7.5a, 7.5b,7.5c, and 7.5d show ternary plots with the ethnic composition of the sample. Own elaboration.

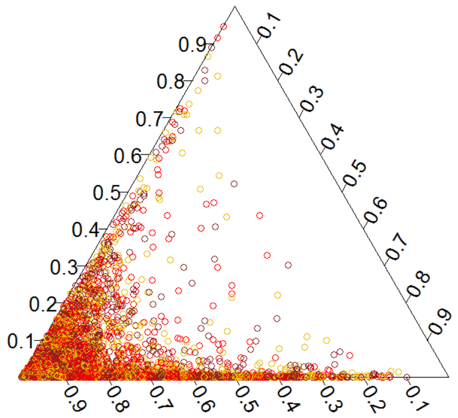
In Figure 7.6a, the observations are concentrated around the point $(1, 0, 0)$ at the bottom left of the graph, indicating that nearly 100% of the data is predicted to belong to the White-Mestizo class. The contributions from Afro-descendants and Indigenous groups are nearly negligible in relative terms. As the proportions of Afro-descendants and Indigenous individuals are increased in the sample through stratified sampling, the predictions shift towards the other corners of the triangle. Conversely, in Figure 7.6d, the data are centered in the ternary plot at the intersection point where each ethnic group contributes 33.3% to the sample, summing up to 100%. Figure 7.5a shows the data without stratification, while Figure 7.5b shows a class ratio of 30 White-Mestizos, 3 Afro-descendants, and 1 Indigenous after stratified sampling. Figure 7.5d showss a balanced dataset where the numbers of White-Mestizos, Afro-descendants, and Indigenous individuals are equal, with a ratio of 1:1:1.

The different approaches for correcting data imbalance affect the speed of error convergence for ethnic classification. Figure 7.7 illustrates the classification error level for each ethnicity using both the weighted and stratified Random Forest (RF) methods. Initially, the error is nearly 100% for Indigenous and Afro-descendant groups and 0% for White-Mestizos. However, as the sample composition becomes more balanced for ethnic minorities, the error converges. The algorithm thus becomes more adept at identifying the specific patterns of Indigenous and Afro-descendant individuals given the increased level of information available, while the accuracy for White-Mestizos decreases as their proportion in the sample diminishes. Consequently, the classification error significantly decreases for Indigenous and Afro-descendant groups while it increases for White-Mestizos as their relative proportion in the sample declines.

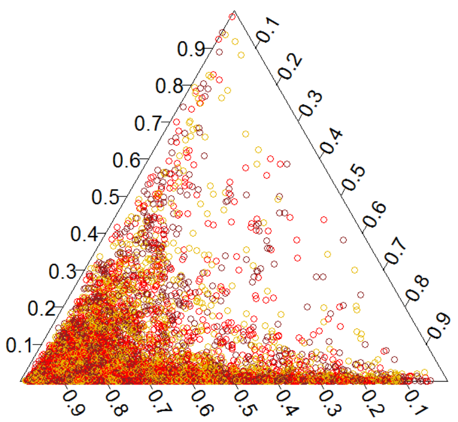
Figure 7.6: Sample Weighting for the Random Forest Model



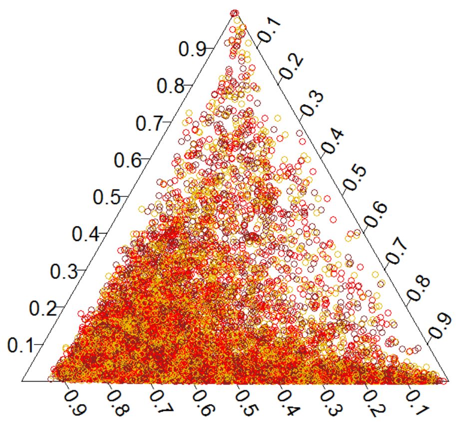
(a) Without Weights



(b) 0.1: 100: 1000



(c) 0.01: 10000: 1e+05

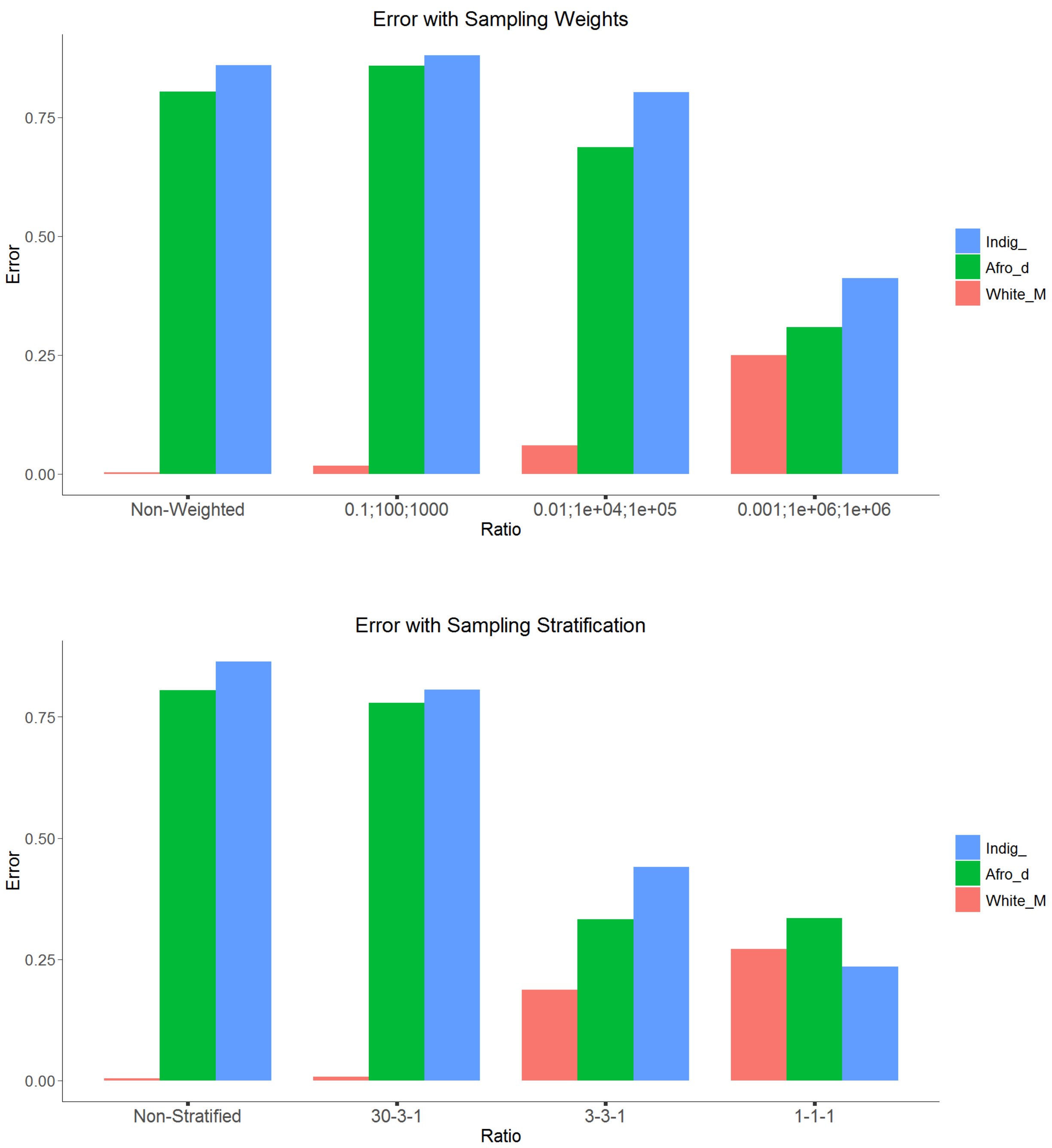


(d) 0.001: 1e+06: 1e+06

Figures 7.6a, 7.6b,7.6c and 7.6d show the changes in the ethnic composition of the sample when different sampling weights are applied. Figure 7.6a shows a sample composition with majority of White-Mestizo people. The inferior corner on the left of the plot represents the intersection point of 100% White-Mestizo, 0% Afro-descendants, and 0% Indigenous. Figure 7.6d shows the most harmonised composition in which the proportion of White-Mestizos, Afro-descendants, and Indigenous in the sample data is relatively similar.

Own elaboration based on Colombian mortality record data.

Figure 7.7: Accuracy of Ethnic Classification with Different Class Ratios and Weights



The accuracy of the algorithm in recognising the ethnicity of deceased people improves when the number of deaths from all ethnic groups is more balanced in the training dataset. Figure 7.7 shows the error for ethnic classification based on sampling from Figures 7.5 and 7.6. The model without stratification or without weights presents low class error for the White-Mestizos while Indigenous and Afro-descendants categories have a level of error around 100%. This means that ethnic recognition for these later groups is completely inaccurate. When sampling weights and sampling stratification are applied, the classification error decreases for Indigenous and Afro-descendants groups while increases for White-Mestizos as these methods level up the proportion of observations for all groups.

Own elaboration based on Colombian mortality record data.

Figure 7.8 presents the ROC curves for the different ethnic groups. This curve plots the rate of true positive values against the rate of false positives. The closer the ROC curve gets to the top left corner of the plot, the better the model does at classifying the data into ethnic categories. The indigenous group shows higher values of the ROC curve and closer to the point $(0, 1)$ of the graph. The higher rate of sensitivity of the Indigenous ethnicity at different threshold levels in both models means that the true positive rate or the proportion of people correctly classified is higher in this group. White-Mestizos and Afro-descendants show lower ROC curves; however, at different points in the Weighted RF, the ROC curves present only slightly differences among ethnic groups. The results in general indicate that the RF model recognises the right ethnicity of deceased individuals with probabilities higher than 80%.

The Area Under Curve (AUC) score in Figure 7.8 should be understood as indicating the model's ability to classify ethnicity correctly. "The AUC is commonly described as the probability that a random individual from the diseased population is more likely to have a higher predicted risk than a random individual from the non-diseased population" (Janssens and Martens, 2020, 1401). Alternatively, it can be interpreted as the expected true positive rate, averaged over all false positive rates (Flach et al., 2011, p.1). A model whose predictions are 100% correct has an AUC of 1.0 and vice versa. In the case of the indigenous group, the AUC of 0.864 represents the probability that a random individual from the indigenous group is more likely to be predicted as Indigenous than a random person from a non-Indigenous group. This indicates a reasonably high classification power of the model. Notably, this probability is higher in the Stratified RF model, which has an AUC of 0.913. Therefore, the probabilities of being classified correctly are higher for Indigenous individuals compared with other groups. Conversely, the lowest predictive ability is observed for the Afro-descendant group, with AUC scores of 0.816 and 0.811 for the Weighted and Stratified RF models, respectively.

The Stratified RF compared to the Weighted RF presents slightly higher AUC scores for ethnic classification except for the Afro-descendant category. The variables with high importance that most contribute to correct ethnic classification are the department of residence, age at death, cause of death, and the municipality of residence. Figure 7.9 shows all the variables used in the model with their contributions in terms of mean decrease in Gini which measures how much more similar observations within splits at a node become if we use variable x to split rather than another variable. In other words, it measures the total decrease in node impurities from splitting on the variable, averaged over all trees (Nembrini et al., 2018; Han et al., 2016). Therefore, the most important variables for providing correct ethnic classification will contribute with the largest mean decrease in Gini.

The most important variable for the RF classification model is the department that makes reference to the geographical location, and the least important is the place of death of the person (hospital, home, or other place of death). The contribution of the variable Department is about 5 times greater than the contribution of the variable Place of Death in both versions of the RF model.

Figure 7.8: ROC Curve by Ethnicity

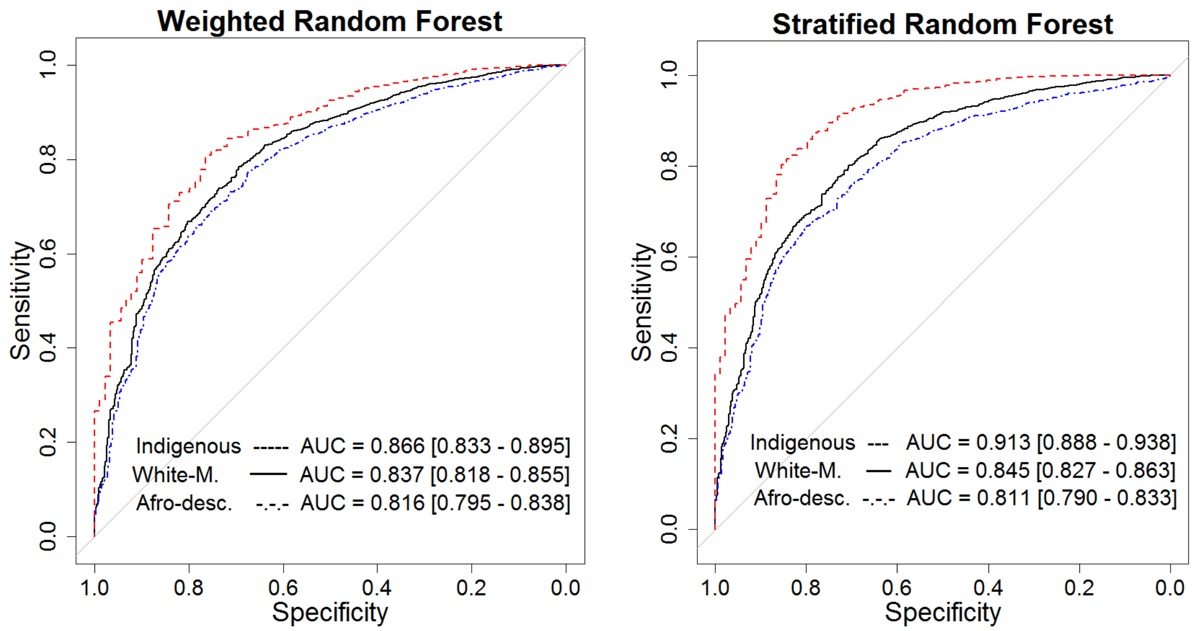
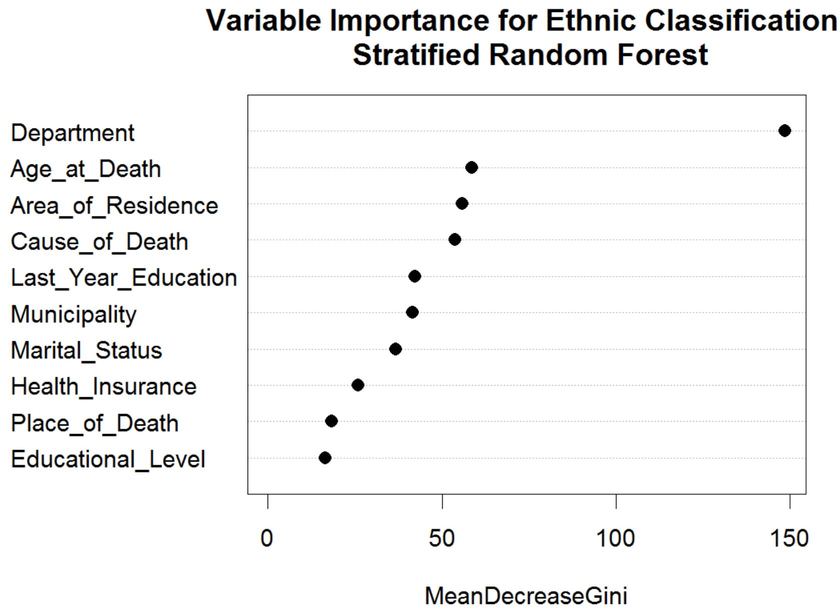


Figure 7.8 presents the ROC curves for ethnic classification. The ROC curve presents the sensitivity of the model (true positive values) plotted against the specificity (false positive values). The higher values in sensitivity in the Indigenous group (dashed line) means that this group has a higher rate of correctly classified cases. White-Mestizos (solid line) and Afro-descendants (dotted line) present lower values of sensitivity at different thresholds. AUC scores for Indigenous and White-Mestizos are higher in the Stratified RF compared to the alternative model, while Afro-descendants presents a slightly advantage in the Weighted RF.
Own elaboration based on Colombian mortality record data.

However,the variable Area of Residence has a higher importance in the weighted model than in the stratified version. In general, the models present very little differences in variable importance structure. The higher values in the x-axis of the weighted model are due to the effect of weights, but the relative difference among the variables remain more or less the same in both models.

Figure 7.9: Variable Importance of the Stratified and Weighted Random Forest Models



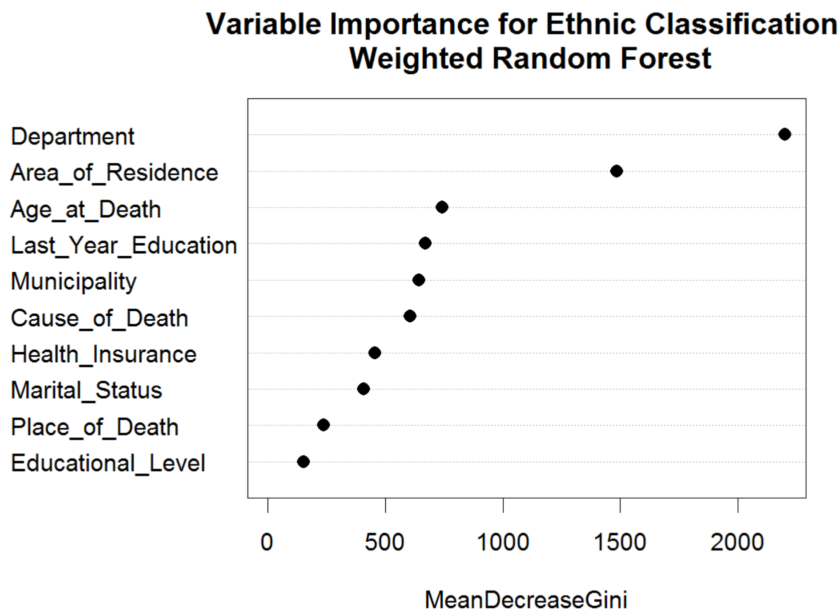


Figure 7.9 shows the contribution of each variable in terms of mean decrease in Gini which measures the rate of incorrect classification. The variable Department in that sense is the information with the highest contribution to the accurate classification of ethnicity in the Colombian context. The importance of the contribution from all the variables shows marginal changes between the two models, but the general order of the variables is very similar. For instance, the five most important contributors to ethnic classification accuracy are the same in both models, but with slight changes in the order.
Own elaboration based on Colombian mortality record data.

7.5 Discussion

The most relevant problem in social research on ethnic minorities is the problem of data quality and data availability. The outcomes of this research provides a significant contribution in addressing the limitations of ethnicity data and revealing ethnic identities when this information has not been collected. The use of experimental methods in mortality analysis such as machine learning models has shown that the probabilities to correctly identified the ethnic identity of a person are over 80%. In the Colombian case this probability reaches as high as 90% for the indigenous population.

Indigenous people show higher probabilities of being correctly classified because their characteristics compared to White/Mestizos and Afro-descendants are more particular. This explains for example, why geographical information about the Department is the most important variable for ethnic classification. Indigenous communities display a higher level of segregation with clearer borders. Although this phenomenon of segregation exists in Afro-descendant communities too, there has been a historical migration and urbanisation process of these communities towards the main cities of the country, so that the relationship between geographical location and Afro-descendant identity has been blurred over time.

The marginally lower difference in probabilities between White-Mestizos and Afro-descendants could be attributed to the fact that a significant proportion of Afro-descendant people reside in the same areas as White-Mestizos. This reduces the effectiveness of geographical location as an ethnic classifier for these groups while increasing its relevance for the Indigenous population. Similarly, information about area of residence and municipality plays an important role, likely being more pertinent for the Indigenous category than for the other groups. Information regarding age at death and cause of death are also among the main classifiers. This indicates that there are significant differences in these aspects among ethnic groups.

Information on marital status, education, health insurance, and place of death surprisingly show lower contributions in ethnic classification compared with other variables. There are important reasons to believe that among ethnic groups there exist considerable differences in socioeconomic aspects, but these variables present serious quality problems in the mortality registers and the level of missingness is relatively high in some ethnic groups. This could be one the reason for the low predictive power of those variables in the model.

The model presents some limitations in the case of multi-ethnic families as well as the case of persons located in unusual geographical areas for their ethnicity. In those case, it is highly probable that the model will misclassify the person into the most common ethnicity of the family or ethnicity of the geographical area. This misclassification error could be reduced if future work can incorporate information related to names of deceased people as names and surnames of White-Mestizos, Afro-descendants, and Indigenous people have an specific cultural background that will allow a higher sensitivity in the model.

The implications and potential extensions of the ethnic classification model represents a major contribution to ethnic studies and to the analysis of health inequalities by race and ethnicity, principally in countries where such data are not being collected. In Colombia, the outcomes of this analysis provide directions for new research and the better understanding on social inequalities. The results also represent an important contribution in terms of solving balanced class problems in ethnic data, and particularly the reconstruction of historical data that allows the analysis of social inequality trends over time and the design of more effective social policies.

CHAPTER 8

Conclusions

8.1 Main findings

A significant outcome of this study is the stark contrasts and inequities observed in mortality patterns and causes of death among different ethnic groups in Colombia. The principal findings of this thesis can be summarised into three key aspects where health disparities are noticeable: mortality under-registration, trends and patterns of mortality and life expectancy, and causes of death. Additionally, the analysis incorporates machine learning techniques to better understand the completeness of ethnic information in mortality registers.

The study documented how White-Mestizos, Afro-descendants, and Indigenous are characterised by regional segregation with considerable social and health inequalities. The poorest regions of the country inhabited by Afro-descendants and Indigenous groups demonstrate evidence of poor health indicators. More importantly, the health differences amongst ethnic minorities have widened over time in recent decades, when compared to White-Mestizos.

In Chapter 4, it was demonstrated that social inequalities and discrimination against Colombian Indigenous and Afro-descendant communities during the lifetime will persist even after death in the form of higher mortality under-registration compared to White-Mestizos. Deaths among ethnic minority groups are registered at lower rates, with under-registration levels particularly pronounced among children, individuals with lower education levels, and residents of rural areas. The findings highlight that the age, gender, and educational level of the head of household play a significant role in the determining the level of under-registration of deaths. Children born in the Indigenous rural communities whose parents are relatively young adults with lower education are at highest risk to be non-registered upon death. The results indicate that ethnic minorities contribute in larger proportions to the total number of unregistered deaths, attributed to factors including the disadvantages living in remote areas, cultural differences in terms of language restrictions, social discrimination, lack of importance attached for death registration, and evidently socioeconomic differences and the absence of the State in regions with higher proportions of ethnic minorities. In addition to the above, the study found an inverse relationship between Under-registration and age at death which is partly explained by a "wealth effect". This works in the form of incentives for registering death at older ages when inheritance and monetary compensation takes precedence. Individual socioeconomic differences determine either better or worst chances to be registered after death. The problem of mortality underestimation is much more complex in Indigenous compared with Afro-descendant communities. However, the biased mortality statistics in both groups represent a major barrier to measure and quantify ethnic mortality inequalities in Colombia.

In Chapter 5, the varying patterns of mortality differences by age, sex and ethnicity where investigated. According to Colombian official vital statistics, White-Mestizos show higher mortality and lower life expectancy when compared to those for Indigenous and Afro-descendants groups. However, this study has shown that the situation reverses when mortality data are corrected for under-registration. Life expectancy at birth for Indigenous is 66.7 years, for Afro-descendants 71.5, and for White-Mestizos 78.9 years. White-Mestizos live on average 12.2, and 7.4 years longer than Indigenous and Afro-descendants respectively. The higher mortality in ethnic minority groups is consistent whether it is estimated using census mortality data or with corrected mortality registers: both datasets show similar results. Additionally, the dispersion of the age of death in ethnic groups have a higher variance across all ages, while age of death in White-Mestizos is concentrated in the older age groups.

Chapter 5 also revealed that life expectancy inequalities for Indigenous and Afro-descendants have their roots across different periods of the lifecycle. In the case of Indigenous, the differences begin in childhood, while in Afro-descendant the differences are concentrated amongst children and young adults. The findings confirm that for the period 2008-2019, the differences in mortality rates between White-Mestizos and Indigenous have moderately increased, whereas that for White-Mestizos and Afro-descendants shows a considerable reduction of the mortality gap. However, the results for Afro-descendants should be interpreted with precaution as mortality data from Afro-Colombians were under-estimated in both the 2018 Census and mortality records.

In addition to estimating mortality differences among ethnic groups, in Chapter 6 the underlying causes of death were analysed. Three clear patterns based on cause of death, age, and ethnicity were identified. The first is related to the White-Mestizo population, in which the main causes of death are degenerative diseases such as cardiovascular diseases and Cancer, as well as respiratory diseases. The age group most affected is the population at older ages, where the majority of the deaths by these causes occur. The second pattern is related to Afro-descendants, and combines a very high mortality of young men between age 15 and 40, principally by homicide, with increased deaths of older people attributed to degenerative diseases. Deaths of Afro-descendants, therefore, present a bimodal distribution with high mortality at young and older ages. Mortality trends from this groups show considerable improvements between 2016 and 2019. However, this could be related to mortality underestimation and data quality problems. The epidemiological patterns among the Indigenous group on the other hand, present a high mortality in newborns and children compared with death at other ages. Likewise, the Indigenous group shows the highest proportion of deaths among women of reproductive age due to causes linked to pregnancy, childbirth, and puerperium. However, the causes of death more common in this ethnic group are primarily specific causes during the perinatal period, and nutritional deficiencies in the case of children, and in groups at older ages, circulatory diseases, cancer, and respiratory diseases are the most common causes of death.

The study found that health differences among Colombian ethnic groups in terms of causes and age at death are similar to differences when comparing rich and poor countries. The epidemiological profile of the White-Mestizo group is consistent with the global mortality transition in which infectious diseases and preventable deaths have been relatively low, while ethnic minorities lag far behind with a higher proportion of deaths in children and young population.

Finally, in Chapter 7, it was shown that the polarisation and the profound differences among ethnic groups related to health and social conditions paradoxically offer the opportunity to improve ethnic identification and data completeness in mortality registers when the ethnic identities of the deceased are missing. The study found that using information from mortality records, it is possible to ascribe ethnicity correctly with probabilities ranging between 80% and 90%.

8.2 Key contributions

The findings of this study represent key contributions in population health studies of ethnic groups in the Latin American region. It is the first analysis of its kind comparing different ethnic categories introduced in the Colombian mortality registers since 2008. The analysis of mortality inequalities taking into account of the measurement and estimation issues is vital to addressing the extent of health inequalities in the region, and the findings of this thesis contribute to the literature on demographic transition in Afro-descendants and Indigenous populations.

The study makes an important contribution in terms of understanding the ethnic bias in demographic data. More importantly, the application of correction methods enabled reliable estimations of life expectancy and mortality trends of Colombian ethnic minority groups, suggesting similar application of the methodology elsewhere in the region and in countries with high proportion of ethnic populations. The present study further contributes to the epidemiological characterisation of the ethnic population in terms of the most common disease patterns and causes of death. In the

Colombian context, the findings clearly highlight the dominant influence of homicide as one of the leading causes of death in the country, and the most affected by violence are the Afro-descendants.

In terms of contributions to research methodology, the present study undertook an experimental interdisciplinary approach which considered machine learning application to identify ethnic backgrounds of deceased population. Relevant population parameters such as age, gender, geographic location and causes of death were used to identify the ethnicity of the person, where appropriate weighting cases to ensure that the random forest algorithm was not biased towards correctly predicting the most prevalent ethnic group. The findings yielded useful insights by identifying the ethnicity correctly, as high as 90% with the highest percentages for Indigenous people and the lowest for Afro-descendants. The ethnic identification method represents thus a key contribution to determine ethnicity in multi-ethnic societies such as Colombia, in the absence of ethnicity data. By reconstructing the ethnic variable, we can design effective health and social policies interventions targeting marginalised ethnic groups.

8.3 Understanding findings through a theoretical lens

The differences in lifespan and causes of death among ethnic groups in Colombia were anticipated, given the evident socioeconomic inequalities among these populations. Such disparities have often been attributed to factors such as migration, poverty, or cultural background in other multiethnic countries. However, racism and ethnic discrimination are frequently overlooked as significant determinants of lifespan inequalities. It is crucial to recognise that both socioeconomic differences and health and mortality inequalities are rooted in the same structural causes, namely racism and discrimination.

Unlike other multiethnic regions such as Europe or North America, where ethnic diversity has arisen due to modern migration intertwined with xenophobia and racism, Colombian society has constructed racism differently. In Colombia, ethnic groups are considered locals and nationals. Nonetheless, the level of discrimination is influenced by a colonial background, where Indigenous and Afro-descendant people are perceived as having a lower social status and, more than that, are seen as individuals who do not deserve the same dignity as White-Mestizo members. This phenomenon often occurs unconsciously as implicit bias, where interactions with Indigenous and Afro-descendant people may be less cordial when they demand services and attention in restaurants, banks, etc. In other cases, the bias is more conscious. For instance, the first Afro-descendant vice president of the country has been harshly criticised for using private flights and national security services, which have historically been reserved for White-Mestizo vice presidents.

The post-colonial society in Colombia, characterised by clear privileges for White-Mestizos and disadvantages for Indigenous and Afro-descendant groups, has led to profound racial inequities. For example, areas predominantly inhabited by White-Mestizos receive more attention from the central government, resulting in significant developmental differences. In contrast, Indigenous and Afro-descendant populations are marginalised in poorer areas, posing serious challenges to accessing education, the labor market, and healthcare. The disadvantages in life expectancy among ethnic groups are directly related to these limitations in access to better living conditions and healthcare, which is evident when examining the causes of death. However, common causes of death, such as infections among Indigenous people and homicides among Afro-descendants, are often perceived as confirming racial stereotypesIndigenous people as unhygienic and Afro-descendants as prone to violence.

If differing living conditions among White-Mestizos, Indigenous, and Afro-descendant populations explain the varying epidemiological profiles among these groups, then the socioeconomic status model could account for the health disadvantages of Indigenous and Afro-descendant people. However, this model does not explain why the "free" market systematically leaves the same ethnic groups as winners and losers. Therefore, health ethnic inequities in Colombian society cannot be solely explained by socioeconomic factors; a deeper mechanism must be considered.

The existence of a social system of racism, in which both socioeconomic differences and health inequalities are generated, represents a more plausible explanation. Racism against Indigenous and Afro-descendant populations causes market distortions due to private allocation of resources and governmental neglect in public policy design. Consequently, economic and health inequalities in the Colombian context are produced by the same structure of colonial racism. Furthermore, the effects of racism and discrimination extend beyond socioeconomic factors to influence health outcomes and mortality patterns. Minority ethnic groups may experience higher rates of preventable diseases, and increased exposure to environmental hazards, all of which contribute to disparities in life expectancy and mortality rates.

Addressing these inequities requires acknowledging the role of racism and discrimination in shaping social and health outcomes and implementing policies and interventions aimed at dismantling systemic barriers and promoting equity for all ethnic groups.

8.4 Strengths and limitations

The present study is the first comprehensive research to address ethnic health inequalities at national level in Colombia. It employed census reported deaths and mortality records which are the most largest and complete datasets in the country. The study is therefore representative of the ethnic diversity in Colombia. The findings could be easily generalised and replicated in other areas of the Latin American region with similar ethnic composition, cultural characteristics and socioeconomic conditions because the study applied robust demographic estimation techniques that have been validated in similar contexts. The findings offer thus explanations about the complex mechanisms that determine the true representation of ethnic groups by quantifying the dimension of the ethnic gap in Colombia.

While population-level data from census and mortality records offer valuable insights, they also come with several limitations. First, the 2018 Census underestimated the Afro-descendant population due to coverage issues, making comparisons between Afro-descendant and other ethnic groups challenging due to under-representation. Second, the observed differences in lifespan and mortality are more pronounced between Indigenous and White-Mestizo groups compared to Afro-descendants and White-Mestizos. Although ethnic disparities for Afro-descendants are still evident, they could have been minimised in some cases which masked the mortality gap when comparing White-Mestizos and Afro-descendants. This phenomenon may be attributed to the serious censal omission in regions where Afro-descendants were not interviewed, while in other regions, Afro-descendants may have been categorised as White-Mestizos, a practice known as *blanqueamiento* [Whitening]. Censal omission poses a significant risk of bias in mortality patterns among Afro-descendants, particularly in age, gender, and causes of death if the omitted population differs structurally from those counted in the census. However, this scenario seems less likely due to the social homogeneity of ethnic groups in the country.

Another notable limitation of this study is its inability to assess the influence of the Colombian internal conflict on demographic patterns and its effect on mortality under-registration among Afro-descendant and Indigenous populations. It is conceivable that mortality under-registration is heightened in conflict-affected areas, potentially resulting in an incomplete recording of deaths and their associated causes. Furthermore, while censuses offer broader coverage, they do not capture data on causes of death, rendering the number of conflict-related deaths unknown. Despite efforts to correct for under-registration, estimates may still be affected by underestimated mortality rates. Nonetheless, the mortality patterns observed within ethnic populations and the distribution of causes of death among ethnicities are derived from national population data and reflect the genuine conditions faced by marginalised groups in the country.

Finally, the use of machine learning methods for ethnic recognition, although representing a promising tool for improving data quality, still necessitates careful considerations regarding racial bias inherent in training datasets, as well as the oversimplification of the complexity of ethnic identities into average or "modal" characteristics. Even when utilising images and names for ethnic ascription, inter-ethnic similarities and the presence of multi-ethnic families introduce significant biases and political considerations. While machine learning methods hold potential for various applications, their use in ethnic recognition should be approached with caution. Additionally, when mortality registers are unavailable, imputing the ethnicity of the head of the household to reported deceased persons in census surveys can serve as a reliable method for ethnic identification in homogeneous populations. However, this approach may introduce racial bias in multi-ethnic families. Addressing these challenges requires not only technical advancements but also careful attention to ethical, social, and legal implications.

8.5 Policy recommendations

Data quality presents a significant challenge for the analysis of health and mortality inequalities among ethnic groups. Skin colour, race, and ethnicity are three categories that are often used to identify ethnic populations. However, all three indicators are sensitive and subject to potential reporting bias. Incomplete and missing data about the different racial and ethnic categories make it difficult to properly compare ethnic groups over time. Ethnicity classification is often determined by sociocultural attributes and economic conditions.

The findings of this study point out the dire need to collect quality racial and ethnic information. Data on ethnicity and race should be compared with other indicators such as skin colour. Self-reported ethnicity does not always capture the way people could be classified; moreover, racism and discrimination fundamentally depend on others' perception. In that sense, skin colour is a piece of central information to understand changes in reported ethnicity. Skin colours palletes [1] have been used in different countries in order to collect information about skin colour, in which interviewees or family members for the case of deceased people, can report the skin tone or a racial category based on the cultural understanding of colour such as *Moreno, Trigueño*, White, etc. Second, surveys and census could use flash cards or images to help respondents identify their race including their family members and those deceased. This will allow researchers to differentiate persons with the same skin colour but from different groups. Third, the questions about ethnic identities should be open ended in the Colombian context allowing multiple answers, so more than one ethnicity can be chosen both within households and for individuals. A large majority of households are made up of parents of different racial and ethnic backgrounds which creates a multi-ethnic identity in descendant members. This could create a problem of inconsistency of ethnic identities if persons have to choose only one of their identities but they use a different identity at different ages and in different surveys.

Collecting information about skin color, race, and ethnicity would mitigate the problem of *blanqueamiento*, wherein the ethnic identity of Afro-descendants and Indigenous individuals is inaccurately recorded as White-Mestizo, thus perpetuating racial bias. Comprehensive data collection would enable researchers to measure, for instance, the proportion of people who report a "White-Mestizo" identity regardless of their skin color. It is crucial to recognise that the effects of racism and discrimination are not contingent upon self-reported ethnic identities. Even if an Afro-descendant or Indigenous individual reports a White-Mestizo identity, this does not "absolve" them from experiencing racism and discrimination. Therefore, additional information about race and ethnicity will empower researchers to investigate whether variations in health and mortality patterns among racial and ethnic groups are attributable to changes in the number of

[1] Previous studies analysed racial discrimination by measuring skin colours using a scale of skin tones in the US (Gullickson2005, 2005; Keith and Herring, 1991) and in some Latin American Countries (Telles et al., 2015). The skin-colour palette used in the Latin American region can be viewed at `http://perla.princeton.edu/surveys/perla-color-palette`. Other studies in Puerto Rico and Brazil employed proxy variables to measure skin colours (Guimarães, 2012; Gravlee, 2005)

people classifying themselves within a particular group or to changes in the social environment and living conditions of groups with specific ethnic identities. This approach allows for a more nuanced understanding of the intersection between identity and the structural impacts of racism and discrimination.

Likewise, in order to improve data collection and registration completeness from ethnic minority groups, it is important to adapt the registration system to the social context of Afro-descendants and Indigenous communities. That is, the registration of vital statistics should also consider language, cultural differences, and geographical access to ethnic communities with the aim of reducing the barriers for registration.

In terms of mortality and causes of death, the present study directs special attention to children and young population of Indigenous and Afro-descendants communities. It is fundamental to design social programmes aimed at these two age-groups, including nulliparous women at reproductive age in order to mitigate the high mortality during the perinatal period due to infectious diseases, nutritional deficiencies and violence/homicide. Equally important is the need to compare data from other sources including registration of deaths by nurses, midwives and other community health workers working in Indigenous and Afro-descendant communities.

8.6 Future research

Future research should examine the specific reasons underlying the non-registration of deaths in Indigenous and Afro-descendants communities. It is therefore necessary to investigate the extent to which this non-registration could be determined by "barriers" in the registration system, such as the lack of access to registration offices especially in remote regions where administrative systems are fragmented or dysfunctional due to poor resources. Alongside, more systematic efforts are required to monitor and record deaths and related relevant data in regions affected by conflict and illegal activities including alcohol or drug abuse. It is important to understand the individual and contextual barriers associated with under reporting of deaths and their causes.

The reasons for abrupt decrease in homicide attributed mortality levels of Afro-descendents. It is unclear if the decrease in registered deaths is due to ethnicity misclassification and registration bias, or due to changes in the number of homicides in young Afro-descendant men. In some years of the period, the number of homicides registered as White-Mestizos is three times higher than those for Afro-descendants especially in regions where Afro-descendants represent over 80% of the total population. In other cases, deaths of Afro-descendants decreased to zero while White-Mestizo deaths increased. Future research should examine and explain these unusual patterns and trends in ethnic mortality.

The machine learning application of ethnic classification model yielded between 80% and 90% of deaths which were correctly classified. However, future improvements of the model should consider the use of names and surnames as input variables. The use of personal names could increase the percentage of correctly classified cases as name patterns are closely related to the cultural ethnic background of Afro-descendants and Indigenous populations.

Future research could consider a mixed methods approach to understand and validate the dynamics of ethnicity classification in Colombia, and further understand the contextual factors and barriers determining reporting of deaths and associated causes.

Bibliography

Abel, T. and Frohlich, K. (2012). Capitals and capabilities: Linking structure and agency to reduce health inequalities. *Social Science & Medicine*, 74:236–244.

Abraido-Lanza, A. F., Dohrenwend, B. P., Ng-Mak, D. S., and Turner, B. J. (1999). The Latino Mortality Paradox: A Test of the "Salmon Bias" and Healthy Migrant Hypotheses. *American Journal of Public Health*, 89(10):1543–1548.

Aburto, J. M., Alvarez, J.-A., Villavicencio, F., and Vaupel, J. W. (2019). The Threshold Age of the Lifetabe Entropy. *Demographic Research*, 41(4):83–102.

Aburto, J. M., Wensink, M., van Raalten, A., and Lindahl-Jacobsen, R. (2018). Potential Gains in Life Expectancy by Reducing Inequality of Lifespans in Denmark: an International Comparison and Cause-of-Death Analysis. *Bio Med Central (BMC) Public Health*, 18(831):1–9.

Adjaye, D., Bednarczyk, R. A., and Davis, R. L. (2014). Using the Bayesian Improved Surname Geocoding Method (BISG) to Create a Working Classification of Race and Ethnicity in a Diverse Managed Care Population: A Validation Study. *Health Research and Educational Trust*, 49(1):268–283.

Adorno, T. W., Frenkel-Brunswik, E., Levinson, D. J., and Sanford, R. N. (1950). *Types and Syndromes*, pages 744–786. Harpers, Cham.

Agrawal, G. and Kandhuja, P. (2015). Influence of Literacy on India's Tendency for Age Misreporting: Evidence from Census 2011. *Journal of Population and Social Studies*, 23(1):47–56.

Agudelo-Suárez, A., Martínez-Herrera, E., Posada-López, A., and Rocha-Buelvas, A. (2016). Ethnicity and Health in Colombia: What do Self-Perceived Health Indicators Tell Us? *Ethnicity & Disease*, 26(2):147–156.

Ajwani, S., Blakely, T., Robson, B., Tobias, M., and Bonne, M. (2003). Decades of Disparity: Ethnic mortality trends in New Zealand 1980–1999. *Ministry of Health and Univesity of Otago*. Wellington.

Akobeng, A. K. (2006). Understanding diagnostic test 1: sensitivity, specificity, and predictive values. *Acta Paediatrica*, 96:338–341.

Albala, C., Vio, F., and Yáñez, M. (1997). Transición epidemiológica en América Latina: comparación de cuatro países [Epidemiological transition in Latin America: a comparison of four countries]. *Revista Médica de Chile*, 125(6):719–727.

Aldridge, R. W., Lewer, D., Katikireddi, S. V., Mathur, R., Pathak, N., Burns, R., Fragaszy, E. B., Johnson, Anne, M., Devakumar, D., Abubakar, I., and Hayward, A. (2020). Black, Asian and Minority Ethnic groups in England are at increased risk of death from Covid-19: direct standardisation of NHS mortality data. *Wellcome Open Research*, 5(88).

Alkire, S., Kanagaratnam, U., and Suppa, N. (2018). The global multidimensional poverty index (mpi): 2018 revision. *OPHI MPI methodological notes*, 46.

Anderson, E. (1990). *Three Worlds of Welfare Capitalism*. Princeton University Press.

Andrews, G. R. (2016). Los Afrodescendientes en los Censos Latinoamericanos, 1776-2011 [Afro-Latin: Black Lives, 1600-2000]. *Claves. Revista de Historia*, 2:257–278.

Anson, J. (2002). Of Entropies and Inequalities: Summary Measures of the Age Distribution of Mortality. In Wunsch, G., Mouchart, M., and Dchêne, J., editors, *The Life Table, Modelling Survival and Death*, chapter 4, pages 95–116. European Asociation for Population Studies - Kluwer Academic Publishers.

Aponte-Gonzáles, J., Rincón, C., and Eslava-Schmalbach, J. (2012). The impact of under-recording on cervical cancer-related mortality in Colombia: an equity analysis involving comparison by provenance. *Revista de Salud Pública*, 14(6):912–921.

Azpiroz, C. n. V. (2022). Etnonacionalismo mapuche y ciudadanía argentina: convivencialidad [Mapuche Ethnonationalism and Argentine Citizenship: Coexistence]. *Pueblos Indígenas Y Educación*, (68):167–176.

Bailey, N. A. (1967). La Violencia in Colombia. *Journal of Inter-American Studies*, 4(10):561–575.

Baker, A. (2015). Race, paternalism, and foreign aid: Evidence from u.s. public opinion. *The American Political Science Review*, 109(1):93–109.

Bamshad, M. J. and Olson, S. E. (2003). Does Race Exist? *Scientif American*, 289(6):78–85.

Banco-De-La-República (1951). Bases de un programa de fomento para Colombia, Informe de una misión dirigida por Lauchlin Currie y asupiciada por el Banco Internacional de Reconstrucción y Fomento, en colaboración con el Gobierno de Colombia. *Banco de la República*. Bogotá, Talleres Editoriales de Librería Voluntad, SA, 2a. Edición.

Banda, Y., Markm, N. K., Thomas, J. H., Stephanie, E. H., Dilrini, R., Hua, T., Chiara, S., Lisa, A. C., Brad, P. D., Mary, H., Carl, o. I., Eric, J., Lawrence, H. K., Dana, L., Diane, O., Charles, P. Q., Sarah, R., Marianne, S., Lori, C. S., Stanley, S., Ling, S., David, S., Carol, P. S., Stephen, K. V. D. E., Lawrence, W., Rachel, A. W., Pui-Yan, K., Catherine, S., and Neil, R. (2015). Characterizing Race/Ethnicity and Genetic Ancestry for 100,000 Subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) Cohort). *Genetics*, 200:1285–1295.

Banguero, H. and Castellar, C. (1993). La población de Colombia. *Universidad del Valle*.

Banguero, L. H. (2005). Estructura y dinámica de la población del departamento del Cauca y de Colombia. 1938-2025. Un análisis comparativo. [Structure and population dinamic of the department of Cauca and Colombia. 1938 - 2025. A comparative analysis]. *El Hombre y la Máquina*, 24:72–91.

Baquero, M. J. (2015). The intersection of race, class, and ethnicity in agrarian inequalities, identities, and the social resistance of peasants in Colombia. *Current Sociology*, 63(7):1017–1036.

Barbary, O. and Urrea, F. (2004). Gente Negra en Colombia: Dinámicas Sociopolíticas en Cali y el Pacífico [Black People in Colombia: Socio-political Dynamics in Cali and the Pacific]. *CIDSE-UNIVALLE-IRD-COLCIENCIAS*.

Barrett, R. E. (2019). *Age Heaping*, pages 1–7. Springer International Publishing.

Beckett, M. (2000). Converging Health Inequalities in Later Life-An Artifact of Mortality Selection? *Journal of Health and Social Behavior*, 41(1):106–109.

Bennet, N. G. and Horiuchi, S. (1984). Mortality Estimation from Registered Deaths in Less Developed Countries. *Demography*, 21(2):217–233.

Benthall, S. and Haynes, B. D. (2019). Racial Categories in Machine Learning. *FAT Association for Computing Machinery*, pages 289–298.

Berger, P. and Luckmann, T. (1966). The Social Construction of Reality. *Allen Lane The Penguin Press*.

Bernal, R. and Cárdenas, M. (2007). Race and Ethnic Inequality in Health and Health Care in Colombia. In Giuffrida, A., editor, *Racial and Ethnic Disparities in Health in Latin America and the Caribbean*, chapter 2, pages 23–64. Inter-American Development Bank. Country Department Andean Group.

Bernard, M. (1987). Race and Class Revisited: The Case of the North America and South Africa. *Africa Development*, 12(1):5–42.

Bertrand, M. and Mullainathan, S. (2004). Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment of Labor Market Discrimination. *The American Economics Review*, 94(4):991–1013.

Better, S. (2008). *Institutional racism: A primer on theory and strategies for social change*. Rowman & Littlefield.

Bhopal, R. (2004). Glossary of Terms Relating to Ethnicity and Race: For Reflection and Debate. *J Epidemiol Community Health*, 58:441–445.

Bhopal, R. and Donaldson, L. (1998). White, European, Western, Caucasian, or What? Inappropriate Labeling in Research on Race, Ethnicity, and Health. *American Journal of Public Health*, 88(9):1303–1307.

Blackwell, D. L., Hayward, M. D., and Crimmins, E. M. (2001). Does childhood health affect chronic morbidity in later life? *Social Science & Medicine*, 52:1269–1284.

Blagus, R. and Lusa, L. (2013). SMOTE for high-dimensional class-imbalaced data. *BioMed Central (BMC) Bioinformatics*, 14(106).

Blakely, T., Atkinson, J., and Fawcett, J. (2008). Ethnic counts on mortality and census data (mostly) agree for 2001- 2004: New Zealand Census-Mortality study update). *Journal of the New Zealand Medical Association*, 121(1281):58–62.

Blakely, T., Woodward, A., and Salmond, C. (2000). Anonymous linkage of New Zealand mortality and Census data. *Australian and New Zealand Journal of Public Health*, 24(1):92–95. Brief Report.

Blaxter, M. (2005). *Health and Lifestyles*, chapter 9–10. Routledge.

Blom, N., Huijts, T., and Kraaykamp, G. (2016). Ethnic health inequalities in Europe. The moderating and amplifying roel of healthcare system characteristics. *Social Sciences and Medicine*, 158:43–51.

Bodnar, Y. C. (2006). Pueblos Indígenas de Colombia: Apuntes Sobre la Diversidad Cultural y la Información Sociodemográfica Disponible [Indigenous peoples of Colombia: notes on cultural diversity and available sociodemographic information]. In CEPAL, editor, *Pueblos Indígenas y Afrodescendientes de América Latina y el Caribe: Información Sociodemográfica para Políticas y Programas*, pages 135–154. CEPAL - United Nations.

Bonilla-Silva, E. (1997). Rethinking racism: Toward a structural interpretation. *American Sociological Review*, 62(3):465–480.

Bonilla-Silva, E. (1999). The Essential Social Fact of Race. *American Sociological Association*, 64(6):899–906.

Bonilla-Silva, E. (2006). *Racism without Racists: Color-Blind Racism and the Persistence of Racial Inequality in the United States*. Rowman & Littlefield Publishers.

Bourdieu, P. (1984). Distinction: A social critique of the judgement of taste. *Harvard University Press*.

Bourdieu, P. (1989). Social Space and Symbolic Power. *American Sociological Association*, 7(1):14–25.

Bouvier-Colle, M., Varnoux, N., Costes, P., and Hatton, F. (1991). Reasons for the Underreporting of Maternal Mortality in France, as Indicated by Survey of All Deaths among Women of Childbearing Age. *International Journal of Epidemiology*, 20(3):717–721.

Bouwhuis, C. and Moll, H. (2003). Determination of ethnicity in children in the Netherlands: Two methods compared. *European Journal of Epidemiology*, 18(5).

Boyd, E. L., Taylor, S. D., Shimp, L. A., and R., S. C. (2000). An Assessment of Home Remedy Use by African Americans. *Journal of The National Medical Association*, 92(7):341–353.

Bradley, A. P. (1997). The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms. *Elsevier*, 30(7):1145–1159.

Bramley, D., Hebert, P., Jackson, R., and Chassin, M. (2004). Indigenous disparities in disease-specifi mortality, a cross-country comparison: New Zealand, Australia, Canada, and the United States. *The New Zealand Medical Journal*, 117(1207).

Brant, L. J. and Boxall, E. (2009). The problem with using computer programmes to assign ethnicity: Immigration decreases sensitivity. *Public Health*, 123:316–320.

Braun, L. (2002). Race, Ethnicity, and Health: Can Genetics Explain Disparities? *John Hopkins University Press*, pages 159–174.

Brave-Heart, M. Y. H. and DeBruyn, L. M. (1991). The American Indian holocaust: Healing historical unresolved grief. *American Indian and Alaska Native Mental Health Research*, 8(2):60–82.

Breiman, L. (2001). Random Forest. *Machine Learning*, 45:5–32.

Breiman, L. and Cutler, A. (2022). Breiman and cutler's random forest for classification and regression'. *Repository CRAN*.

Brown, T. N., Williams, D. R., Jackson, J. S., Neighbors, H. W., Torres, M., Sellers, S. L., and Brown, K. T. (1999). Bein Black and Feeling Blue: The Mental Health Consequences of Racial Discrimination. *Race & Society*, 2(2):117–131.

Burrel, J. and Fourcade, M. (2021). The Society of Algorithms. *Annual Review of Sociolgy*, 47:213–237.

Buxbaum, J., Jacobson, D. R., Tagoe, C., Alexander, A., Kitzman, D. W., Greenberg, B., Thaneemit-Chen, S., and Lavori, P. (2006). Transthyretin v122i in african americans with congestive heart failure. *Journal of the American College of Cardiology*, 47(8):1724–1725.

Buxbaum, J. N. and Ruberg, F. L. (2017). Transthyretin V122I (pV142I) cardiac amyloidosis: an age–dependent autosomal dominant cardiomyopathy too common to be overlooked as a cause of significant heart disease in elderly African Americans. *Genet Med*, 19(7):733–742.

Bécares, L., Nazroo, J., and Stafford, M. (2009). The buffering effects of ethnic density on experienced racism and health. *Health Place*, 15(3):700–708.

Campos de Lima, E. M. and Lanza, Q. B. (2014). Evolution of the deaths registry system in Brazil: associations with changes in the mortality profile, under-registration of death counts, and ill-defined causes of death. *Cad. Saúde Pública*, 30(8):1721–1730.

Carabali, H. B., Beltrán, S. H., Aidar, T., and Palloni, A. (2021). Estimación de la tasa de mortalidad en contexto de altos homicidios: caso de la región Pacífica colombiana, 1993-2013 [Mortality rates estimates in the context of high homicide levels and poor data quality on deaths: The case of the Pacific region in Colombia between 1993 and 2013]. *RELAP - Revista Latinoamericana de Población*, 15(29):140–165.

Cárdenas, R. (2018). Thanks to my forced displacement: blackness and the politics of Colombia's war victims. *Latin American and Caribbean Ethnic Studies*, 13(1):72–93.

CDC (2020). International Classification of Diseases, Tenth Revision (ICD-10). `https://www.cdc.gov/nchs/icd/icd10.htm`. National Center for Health Statistics, USA.

Cell, J. W. (1982). The Highest Stage of White Supremacy: The Origin of Segregation in South Africa and the America South. *Cambridge University Press*.

Cendales, R. and Pardo, C. (2011). La calidad de certificación de la mortalidad en Colombia, 2002-2006. *Revista de Salud Pública*, 13(2):229–238.

Chang, M.-H., Moonesinghe, R., Athar, H. M., and Truman, B. I. (2016). Trends in Disparity by Sex and Race/Ethnicity for the Leading Causes of Death in the United States 1999-2010. *J Public Health Management Practice*, 22(1):13–24.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, P. W. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Chouhan, K. and Nazroo, J. (2020). Health inequalities. In Byrne, B., Alexander, C., Khan, O., Nazroo, J., and Shankley, W., editors, *Ethnicity, Race and Inequality in the UK: State of the Nation*, chapter 4, pages 73–92. Policy Press, University of Bristol.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments.

Cooper, R., Rotimi, C., Ataman, S., McGee, D., Osotimehin, B., Kadiri, S., Muna, W., Kingue, S., Fraser, H., Forrester, T., Bennet, F., and Wilks, R. (1997). The Prevalence of Hypertension in Seven Populations of West African Origin. *American Journal of Public Health*, 87(2):160–168.

Cornell, S. and Hartmann, D. (2007). Ethnicity and race : making identities in a changing world. *Pine Forge Press*. Second Edition.

Crimmins, E. M., Hayward, M. D., and E., S. T. (2004a). Race/Ethnicity, Socioeconomic Status, and Health. In National-Research-Council, editor, *Critical Perspectives on Racial and Ethnic Differences in Health in Late Life*, chapter 9, pages 310–352. National Academy of Sciences.

Crimmins, E. M., Hayward, M. D., and E., S. T. (2004b). Race/Ethnicity, Socioeconomic Status, and Health. In Council, N. R., editor, *Critical Perspectives on Racial and Ethnic Differences in Health in Late Life*, chapter 9, pages 310–352. National Academy of Sciences, Washington, DC.

Cristancho, F. C. (2017). Niveles, tendencias y determinantes de la mortalidad reciente en Colombia. `https://ddd.uab.cat/pub/tesis/2017/hdl_10803_406099/ccf1de1.pdf`. Ph.D. Thesis [Accessed: 2021-02-22].

Cunneen, C. (2006). Racism, discrimination and the over-representation of indigenous people in the criminal justice system: Some conceptual and explanatory issues. *Current issues in criminal justice*, 17(3):329–346.

de Benoist, A. (1999). What is racism? *Telos: Critical Theory of the Contemporary*, 1999(114):11–48.

Del Popolo, F., Garcia de Pinto da Cunha, E. M., Ribbota, B., and Azevedo, M. (2011). *Pueblos Indígenas y afrodescendientes en américa Latina: dinámicas poblacionales diversas y desafíos comunes[Indigenous and Afro-descendant folks in Latin America: diverse population dynamics and common challenges]*. UNFPA–United Nations. Editor: ALAP.

Del-Popolo, F. and Oyarce, A. M. (2005). Población indígena de América Latina: perfíl sociodemográfico en el marco de la CIPD y de las Metas del Milenio. CELADE–CEPAL.

Deneux-Tharaux, C., Berg, C., Bouvier-Colle, M.-H., Gissler, M., Harper, M., Nannini, A., Alexander, S., Wildman, K., Breart, G., and Buekens, P. (2005). Underreporting of Pregnancy-Related Mortality in the United States and Europe. *Obstetrics & Gynecology*, 106(4):684–692.

Departamento-Administrativo-Nacional-De-Estadística (2006). La Mortalidad Materna y Perinatal en Colombia en los Albores del Siglo XXI. Estimación del Subregistro de Nacimientos y Defunciones y Estimaciones Ajustadas de Nacimientos, Mortalidad Materna y Perinatal por Departamentos. `https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/INEC/IGUB/col-unfpa-dane-estudiocabalidad-mm-2003.pdf`. DANE, UNFPA, Ministerio de Protección Social. [Accessed: 2021-02-25].

Departamento-Administrativo-Nacional-De-Estadística (2007). Colombia una nación multicultural: su diversidad étnica [Colombia a multicultural nation: its ethnic diversity]. `https://www.dane.gov.co/files/censo2005/etnia/sys/colombia_nacion.pdf`. Departamento Nacional de Estadística (DANE)). [Accessed: 2021-09-24].

Departamento-Administrativo-Nacional-De-Estadística (2017). Pobreza Monetaria y Multidimensional en Colombia [Monetary and Multidimensional Poverty in Colombia]. https://www.dane.gov.co/files/investigaciones/condiciones_vida/pobreza/bol_pobreza_17_v2.pdf. DANE, UNFPA, Ministerio de Protección Social. [Accessed: 2024-05-23].

Departamento-Administrativo-Nacional-De-Estadística (2019a). Población Indígena de Colombia: Resultados Del Censo Nacional de Población y Vivienda 2018 [Indigenous Population: Results of the National Census 2018]. https://www.dane.gov.co/files/investigaciones/boletines/grupos-etnicos/presentacion-grupos-etnicos-2019.pdf. [Accessed: 2021-07-27].

Departamento-Administrativo-Nacional-De-Estadística (2019b). Población Negra, Afrocolombiana, Raizal, y Palenquera: Resultados Del Censo Nacional de Población y Vivienda 2018 [Black, Afro-Colombian, Black-Caribbean, and Black-Maroon Population: Results of the National Census 2018]. https://www.dane.gov.co/files/investigaciones/boletines/grupos-etnicos/presentacion-grupos-etnicos-poblacion-NARP-2019.pdf. [Accessed: 2021-07-27].

Departamento-Administrativo-Nacional-De-Estadística (2021). Estimación de la Población Étnica a Nivel Subnacional [Estimation of Ethnic Population at Sub-Population level]. https://www.dane.gov.co/files/investigaciones/experimentales/demografia-poblacion/pertenencia-etnica/pertenencia-etnica-documento-metodologico.pdf. [Accessed: 2021-12-03].

Derrington, R. (2013a). The Brass Growth Balance method. In Moultrie, T. A., Dorrington, R. E., Hill, A. G., Hill, K., Timaeus, I. M., and Zaba, B., editors, *Tools for demographic estimation*, chapter 20, pages 196–208. IUSSP-UNFPA.

Derrington, R. (2013b). The synthetic extinct generations method. In Moultrie, T. A., Dorrington, R. E., Hill, A. G., Hill, K., Timaeus, I. M., and Zaba, B., editors, *Tools for demographic estimation*, chapter 25, pages 275–292. IUSSP-UNFPA.

Di Cesare, M. (2010). El Perfil Epidemiológico de América Latina y el Caribe: Desafíos, Límites y Acciones. *Comisión Económica para América Latina y el Caribe - CEPAL*. Available online at http://www.eclac.cl/publicaciones/xml/9/44309/lcw395.pdf; visited on 28 January 2014.

DNP (1989). Los Pueblos Indígenas' de Colombia en el Umbral Ddel Nuevo Milenio [The Indigenous Peoples of Colombia on the Threshold of the New Millennium]. https://web.archive.org/web/20070610220752/http://www.dnp.gov.co/paginas_detalle.aspx?idp=452. [Accessed: 2024-03-24].

Dorrington, R. (2013a). The Brass Growth Balance Method. In Moultrie, T., Dorrington, R., Hill, A., Hill, K., Timaeus, I., and Zaba, B., editors, *Tools for Demographic Estimation*, chapter 20, pages 196–208. International Union for the Scientific Study of Population (IUSSP)), Paris.

Dorrington, R. (2013b). The Generalised Growth Balance Method. In Moultrie, T., Dorrington, R., Hill, A., Hill, K., Timaeus, I., and Zaba, B., editors, *Tools for Demographic Estimation*, chapter 24, pages 258–274. International Union for the Scientific Study of Population (IUSSP)), Paris.

Dressler, W. W. (1993). Health in the African American Community Accounting for Health Inequalities. *Medical Anthropology Quarterly*, 7(4):325–345.

Dressler, W. W., Oths, K. S., and Gravlee, C. C. (2005). Race and Ethnicity in Public Health Research: Models to Explain Health Disparities. *The Annual Review of Anthropology*, 34:231–252.

Dubowitz, T., Heron, M., Basurto-Davila, R., Bird, C. E., Lurie, N., and Escarce, J. J. (2011). Racial/Ethnic Differences in U.S. Health Behaviors: A Descomposition Analysis. *Am J Health Behav.*

Duryea, S., Olgiati, A., and Stone, L. (2006). The Under-Registration of Births in Latin America. `https://publications.iadb.org/publications/english/document/The-Under-Registration-of-Births-in-Latin-America.pdf`. Working Paper No. 551 [Accessed: 2021-04-08].

ECLAC (1990). Boletín Demográfico [Demografic Bulletin]. *ECLAC, United Nations*, 45.

ECLAC (2007). Mortalidad [mortality]. *ECLAC, United Nations*, 4.

ECLAC (2010). Population and health in Latin America and the Caribbean: outstanding matters, new challenges. `https://repositorio.cepal.org/bitstream/handle/11362/2946/1/S2010206_en.pdf`. ECLAC Ad Hoc Committee on Population and Development.

Edwards, F., Lee, H., and Esposito, M. (2019). Risk of being Killed by Police Use of Force in the United States by Age, Race-Ethnicity, and Sex . *PNAS*, 116(34).

Ehrlich, P. R. and Holm, R. W. (1964). A Biological View of Race. In Montagu, A., editor, *The Concept of Race*, pages 153–179. The Free Press of Glencoe.

Elliott, M. N., Fremont, A., Morrison, P. A., Pantoja, P., and Lurie, N. (2008). A New Method for Estimating Race/ Ethnicity and Associated Disparities Where Administrative Records Lack Self-Reported Race/Ethnicity. *Health Services Research*, 43(5).

Escobar, A. (2003). Displacement, development, and modernity in the colombian pacific. *International Social Science Journal*, 55(175):157–167.

Evans-Campbell, T. (2008). Historical trauma in american indian/native alaska communities: A multilevel framework for exploring impacts on individuals, families, and communities. *Journal of Interpersonal Violence*, 23(3):316–338.

Fanon, F. (1961). The wretched of the earth. *Grove Weidenfeld*. New York.

Feagin, J. and Bennefield, Z. (2014). Systemic racism and u.s. health care. *Social Science Medicine*, 103:7–14. Structural Stigma and Population Health.

Flach, P., Orallo-Hernández, J., and Ferrari, C. (2011). A Coherent Interpretation of AUC as a Measure of Aggregated Classfication Performance. *Proceeding of the 28th International Conference on Machine Learning*. WA, USA.

Florez, C. E., H, R., Méndez, R., and de la República (Colombia), B. (2000). *Las transformaciones sociodemográficas en Colombia durante el siglo XX*. Economía del siglo XX. Banco de la República.

Flórez, C. E., Medina, C., and Urrea-Giraldo, F. (2003). Los costos de la exclusión social por raza o etnia en América Latina y el Caribe. *Coyuntura Social*, 29.

Florez, E. C. (1987). La transición demográfica en Colombia. *Uniandes*.

Flórez, J. E. (1995). Transición Demográfica en Colombia? *Cuadernos de Geografía*, 5(2):133–146.

Franco, A. S. (1997). Violence and health in Colombia. *Pan Am J Public Health*, 2(3):170–180.

Frenk, J., Bobadilla, J. L., Sepulveda, J., and Lopez, C. M. (1989). Health Transition in Middle-income Countries: New Challenges for Health Care. *Health Policy and Planning*, 4:29–39.

Frenk, J., Bobadilla, J. L., Stern, C., Frejka, T., and Lozano, R. (1991a). Elements for a theory of the health transition. *National Center for Epidemiology and Population Health (NCEPH)*, 1(1):21–38. The Australian National University.

Frenk, J., Lozano, R., and Bobadilla, J. L. (1991b). La Transición Epidemiológica en America Latina. *The World Bank*, 1(1):79–101. The Australian National University.

Frohlich, K. L., Corin, E., and Potvin, L. (2001). A theoretical proposal for the relationship between context and disease. *Sociology of Health & Illness*, 23(6):776–797.

Gakidou, E., Horgan, M., and Lopez, A. D. (2004). Adult Mortality: Time for a Reappraisal. *International Journal of Epidemiology*, 33:710–717.

Garner, S. (2017). Racismus: An Introduction. *SAGE*. Second Edition.

Gee, G., S., S.-M., and Chae, D. (2009). Racial discrimination and health among Asian americans: evidence, assessment, and directions for future research. *Epidemiol Rev.*, 31:130–151.

Gillespie, T. (2013). *The Relevance of Algorithms*, chapter 9, pages 167–193. MIT Press.

Giraldo, D., Atehortúa, A., García-Arteaga, J. D., Díaz-Jiménez, D. P., Romero, E., and Rodríguez, J. (2017). Modelo para el análisis de la mortalidad en colombia 2000-2012 [model for mortality analysis in colombia 2000-2012]. *Revista Salud Pública*, 19(2):241–249.

Giufrida, A., Bernal, R., Cárdenas, M., Handa, A., Trujillo, A. J., Vernon, J. A., Rodriguez, W. L., Angeles, G., Mayer-Foulkes, D., and Larrea, C. (2007). Racial and Ethnic Disparities in Health in Latin America and the Caribbean. *Country Department Andean Group. Inter-American Development Bank*. Edited by Antonio Giuffrida.

Glei, D. A., Barajas, P. A., Aburto, J. M., and Barbieri, M. (2021). Mexican mortality 1990-2016: Comparison of unadjusted and adjusted estimates. *Demographic Research*, 44(30).

Gómez, I. F. (2014). El desplazamiento forzado de los pueblos indígenas en Colombia [The forced displacement of Indigenous peoples in Colombia]. *Revista del Departamento de Derecho*, 30:431–455. Universidad de Navarra, Pamplona-España.

Gonzáles, B. E., Ziv, E., Coyle, N., Gomez, S. L., Tang, H., Karter, A. J., Mountain, J. L., Pérez-Stable, E. J., Sheppard, D., and Risch, N. (2003). The importance of Race and Ethnic Background in Biomedical Research and Clinical Practice. *The New England Journal of Medicine*, 348(12):1170–1175.

Goodman, A. H. (2000). Why Genes Don't Count (for Racial Differences in Health). *American Journal of Public Health*, 90(11):1699–1702.

Gravlee, C. C. (2005). Ethnic Classification in Southeastern Puerto Rico: The Cultural Model of "Color". *Social Forces*, 83(3):949–970.

Gravlee, C. C. (2009). How Race Becomes Biology: Embodiment of Social Inequality. *American Journal of Physical Anthropology*, 139:47–57.

Groothuis, S. (2020). Research race, racialisation, and racism in critical terrorism studies: clarifying conceptual ambibuities. *Springer Nature*, 10(35):3097–9.

Gross, C. P., Essien, U. R., Pasha, S., Gross, J. R., Wang, S.-y., and Nunez-Smith, M. (2020). Racial and Ethnic Disparities in Population-Level Covid-19 Mortality. *General International Medicine*, 35(10).

Guillot, M., Gavrilova, N., and Pudrovska, T. (2011). Understanding the "Russian Mortatality Paradox" in Central Asia: Evidence from Kyrgystan. *Demography*, 48(3):1081–1104.

Guimarães, A. S. (2012). The Brazilian System of Racial Classification. *Ethnic and Racial Studies*, 35(7):1157–1162.

Gullickson2005 (2005). The Significance of Color Declines: A Re-Analysis of Skin Tone Differentials in Post-Civil Rights America. *Social Forces*, 84(1):157–180.

Guo, K., Ying, P., Wang, L., Ji, Y., Li, Q., Bishai, D., Liu, S., Liu, Y., Astell-Burt, T., Feng, X., You, J., Liu, J., and Zhou, M. (2015). Propensity score weighting for addressing under-reporting in mortality surveillance: a proof-of-concept study using the nationally representative mortality data in China. *Population Health Metrics*, 13(16):1–11.

Guzmán, G., Fals-Borda, O., and Umaña Luna, E. (1962). La Violencia en Colombia; Estudio de un proceso social [The Violence in Colombia; Study on a social process]. *Ediciones Tercer Mundo*, page Pp 430.

Hall, S. (2018). *6. Race, Articulation, and Societies Structured in Dominance [1980]*, pages 172–221. Duke University Press, New York, USA.

Hall, W., Chapman, M., Lee, K., Merino, Y., Thomas, T., Payne, B., Eng, E., Day, S., and Coyne-Beasley, T. (2015). Implicit Racial/Ethnic Bias Among Health Care Professionals and Its Influence on Health Care Outcomes: A Systematic Review. *American Journal of Public Health*, 105(12):60–76.

Haller, J. S. (1970). The physician versus the negro: medical and anthropological concepts of race in the late nineteenth century. *The Johns Hopkins University Press*, 44(2):154–167.

Han, H., Guo, X., and Yu, H. (2016). Variable selection using mean decrease accuracy and mean decrease gini based on random forest. In *2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, pages 219–224.

Hand, D. J. and Till, R. J. (2001). A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classfication Problems. *Machine Learning*, 45:171–186.

Hankerson, S. H., Moise, N., Wilson, D., Waller, B. Y., Arnold, K. T., Duarte, C., Lugo-Candelas, C., Weissman, M. M., Wainberg, M., Yehuda, R., and Shim, R. (2022). The intergenerational impact of structural racism and cumulative trauma on depression. *American Journal of Psychiatry*, 179(6):434–440.

Hanushek, E. A. and Rivkin, S. G. (2006). School Quality and The Black-White Achievement Gap. *National Bureau of Economic Research*, Working Paper 12651(2–43). NBER Working Paper Series.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. *Springer*.

Hayward, M. D., Miles, T. P., M., C. E., and Yang, Y. (2000). The Significance of Socioeconomic Status in Explaining the Racial Gap in Chronic Health Conditions. *American Sociological Review*, 65(6):910–930.

Hernández, A. M., Obregón, T. D., Miranda, C. N., García, C. M., Eslava, J. C., Romero, V. R., Hernández, T. J., Vega, V. M., Villamizar, V. C., and Barbosa, M. D. (2002). La OPS y el Estado colombiano: Cien años de historia [The PAHO and the Colombian state: one hundred years of history]. *PAHO-WHO*.

Hernández, V. J. A., Ramírez, B. P., Gil, Q. A. M., Valbuena, A. M., Acuña, L., and Gonzáles, J. A. (2020). Patterns of breast, prostate and cervical cancer incidence and mortality in Colombia: an adminstrative registry data analysis. *BioMed Central (BMC) Cancer*, 20(1097):3–9.

Hopenhayn, M., Bello, A., and Miranda, F. (2006). Los pueblos indígenas y afrodescendientes ante el nuevo milenio. *ECLAC, United Nations*, 118:1–60.

Horiuchi, S., Wilmoth, J. R., and D., P. S. (2008). A Decomposition Method Based on a Model of Continuous Change. *Demography*, 45(4):785–801.

Human-Right-Watch and Johns-Hopkins-Center (2020). People of Resilience: Colombias Wayuu Indigenous Community Confronts a Malnutrition Crisis Amid Covid-19. `https://www.hrw.org/video-photos/interactive/2020/07/29/people-resilience-colombias-wayuu-indigenous-community`. Accessed 10–02–2023.

Hummer, R. A., Maureen, B. R., and G., R. R. (2004). Racial and Ethnic Disparities in Health and Mortality Among the U.S Elderly Population. In Council, N. R., editor, *Critical Perspectives on Racial and Ethnic Differences in Health in Late Life*, chapter 3, pages 53–94. National Academy of Sciences, Washington DC.

Imai, K. and Khanna, K. (2016). Improving Ecological Inference by Predicting Individual Ethnicity from Voter Registration Records. *Political Analysis*, 24:263–272.

Jakobsson, M., Scholz, S. W., Scheet, P., Gibbs, R. J., VanLiere, J. M., Fung, H.-C., Szpiech, Z. A., Degnan, J. H., Wang, K., Guerreiro, R., Bras, J. M., Schymick, J. C., Hernandez, D. G., Traynor, B. J., Simon-Sanchez, J., Matarin, M., Britton, A., vande, L. J., Rafferty, I., Bucan, M., Cann, H. M., Hardy, J. A., Rosenberg, N. A., and Singleton, A. B. (2008). Genotype, haplotype and copy–number variation in worldwide human population. *Nature*, 451(998–1003).

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). Tree-Based Methods. In *An Introduction to Statistical Learning*, chapter 8, pages 303–336. Springer.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2014). Statistical learning. In *An Introduction to Statistical Learning*, chapter 2, pages 15–58. Springer.

Janssens, C. A. and Martens, F. K. (2020). Reflection on modern methods: Revisiting the area under the ROC Curve. *International Journal of Epidemiology*, 49(4):1397–1403.

Jaramillo, U. J. (1964). La población indígena de colombia en el momento de la conquista y sus transformaciones posteriores: primera parte [the indigenous population of colombia at the time of the conquest and its subsequent transformations: Part one].

Jiménez-Peña, O. (2014). La transición sanitaria en Colombia, 1946-2001. `http://rua.ua.es/dspace/bitstream/10045/41904/3/tesis_oscar_mauricio_jimenez_pe%c3%b1a.pdf`. Ph.D. Thesis. Departamento de Enfermería Comunitaria, Medicina

Preventiva y Salud Pública e Historia de la Ciencia. Universidad de Alicante. [Accessed: 2021-02-22].

Johnston, D. W. and Lordan, G. (2016). Discrimination Makes Me Sick! An Examination of the Discrimination Health–Relationship. *Jounal of Health Economics*, 31:99–111.

Jones, C. P. (2000). Levels of Racism: A Theory Framework and a Gardener's Tale. *J Public Health*, 90(8):1212–1215.

Jordan, J. (2004). Afro-colombian. *Souls*, 6(2):19–30.

Jorde, L. B. and Wooding, S. P. (2004). Genetic Variation, Classificationand "Race". *Nature Genetics*, 36(11).

Kalmanovitz, S. and López, E. (2007). Economía colombiana del siglo XX. Un análisis cuantitativo [Colombian economy in the XX century. A cuantitative analysis]. In Robinson, J. and Urrutia, M., editors, *Aspectos de la agricultura colombiana en el siglo XX*, pages 127–172. Fondo de Cultura Económica, Bogotá.

Kateri, H. T. (2013). Brazil: At the forefront of latin american race-based affirmative action policies and census racial data collection. In *Racial Subordination in Latin America: The Role of the State, Customary Law, and the New Civil Rights Response*, chapter 6. Cmabridge University Press.

Kaufman, S. J. (2018). Ethnic Conflict. In Williams, P. D. and McDonald, M., editors, *Security Studies: An Introduction*, chapter 25. Routledge, 3rd edition.

Keita, S. and Kittles, R. A. (1997). The persistence of racial thinking and the myth of racial divergence. *American Anthropologist*, 99(3):534–544.

Keith, V. M. and Herring, C. (1991). Skin Tone and Stratification in the Black Community. *American Journal of Sociology*, 97(3):760–778.

Kemp, A. H., Koenig, J., Thayer, J. F., Bittencourt, M. S., Pereira, A. C., Santos, I. S., Dantas, E. M., Mill, J. G., Chor, D., Ribeiro, A. L., Bensenor, I. M., and Lotufo, P. A. (2016). Race and Resting–state Heart Rate Variability in Brazilian Civil Servants and the Mediating Effects of Discrimination: An ELSA–Brasil Cohort Study. *Psychosomatic Medicine*, 78:950–958.

Kim, J.-S., Gao, X., and Andrey, R. (2018). RIDDLE: Race and ethnicity Imputation from Disease history with Deep LEarning. *PLoS Computational Biology*, 14(4).

King, M., Smith, A., and Gracey, M. (2009). Indigenous health part 2: the underlying causes of the health gap. *Lancet*, 374:76–85.

Kirk, D. (1996). Demographic Transition Theory. *Population Studies*, 50:361–387.

Komahan, K. and Reidpath, D. (2014). A "Roziah" by Any Other Name: A Simple

Bayesian Method for Determining Ethnicity from Names. *American Journal of Epidemiology*, 180(3):325–329.

Kornienko, O., Rambaran, A. J., and Rivas-Drake, D. (2022). Interpersonal racism and peer relationships: An integrative framework and directions for research. *Journal of Applied Developmental Psychology*, 80:101414.

Krieger, N. (1994). Epidemiology and the web of causation: Has anyone seen the spider? *Social Science Medicine*, 39(7):887–903.

Krieger, N. (1999). Embodying inequality: A review of concepts, measures, and methods for studying health consequences of discrimination. *International Journal of Health Services*, 29(2):295–352.

Krieger, N. and Zierler, S. (1997). Accounting for health of women. *Critical Public Health*, 7(1-2):38–49.

Krueger, P. M., Saint Onge, J. M., and Chang, V. W. (2011). Race/ethnicity Differences in adult mortality: The role of perceived stress and health behaviors . *Social Sciences & Medicine*.

Lakha, F., Gorman, D. R., and Mateos, P. (2011). Name analysis to classify populations by ethnicity in public health: Validation of Onomap in Scotland. *Public Health*, 125:688–696.

Landry, A. (1934). La Révolution démographique. *Librairie du Recueil Sirey*. Ed. INED, Paris, 1982.

Lariscy, J. T., Tasmim, S., and Collins, S. (2020). Racial and Ethnic Disparities in Health. In D. Gu, M. D., editor, *Encyclopedia of Gerontology and Population Aging*, pages 1–9. Springer Nature.

Lauderdale, D. and Kestenbaum, B. (2000). Asian American ethnic identification by surname. *Population Research and Policy Review*, 19(3):283–300.

Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S., Cann, H. M., Barsh, G. S., Feldman, M., Cavalli-Sforza, L. L., and Myers, R. M. (2008). Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation. *Science*, 319(5866):1100–1104.

Link, B. G. and Phelan, J. C. (1995). Social Conditions As Fundamental Causes Of Disease. *Journal of Health and Social Behavior*, 35:80–94.

Lotufo, P. A., Goulart, A. C., and Bensenor, I. M. (2007). Race, Gender and Stroke Subtypes Mortality in São Paulo, Brazil. *Arq Neuropsiquiatr*, 65(3-B):752–757.

Lumbiganon, P., Panamonta, M., Laopaiboon, M., Pothinam, s., and Patithat, N. (1990). Why Are Thai Official Perinatal and Infant Mortality Rates So Los? *International Journal of Epidemiology*, 19(4):997–1000.

Lynch, S. M. (2003). Cohort and Life-Course Patterns in the Relationship between Education and Health: A Hierarchical Approach. *Demography*, 40(2):309–331.

Macfarlane, G. J., Lunt, M., Palmer, B., Afzal, C., Silman, A. J., and Esmail, A. (2007). Determining aspects of ethnicity amongst persons of South Asian origin: The use of a surname-classification programme (Nam Pehchan). *Public Health*, 121:231–236.

Manica, A., Prugnolle, F., and Balloux, F. (2005). Goegraphy is a better determinant of human genetic differentiation than ethnicity. *Hum Genet.*, 118(3–4):1–11.

Maniscalco, L., Miceli, S., Bono, F., and Matranga, D. (2020). Sel-Perceived Health, Objective Health, and Quality of Life among People Aged 50 and Over: Interrelationship among Health Indicators in Italy, Spain, and Greece. *International Journal of Environment Research and Public Health*, 17:1–15.

Manish, S. (2017). Understading Digit Preferences in India Using Modified Whipple Index: An Analysis of 640 Districts of India. *International Journal of Current Research*, 9(01). 45144-45152.

Markides, K. S. and Coreil, J. (1986). The Health of Hispanics in the Southwestern United States: an Epidemiologic Paradox. *Public Health Reports*, 101(3):253–265.

Mateos, P., Longley, P. A., and O'Sullivan, D. (2011). Ethnicity and Population Structure in Personal Naming Networks. *PLoS ONE*, 6(9).

McKinley, M. A. (2012). *The Unbearable Lightness of Being (Black): Legal and Cultural Constructions of Race and Nation in Colonial Latin America*, pages 116–142. University of California Press, 1 edition.

Melguizo, R. C. and Cronshaw, F. (2001). The Evolution of Armed Conflict in Medellín: An Analysis of the Major Actors. *Latin American Perspectives*, 28(1):110–131.

Memmi, A., Appiah, K. A., and Martinot, S. (2000). *Definition*, pages 89–122. University of Minnesota Press.

Mendez, F., Urreal-Giraldo, F., and Ortega, D. (2020). Skin color, social inequalities and health in older adults: an based on the SABE survey in Colombia. *Cadernos de Saúde Pública [Reports in Pulbic Health]*, 36(10).

Mikkelsen, L., Phillips, D. E., AbouZahr, C., Setel, P. W., de Savigny, D., Lozano, R., and Lopez, A. D. (2015). A Global Assessment of Civil Registration and Vital Statistics Systems: Monitoring Data Quality and Progress. *Lancet*, 386:1395–406.

Miller, G. (2005). Contraception as Development? New Evidence from Family Planning in Colombia. *National Bureau of Economic Research*. Working Paper 11704.

Montenegro, R. A. and Stephens, C. (2006). Indigenous health in Latin America and the Caribbean. *Lancet*, 367:1859–69.

Moreno, A. R. (2006). Clima change and human health in Latin America: Drivers, effects, and policies. *Regional Environmental Change*, 6:157–164.

Moubarac, J. (2013). Persisting Problems Related to Race and Ethnicity in Public Health and Epidemiology Research. *Revista de Saúde Pública*, 47(1):105–116.

Mukaz, D. K., Zakai, N. A., Cruz-Florez, S., McCullough, L. D., and Cushman, M. (2020). Identifying Genetic and Biological Determinants of Race-Ethnic Disparities in Stroke in the United States. *Stroke*, 51:3417–3424.

Myers, H. F. and Hwang, W.-C. (2004). Cummulative Psychosocial Risk and Resilience: A Conceptual Perspective on Ethnic Health Disparities in Late Life. In National-Research-Council, editor, *Critical Perspectives on Racial and Ethnic Differences in Health in Late Life*, chapter 13, pages 492–539. National Academy of Sciences.

Myers, R. J. (1954). Accuracy of Age Reporting in the 1950 United States Census. *Journal of the American Statistical Association*, 49(268):826–831.

Nanchahal, K., Mangtani, P., Alston, M., and Silva, d. S. I. (2001). Development and validation of a computerized South Asian Names and Group Recognition Algorithm (SANGRA) for use in British Health-related studies. *Journal Of Public Health Medicine*, 23(4):278–285.

National-Research-Council (2004a). Perspectives on Racial and Ethnic Differences. In Bulatao, R. A. and Anderson, N. B., editors, *Understanding Racial and Ethnic Differences in Health in Late Life: A Research Agenda*, chapter 2. The National Academic Press, Washington D.C. Committee on Population, Division of Behavioral and Social Sciences and Education.

National-Research-Council (2004b). Prejudice and Discrimination. In NB, B. R. A., editor, *Understanding Racial and Ethnic Differences in Health in Late Life : A Research Agenda*, chapter 7, pages 76–81. National Academic Press.

National-Research-Council (2004c). Socioeconomic Status. In NB, B. R. A., editor, *Understanding Racial and Ethnic Differences in Health in Late Life : A Research Agenda*, chapter 4, pages 54–60. National Academic Press.

National-Research-Council (2004d). Stress. In NB, B. R. A., editor, *Understanding Racial and Ethnic Differences in Health in Late Life : A Research Agenda*, chapter 8, pages 82–85. National Academic Press.

National-Research-Council (2004e). The Nature of Racial and Ethnic Differences. In Bulatao, R. A. and Anderson, N. B., editors, *Understanding Racial and Ethnic Differences in Health in Late Life*, chapter 1, pages 7–31. The National Academies Press.

National-Research-Council (2004f). The Nature of Racial and Ethnic Differences. In A., B. R. and B., A. N., editors, *Understanding Racial and Ethnic Differences in Health in Late Life : A Research Agenda*, chapter 1, pages 7–31. National Academic Press.

National-Statistics-Office-Thailand (1969). Final report: the survey of population changes; supplementary survey on knowledge of attitude toward, and practice concerning official registration of births and deaths, March 21-April 22, 1966. *National Statistical Office*, pages 01–84. Thailand.

Nazroo, J. and Becares, L. (2020). Evidenceafor Ethnic Inequalities in Mortality Related to Covid-19 Infections: Findings from an Ecological Analysis of England. *British Medical Journal (BMJ) Open*, 10:1–6.

Nazroo, J., Bhui, K., and Rhodes, J. (2020). Where next for understanding race/ethnic inequalities in severe mental illness? Structural, interpersonal and institutional racism. *Sociol Health Illn.*, 42(2):262–276.

Nazroo, J., Williams, D., Marmot, M., and Wilkinson, R. (2006). *The Social Determination of Ethnic/Racial Inequalities in Health*, chapter 12, pages 238–266. Oxford University Press, Oxford, United Kingdom.

Nazroo, J. Y. and Williams, D. R. (2006). *The Social Determination of Ethnic/Racial Inequalities in Health*, chapter 12, pages 238–266. Oxford University Press.

Nembrini, S., König, I. R., and Wright, M. N. (2018). The Revival of the Gini Importance. *Bioinformatics*, 34(21):3711–3718.

Nengwekhulu, R. H. (1986). Race, Class and Ethnicity in South Africa. *African Journal of Political Economy*, 1(1):29–39.

Notestein, F. (1945). Population : the Long View. In Schultz, T., editor, *Food for the World*. Chicago: Univ. of Chicago Press.

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366:447–453.

Office for National Statistics, O. (2020). Coronavirus and the social impacts on different ethnic groups in the UK: 2020. Estimates from the Understanding Society: COVID-19 Study, 2020, UK Household Longitudinal Study (UKHLS) and Wealth and Assets Survey (WAS) to explore the social impacts of the coronavirus (COVID-19) pandemic on people from different ethnic groups in the UK. `https://www.ons.gov.uk/peoplepopulationandcommunity/culturalidentity/ethnicity/articles/coronavirusandthesocialimpactsondifferentethnicgroupsintheuk/2020/pdf`. [Accessed: 2023-04-22].

Office for National Statistics, O. (2021). Review of the current evidence base for population estimates by ethnic group: December 2021. `https://www.ons.gov.uk/peoplepop` ulationandcommunity/culturalidentity/ethnicity/articles/reviewofthecurrentevidencebaseforpopulationestimatesbyethnicgroup/2021-12-16/pdf. [Accessed: 2023-04-22].

Office of Management and Budget (1997). Revisions to the Standards for the Classification of Federal Data on Race and Ethnicity. https://ddd.uab.cat/pub/tesis/2017/hdl_10803_406099/ccf1de1.pdf. [Accessed: 2022-04-22].

Ogedegbe, G., Ravenell, J., Adhikari, S., Butler, M., Cook, T., Francois, F., Iturrate, E., Jean-Louis, G., Jones, S. A., Onakomaiya, D., Petrilli, Christopher, M., Pulgarin, C., Regan, S., Reynolds, H., Seixas, A., Volpicelli, F. M., and Horwitz, L. I. (2020). Assessment of Racial/Ethnic Disparities in Hospitalization and Mortality in Patients With COVID-19 in New York City. *Jama Network*, 12(3):1–14.

O'Grady, A., Balmer, N., Carter, B., Pleasence, P., Buck, A., and Genn, H. (2005). Institutional Racism and Civil Justice, Ethnic and Racial Studies. *Milbank Q*, 28(4):620–628.

Olshanksy, J. S. and Ault, B. A. (1986). The Fourth Stage of the Epidemiologic Transition: The Age of Delayed Degenerative Diseases. *Wiley on behalf of Milbank Memorial Fund*, 64(3):355–391.

Olshansky, J. S., Carnes, B. A., Rogers, R. G., and Smith, L. (1998). Emerging infectious diseases: the Fifth stage of the epidemiologic transition? *World Health Statistics Quarterly*, 51(2-3-4):207–217.

Omran, A. R. (1971). The epidemiologic transition: a theory of the epidemiology of population change. *Milbank Q*, 83(4):731–57.

O'Neil, C. (2016). Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. *Crown New York*.

Oslender, U. (2007). Violence in development: the logic of forced displacement on colombia's pacific coast. *Development in Practice*, 17(6):752–764.

Østergard, U. (1992). What is national and ethnic identity? In *Ethnicity in Hellenistic Egypt*. Aarhus University Press.

Otero, A. (2013). Differencias departamentales en las causes de mortalidad en Colombia [Departamental differences in causes of death in Colombia]. *Centro de Estudios Económicos Regionales (CEER). Banco de la República*, 187. Documentos de trabajo sobre economía regional.

ONeill, L., Fraser, T., Kitchenham, A., and McDonald, V. (2016). Hidden Burdens: a Review of Intergenerational, Historical and Complex Trauma, Implications for Indigenous Families. J Child Adolesc Trauma. *Journ Child Adol Trauma*, 11(2):173–186.

Paéz, G. N., Franco, R. C., and Jaramillo, Q. L. (2012). Estudio sobre la Geografía Sanitaria de

Colombia [Study about the Colombian Health Geography]. *Ministerio de Salud y Protección Social*.

Pager, D. and Shepherd, H. (2008). The Sociology of Discrimination: Racial Discrimination in Employment, Housing, Credit, and Consumer Markets. *Annual Review of Sociology*, 34:181–209.

Palacios, C. A. (2018). The color of child survival in Colombia 1955–2005. *Ethnicity & Health*, 23(2):207–220.

Palacios, M. and Safford, F. (2002). La violencia política en la segunda mitad del siglo xx. In *Historia de Colombia. País fragmentado, sociedad dividida*, chapter 15, pages 493–530. Norma.

Palloni, A. and Arias, E. (2004). Paradox lost: Explaining the hispanic adult mortality advantage. *Demography*, 41(3):385–415.

Paradies, E. A. (2021). The Costs of Institutional Racism and its Ethical Implications for Healthcare. *Bioethical Inquiry*, 18:45–58.

Paradies, Y. (2016). Colonisation, racism and indigenous health. *J Pop Research*, 33:83–96.

Pechholdová, M. (2009). Results and Observations from the Reconstruction of Continuous Time Series of Mortality by Cause of Death: Case of West Germany, 1968-1997. *Demographic Research*, 21(18):535–568. Stand 28.08.2021.

Platt, L. and Warwick, R. (2020). Are some ethnic groups more vulnerable to COVID-19 than others? *Institute for Fiscal Studies*.

Possas, C. (1991). *Epidemiologia e Sociedade: Heterogeneidade Estructural e Saúde no Brazil*. Hucitec, Sao Paulo.

Powe, C. E., Evans, M. K., Wenger, J., Zonderman, A. B., Berg, A. H., Nalls, M., Tamez, H., Zhang, D., Bhan, I., Karumanchi, S. A., Powe, N. R., and Thadhani, R. (2013). Vitamin D–Binding Protein and Vitamin D Status of Black Americans and White Americans. *N Engl J Med.*, 369(21):1–17.

Prasartkul, P. and Vapattanawong, P. (2006). The Completeness of Death Registration in Thailand: Evidence from Demographic Survillance System of the Kanchanaburi Project. *World Health and Population*, 8(3):43–51.

Preston, S. H., Elo, I. T., Rosenwaike, I., and Hill, M. (1996). African-American Mortality at Older Ages: Results of a Matching Study. *Springer*, 33(2):193–209.

Preston, S. H., Elo, I. T., and Stewart, Q. (1999). Effects of Age Misreporting on Mortality Estimates at older Ages. *Population Studies*, 53(2):165–177.

Prohmmo, A. and Guest, P. (1996). Determinants of Reporting of Mortality Data. *Health Systems Research Journal*, 4(4):258–280.

Provost, F. and Fawcett, T. (2001). Robust Classification fro Imprecise Environments. *Machine Learning*, 42:203–231.

Psacharopoulos, G. and Patrinos, H. A. (1994). Pueblos indígenas y pobreza en América Latina: Un análisis empírico. In *Estudios Sociodemográficos de Pueblos Indígenas*, volume 40, pages 417–430. CELADE–UNO, Centro Latinoamericano de Demografía CELADE).

Rahman, A., Nesha, K., Akter, M., and Uddin, S. G. (2013). Application of artifitial neural network and binary Logistic regression in detection of Diabetes status. *Science Journal of Public Health*, 1(1):39–43.

Rao, C., Osterberger, B., Anh, T. D., MacDonald, M., Kim Chúc, N. T., and Hill, P. S. (2009). Compiling mortality statistics from civil registration systems in Viet Nam: the long road ahead. *Bull World Health Organ*, 87:58–65.

Razum O, Zeeb H, A. S. (2001). How useful is a name-based algorithm in health research among Turkish migrants in Germany? *Tropical Medicine and International Health*, 6(8):654–661.

Reskin, B. (2012). The race discrimination system. *Annual Review of Sociology*, 38(Volume 38, 2012):17–35.

Ribotta, B. S., Salazar, A. L., and Bertone, C. L. (2019). Evaluations of the Life Statistics Coverage at a Subnational Level. Recent Studies in Latin America. *Rev. Gerenc. Polit. Salud*, 18(36):1–36.

Robey, B., Rutstein, S. O., and Morris, L. (1993). The Fertility Decline in Developing Countries. *Scientific American*, 269(6):60–67.

Rodríguez, M. M. M. (2010). La invisibilidad estadística étnico-racial negra, afrocolombiana, raizal y palenquera en Colombia [The statistical invisibility of black, Afro-Colombian and Palenque-rooted populations in Colombia]. *Trabajo Social*, 12.

Rogers, R. G. and Hackenberg, R. (1987). Extending epidemiologic transition theory: A new stage. *Social Biology*, 34(3–4):234–243.

Rogers, R. G., Hummer, R. A., and Nam, C. B. (2000). The Effects of Basic Socioeconomic Factors on Mortality. In Elsevier, editor, *Living and Dying in the USA: Behavioral, Health, and Social Differentials of Adult Mortality*, chapter 7, pages 115–139. Academic Press.

Ronan, F. (2014). Indigenous Narratives of Health: (Re)placing Folk-Medicine within Irish Health Histories. *J Med Humanit*, 36:5–18.

Ruiz, S. M. (1988). El Registro de Defunción con Particular Referencia a Colombia. *Biomédica*, 8(3–4):94–98.

Saint Onge, J. M. and Krueger, P. M. (2017). Health Lifestyle behaviors among U.S. adults. *SSM-Population Health*.

Sankar, P., Cho, M. K., Condit, C. M., Hunt, L. M., Koenig, B., Marshall, P., Lee, S.-J. S., and Spicer, P. (2004). Genetic Research and Health Disparities. *JAMA*, 291(24):2985–2989.

Sarmiento, M. C. (2000). Enfermedades Transmisibles en Colombia: Cambios Ambivalentes [Communicable Diseases in Colombia: Ambivalent Changes]. *Salud pública*, 2(1):82–93.

Schöley, J. and Willekens, F. (2017). Visualizing compositional data on the Lexis surface. *Demographic Research*, 36(21):627–658.

Scholz, R. and Jdanov, D. (2008). Weniger Hochbetagte als gedacht: Korrekturen in der amtlichen Statistik für Westdeutschland notwendig [Fewer elderly people than expected: necessary corrections in official statistics for West Germany]. *Demografische Forschung, Aus Erster Hand*, 5(1).

Schuitemaker, N., Van Roosmalen, J., Dekker, G., Van Dongen, P., Van Geijn, H., and Gravenhorst, J. B. (1997). Underreporting of Maternal Mortality in The Netherlands. *Obstretrics & Gynecology*, 90(1).

Schweitzer, A. (1947). *An Anthology*. The Beacon Press.

Seffens, W., Evans, C., Minority Health-GRID, N., and Taylor, H. (2015). Machine Learning Data Imputation and Classification in a Multicohort Hypertension Clinical Study. *Bioinformatics and Biology Insights*, 9:43–54.

Setel, P. W., Sankoh, O., Rao, C., Velkoff, V. A., Mathers, C., Gonghuan, Y., Hemed, Y., Jha, P., and Lopez, A. D. (2005). Sample registration of vital events with verbal autopsy: a renewed commitment to measuring and monitoring vital statics. *Bulletin of the World Health Organization*, 83:611–617.

Sheth, T., Nair, C., Nargundkar, M., Anand, S., and Yusuf, S. (1999). Cardiovascular and cancer mortality among Canadians of European, south Asian and Chinese origin from 1979 to 1993: an analysis of 1.2 million deaths. *Canadian Medical Association Journal*, 161(2).

Silva, S. and Kenney, M. (2018). Algorithms, platforms, and ethnic bias: An integrative essay. *Phylon (1960-)*, 55(1 & 2):9–37.

Sirugo, G., Williams, S. M., and Tishkoff, S. A. (2019). The missing diversity in human genetic studies. *Cell*, 177:26–31.

Smedley, A. (2007). *Race in North America : origin and evolution of a worldview*. Westview Press, Boulder, CO, 3th ed. edition.

Smith, J. P. (1999). Healthy Bodies and Thick Wallets: The Dual Relation Between Health and Economic Status. *Journal of Economic Perspectives*, 13(2):145–166.

Soledad, S. J. and Jiménez, E. C. (2011). Desplazamiento forzado de los grupos Ãľtnicos en Colombia [Forced displacement of ethnic groups in Colombia]. In ALAP, editor, *Pueblos In-*

dígenas y afrodescendientes en américa Latina: dinámicas poblacionales diversas y desafíos comunes. UNFPA–United Nations, first edition. Coordinators: Del Popolo Fabiana, Estela Maria Garcia de Pinto da Cunha, Bruno Ribbota and Marta Azevedo.

Spijker, J., Recaño, J., Martínez, S., and Carioli, A. (2020). Mortality by cause of death in Colombia: a local analysis using spatial econometrics. *Journal of Geographical Systems*.

Stopforth, S., Kapadia, D., Nazroo, J., and Bécares, L. (2022). The enduring effects of racism on health: Understanding direct and indirect effects over time. *SSM Popul Health*, 28(19):1–11.

Sun, J., Guo, X., Lu, Z., Fu, Z., Li, X., Chu, J., Zhang, G., Kou, K., Xue, F., and Xu, A. (2018). The gap between cause-of-death statistics and Household Registration reports in Shandong, China during 2011-2013: Evaluation and adjustment for underreporting in the mortality data for 262 subcounty level populations. *PLoS ONE*, 13(6):1–12.

Tamargo, M. d. C. (2007). El Subregistro de Nacimientos: El análisis de las variables de género y etnia en Ecuador. *Banco Interamericano de Desarrollo*, pages 1–58.

Telles, E., Rene, F. D., and Fernando, U.-G. (2015). Pigmentocracies: Educational inequality, skin color and census ethnoracial identification in eight Latin American countries. *Research in Social Stratification and Mobility*, 40:39–58.

Thoma, Martin (2018). Receiver Operating Characteristic (ROC) curve with False Positive Rate and True Positive Rate. `https://commons.wikimedia.org/wiki/File:Roc-draft-xkcd-style.svg`, Last accessed on 2023-07-03.

Tin Oung, M., Richter, K., Prasartkul, P., and Tangcharoensathien, V. (2017). Myanmar mortality registration: an assessment for system improvement. *Population Health Metrics*, 15(34).

Trovato, F. (2001). Aborginal Mortality In Canada, The United States and New Zealand. *J. biosoc. Sci.*, 33:67–86.

Tsung-Hsueh, L., Craig, J. R., Meng-Chih, L., Ming-Chih, C., and Tai-Ping, S. (2002). High frequency death in Taiwan: a sociocultural product. *Social Science and Medicine*, 55:1663–1669.

UNHCR (2020). Internal Displacement/Colombia Large-Group Internal Displacement. `https://data2.unhcr.org/en/documents/download/84007`. [Accessed: 2023-30-05].

United-Nations (2010). Mortalidad infantil y en la niñez de pueblos indígenas a afrodescendientes de América Latina: Inequidades estructurales, patrones diversos y evidencia de derechos no cumplidos [Child mortality and Childhood in Indigenous and Afro–descendant populations in Latin American countries. Structural inequalities, diverse patterns, evidence of unfulfilled rights]. *United Nations*. CEPAL, UNFPA, PAHO.

Urdinola, P. B., Torres, A. F., and Velasco, J. A. (2017). The Homicide Atlas in Colombia: Contagion and Under-Registration for Small Areas. *Revista Colombiana de Geografía*, 26:101–118.

Uribe, M. M. (1998). Los grupos étnicos de Colombia: intentos de cuantificación y criterios para el censo 1993 [The ethnic groups in Colombia: Attempts of quantification and criterions for the Census 1993]. *Departamento Administrativo Nacional de Estadistica (DANE)).* Thesis in Anthropology. Universidad de los Andes.

Urrea-Giraldo, F., Carabalí, H. B., Quiroz, c. L., Valoyes, V. V., Romero, B. D., Guzmán, G. S., and Obando, J. S. (2020). Análisis de la Dinámica Intercensal del Autorreconocimiento en la Población Negra, Afrocolombiana, Raizal y Palenquera en el Periodo 2005-2018 [Analysis of the Intercensal Dynamics Regarding Self-Reported Ethnicity in the Black Population, Afro-Colombian, Black Maroon and Black Caribbean During the Period 2005-2018]. *Working Paper*, pages 01–99. Non-published document.

Urrea-Giraldo, F., G., B. P., Carabalí, H. B., and Muñoz, V. V. (2015). Patrones de mortalidad comparativos entre la población afrodescendiente y la blanca–mestiza para Cali y el Valle del Cauca. *Revista CS Universidad ICESI*, pages 135–171.

Urrea-Giraldo, F. and Hurtado, S. T. (2002). La Construcción de las Etnicidades en la Sociedad Colombiana Contenporánea: Un Caso Ejemplar para una Discusión Sobre Etnicidad y Grupos Raciales [The Construction of Ethnicities in Contemporary Colombian Society: An Exemplary Case for a Discussion on Ethnicity and Racial Groups] . *Interculturalidad y Política, Pontificia Universidad Católica del Perú*, pages 1–26.

Urrea-Giraldo, F., Ramírez, H. F., Carabalí, H. B., Ortega, L. D., Loja, J., Quiroz, c. L., Valoyes, V. V., Romero, B. D., and González, O. N. (2021). Brechas Étnico Raciales en Colombia [Racial-Ethnic Social Gap in Colombia]. *Report*, pages 01–153. Non-published document.

Urrea-Giraldo, F., Viafara, L. C. A., and Mara, V. V. (2014). From Whitened Miscegenation to Tri-Ethnic Multiculturalism. In Telles, E., editor, *Pigmentocracies, Ethnicity, Race, and Color in Latin America*, chapter 3. The University of North Carolina Press, Chapel Hill.

Vallin, J. and Meslé, F. (2004). Convergences and divergences in mortality. A new approach to health transition. *Demoraphic Research*, 2(2):11–44.

van Raalte, A. and Caswell, H. (2013). Perturbation Analysis of Indices of Lifespan Variability. *Demography*, 50(6):1615–1640.

VanDeVeer, D. (2014). *Paternalistic intervention: The moral bounds on benevolence*, volume 59. Princeton University Press.

Vaupel, J. W. (2001). Mortality of the Oldest Old. In SMELSER, N. J. and BALTES, P. B., editors, *International Encyclopedia of the Social and Behavioral Sciences*, volume 15, pages 10075–10079.

Vaupel, J. W. and Canudas-Romo, V. (2003). Decomposing Change in Lige Expectancy: A Bouquet of Formulas in Honor of Nathan Keyfitz's 90th Birthday. *Demography*, 40(2):201–216.

Vaupel, J. W., Zhang, Z., and van Raalte, A. (2011). Life Expectancy and Disparity: An International Comparison of Life Table Data. *British Medical Journal (BMJ) Open*, 1(1):1–6. Publication Includes Supplementary Material.

Viáfara, L. C. A. (2017). Diferenciales de ingreso por el color de la piel y desigualdad de oportunidades en Colombia. *Revista de Economía del Rosario*, 20(1):97–126.

Vilella, A. and Trilla, A. (2021). The Covid-19 Pandemic -an Epidemiological Perspective. *Current Allergy and Asthma Reports*, 21(29).

Voicu, I. (2018). Using First Name Information to Improve Race and Ethnicity Classification. *Statistics and Public Policy*, 5(1):1–13.

Wade, P. (1985). Race and Class: the case of South American blacks. *Ethnic and Racial Studies*, 8(2):233–249.

Wade, P. (1993a). Blackness and Race Mixture: The Dynamics of Racial Identity in Colombia. *Baltimore Johns Hopkins University Press*, page 416.

Wade, P. (1993b). The Racial Order and National Identity. In Scott, J. R., Mintz, S., and Trouillot, M.-R., editors, *Blackness and Race Mixture, The Dynamics of Racial Identity in Colombia*. Baltimore Johns Hopkins University Press.

Wade, P. (2008). Race in Latin America. In Poole, D., editor, *A Companion to Latin American Anthropology*, pages 177–192. Blackwell Publishing.

Wade, P. (2009). *Race and Sex in Colonial Latin America*, pages 61–109. Pluto Press.

Wade, P. (2010a). *Blacks and Indigenous People in Latin America*, pages 24–40. Pluto Press.

Wade, P. (2010b). Blacks and Indigenous People in the Postmodern and Postcolonial Nation and Beyond. In *Race and Ethnicity in Latin America*. Pluto Press. JSTOR.

Wade, P. (2012). Skin colour and race as analytic concepts . *Ethnic and Racial Studies*, 35(7):1169–1173.

Wakefield, J. (2004). A critique of statistical aspects of ecological studies in spatial epidemiology. *Environmental and Ecological Statistics*, 11:31–54.

Warner, D. F. and Hayward, M. D. (2006). Early–Life Origins of the Race Gap in Mens's Mortality'. *J Health Soc Behav.*, 47(3):209–226.

Warren, T. (1929). Population. *American Journal of Sociology*, 34(6):959–975.

WHO (2021). International Statistical Classification of Diseases and Related Health Poblems (ICD). `https://www.who.int/standards/classifications/classification-of-diseases`.

Williams, D. R. (1999). Race, socioeconomic status, and health. The added effects of racism and discrimination. *Ann N Y Acad Sci*, pages 173–188.

Williams, D. R. and Collins, C. (2001). Racial Residential Segregation: A Fundamental Cause of Racial Disparities in Health. *Public Health Reports*, 116:404–416.

Williams, D. R., Neighbors, H. W., and Jackson, J. S. (2003). Racial/Ethnic Discrimination and Health: Findings From Community Studies . *American Journal of Public Health*.

Wilmoth, J. R. and Horiuchi, S. (1999). Rectanguralization Revisted: Variability of Age at Death Within Human Populations. *Demography*, 36(34):475–495.

Wilson, E. O. and Brown, J. W. L. (1953). The Subspecies Concept and Its Taxonomic Application. *Taylor & Francis*, 2(3):97–111.

Winkleby, M. A. and Cubbin, C. (2004). Racial/Ethnic Disparities in Health Behaviors: A Challenge to Current Assumptions. In Council, N. R., editor, *Critical Perspectives on Racial and Ethnic Differences in Health in Late Life*, chapter 12, pages 450–491. National Academy of Sciences.

Winner, L. (1980). Do Artifacts Have Politics? *The MIT Press*, 109(1):121–136.

Wohland, P., Rees, P., Nazroo, J., and Jagger, C. (2015). Inequalities in healthy life expectancy between ethnic groups in England and Wales in 2001. *Ethnicity & Health*, 20(4):341–353.

Womack, L. S., Rossen, L. M., and Hirai, A. H. (2020). Urban–Rural Infant Mortality Disparities by Race and Ethnicity and Cause of Death. *American Journal of Preventive Medicine*, 58(2):254–260.

Wu, S., Wang, R., Zhao, Y., Ma, X., Wu, M., Yan, X., and He, J. (2013). The relationship between self-rated health and Objective health status: a population-based study. *BioMed Central (BMC) Public Health*, 13(320):1–9.

Wu, X., Nethery, R. C., Sabath, M. B., Braun, D., and Dominici, F. (2020). Air pollution and covid-19 mortality in the united states: Strengths and limitations of an ecological regression analysis. *Science Advances*, 6(45):eabd4049.

Xia, H., Daley, B. J., Petrie, A., and Zhao, X. (2012). A Neural Network Model for Mortality Prediction in ICU. *Computing in Cardiology*, 39:261–264.

Xue, Y., Harel, O., and Aseltine, R. J. (2019). Comparison of Imputation Methods for Race and Ethnic Information in Administrative Health Data. In *13th International conference on Sampling Theory and Applications (SampTA)*, pages 1–4. IEEE.

Yamada, C. and Takahashi, E. (2012). Self-rated health as comprehensive indicator of lifestyle-related health status. *Environ Health Preventive Medicine*, 17:457–462.

Yang, G., Hu, J., Rao, K. Q., Ma, J., Rao, C., and Lopez, A. D. (2005). Mortality registration and surveillance in China: History, current situation and challenges. *Population Health Metrics*, 3(3):1–9.

Yehuda, R. and Bierer, L. M. (2009). The relevance of epigenetics to ptsd: Implications for the dsm-v. *Journal of Traumatic Stress*, 22(5):427–434.

Zhang, Z. and Vaupel, J. W. (2009). The Age Separating Early Deaths from Late Deaths. *Demographic Research*, 20(29):721–730.

Zschirnt, E. and Ruedin, D. (2016). Ethnic Discrimination in Hiring Decisions: A Meta–Analysis of Correspondence Test 1990–2015. *Journal of Ethnic and Migration Studies*.

*

APPENDIX A

Variables of Death Certificates

A.1 Death Certificate Issued from 1980 to 1991

Death certificates from 1980 to 1991 have $1,758,625$ registers (25% of the total deaths). This certificate included 15 variables, as follows:

- **Variable 1**: Province where the death occurred
- **Variable 2**: Municipality where the death occurred
- **Variable 3**: Area where the death occurred (1 = urban, 2 = rural, 3 = no information).
- **Variable 4**: Month when the death occurred
- **Variable 5**: Year when the death occurred
- **Variable 6**: Sex (1 = male, 2 = female)
- **Variable 7**: Age-group of death (1 = less than one day, 2 = between 1 and 6 days, 3 = between 7 and 29 days, 4 = between 1 and 5 months, 5 = between 6 and 11 months, 6 = one year, 7 = between 2 and 4 years, 8 = between 5 and 9 years....23 = between 80 and 84 years, 24 = more than 85 years, 25=unknown

- **Variable 8**: Age of death, single age

- **Variable 9**: Marital status (1 = single, 2 = married, 3 = widowed, 4 = cohabitation, divorced, other, 5 = no information)

- **Variable 10**: Province of permanent residence of the deceased person

- **Variable 11**: Municipality of permanent residence of the deceased person

- **Variable 12**: Place of death (1 = hospital or clinic, 2 = home, 3 = other, 4 = no information

- **Variable 13**: Main cause of death (according to ICD-9)

- **Variable 14**: Person who certify the death 1 = family doctor, 2 = other doctor, 3 = no certified)

- **Variable 15**: Cause of death (according to the list 105 Colombia).

A.2 Death Certificate Issued from 1992 to 1996

Death certificate issued from 1992 to 1996 contains 18 variables and $848,360$ registers (12% of the total deaths). three new variables were added to the former certificate:

- **Variable 16**: Time of residence in the municipality where the death occurred (1 = less than one day, 2 = between 1 and 6 days, 3 = between 7 and 29 days, 4 = between 1 and 5 months, 5 = between 6 and 11 months, 6 = unknown, 7 = more than 1 year)

- **Variable 17**: Area of permanent residence of the deceased person (1 = urban, 2 = rural, 3 = no information)

- **Variable 18**: Age group (1 = <1 year, 2 = 1-4 years, 3 = 5-14 years, 4 = 15-44 years, 5 = 45-64 years, 6 = 65+, 7 = unknown.

A.3 Death Certificate Issued in 1997

Death certificate from 1997 onwards included a new variable, and $170,753$ registers (3% of the total deaths):

- **Variable 19**:Kind of death (1 = natural death, 2 = by external causes, 3 = in process to be determined.

A.4 Death Certificate Issued from 1998 to 2000

Thirty new variables were added to the mortality registration form and $546,348$ (8%) deaths were registered during this period.

- **Variable 20**: Kind of death 2 (1 = fetal death, 2 = non-fetal death)
- **Variable 21**: Hospital/clinic reference code
- **Variable 22**: Hospital/clinic name
- **Variable 23**:Educational level of the deceased person (1 = preschool, 2 = primary school incomplete, 3 = primary school complete, 4 = Secondary school complete, 5 = secondary school incomplete, 6 = university education complete, 7 = university education incomplete, 8 = no formal education, 9 = no information)
- **Variable 24**: Social insurance of the deceased person (1 = contributor scheme, 2 = subsidized scheme, 3 = affiliated, 4 = non-affiliated, 5 = no information)
- **Variable 25**:country of residence of the deceased person
- **Variable 26**: Childbirth-related death (1 = before childbirth, 2 = during childbirth, 3 = after childbirth, 4 = ignored, 5 = no information)
- **Variable 27**: Type of childbirth (1 = spontaneous, 2 = cesarean, 3 = ignored, 4 = no information)
- **Variable 28**: Type of pregnancy (1 = simple, 2 = multiple, 3 = no information)
- **Variable 29**: Gestation time (1 = $<$ 20 weeks, 2 = 20-27 weeks, 3 = 28+ weeks, 4 = ignored, 5 = no information)
- **Variable 30**: weight, weigh in grams of the deceased newborn
- **Variable 31**:Age of the mother of the deceased newborn
- **Variable 32**: Number of live births of the mother
- **variable 33**: Number of stillborn of the mother, including the current deceased newborn
- **Variable 34**: Marital status of the mother ((1 = single, 2 = married, 3 = widowed, 4 = cohabitation, divorced, other, 5 = no information)

- **Variable 35**: Educational level of the mother (1 = preschool, 2 = primary school incomplete, 3 = primary school complete, 4 = Secondary school complete, 5 = secondary school incomplete, 6 = university education complete, 7 = university education incomplete, 8 = no formal education, 9 = no information)

- **Variable 36**: Pregnancy in deceased women (1 = if the woman was pregnant, 2 = no pregnant, 3 = no information)

- **Variable 37**: Pregnancy in the last week before death (1 = if the woman was pregnant one week before death, 2 = if she was not pregnant, 3 = no information)

- **Variable 38**: Pregnancy in the last month before death (1 = if the woman was pregnant one month before death, 2 = if she was not pregnant, 3 = no information)

- **Variable 39**: Kind of external cause of death (1 = suicide, 2 = homicide, 3 = traffic accident, 4 = other accidents, 5 = in process to be determined, 6 = no information)

- **Variable 40**: Province where the external cause occurred

- **Variable 41**: Municipality where the external cause occurred

- **Variable 42**: Mean by which the caused of death was determined (1 = necropsy, 2 = medical history, 3 = laboratory test, 4 = through family or friends, 5 = no information)

- **Variable 43**: Medical assistance before death (1 = there was medical assistance, 2 = there was not medical assistance, 3 = ignored information, 4 = no information)

- **Variable 44**: Direct cause of death according to ICD-10

- **Variable 45**:Related cause 1 of death according to ICD-10

- **Variable 46**: Related cause 2 of death according to ICD-10

- **Variable 47**: Related cause 3 of death according to ICD-10

- **Variable 48**: Other important pathological conditions according to ICD-10

- **Variable 49**: Group of causes of death according to the list 6/67 of the Pan American Health Organisation (PAHO) / World Health Organisation (WHO)

A.5 Death Certificate Issued from 2001 to 2007

In this period $1,340,601$ (19% of the total deaths) deaths were registered, and five additional questions were added, but they were discontinued in the year 2008

- **Variable 49a**: Date of birth of the deceased person.
- **Variable 49b**: Number of weeks of pregnancy
- **Variable 49c**: Number of weeks of the gestational pregnancy: 1 = less than 22 weeks, 2 = between 22 and 27 weeks, 3 = more than 28 weeks, 4 = ignored, 5 = no information
- **Variable 49d**: Mother's ID-number' of the deceased child
- **Variable 49e**: Date of issue of the death certificate.

A.6 Death Certificate Issued from 2008 to 2018

Death certificate from 2008 to 2018 contains $2,309,658$ (33% of the total) registered deaths and 67 variables, that means, 18 variables more than the former certificate.

- **Variable 50**: Place of death (other place)
- **Variable 51**: Hour in which the death occurred
- **Variable 52**: Minute in which the death occurred
- **Variable 53**: Last year of formal education completed
- **Variable 54**: Work-related death (1 = yes, it is related, 2 = no, it is not related)
- **Variable 55**: For work-related deaths (1 = accident, 2 = disease)
- **Variable 56**: Last permanent occupation of the deceased person
- **Variable 57**: Ethnicity of the deceased person (1 = Indigenous, 2 = ROM (Gipsy), 3 = Black of the Colombian Caribbean islands, 4 = Black of San Basilio Town, 5 = Black, Mulatto, Afro-descendant, 6 = Non-ethnic)
- **Variable 58**: Type of health insurance company
- **Variable 59**: Name of the health insurance company

- **Variable 60**: For newborn deaths, last year of education completed of the mother
- **Variable 61**:Means by which the cause of death was determined (A = autopsy, B = medical history, C = laboratory test, D = through family or friends, E = no information)
- **Variable 62**: Second direct cause of death according to ICD-10
- **Variable 63**: Related cause 4 of death according to ICD-10
- **Variable 64**: Related cause 5 of death according to ICD-10
- **Variable 65**: Related cause 6 of death according to ICD-10
- **Variable 66**: Other pathological conditions 2 according to ICD-10
- **Variable 67**: Occupation of the person who certified the death (1 = physician, 2 = nurse, 3 = nursing assistant, 4 = health promoter, 5 = civil registry official)

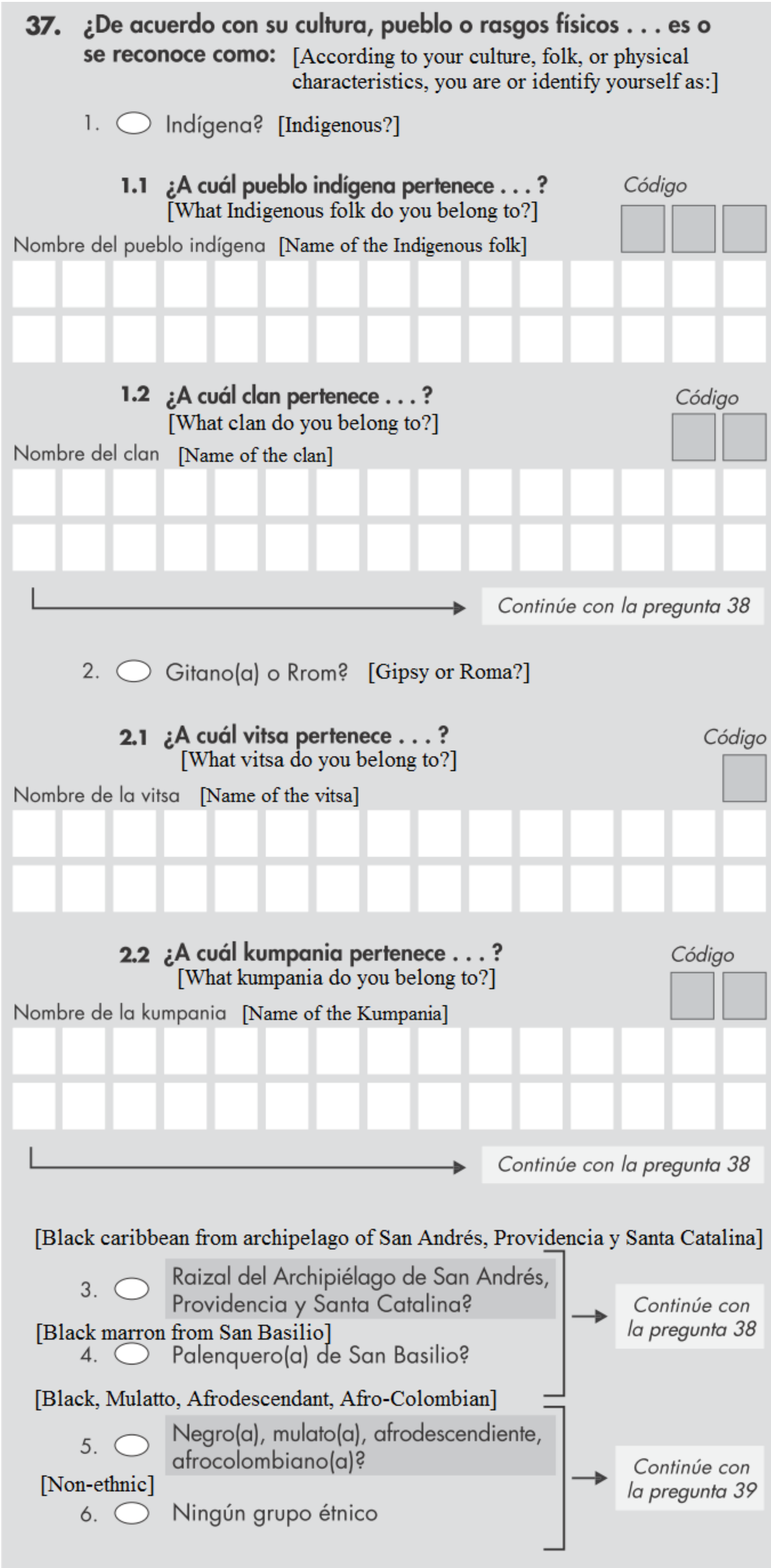
The use of the datasets was approved by the Ethics and Research Governance of the University of Southampton.

APPENDIX B

Variables of the Census

B.1 Questions about Ethnicity and Mortality in 2018 Census

Figure B.1: Colombian Census 2018, Question for Collection of Ethnic Information



37. ¿De acuerdo con su cultura, pueblo o rasgos físicos . . . es o se reconoce como: [According to your culture, folk, or physical characteristics, you are or identify yourself as:]

1. Indígena? [Indigenous?]

1.1 ¿A cuál pueblo indígena pertenece . . . ? [What Indigenous folk do you belong to?] *Código*

Nombre del pueblo indígena [Name of the Indigenous folk]

1.2 ¿A cuál clan pertenece . . . ? [What clan do you belong to?] *Código*

Nombre del clan [Name of the clan]

Continúe con la pregunta 38

2. Gitano(a) o Rrom? [Gipsy or Roma?]

2.1 ¿A cuál vitsa pertenece . . . ? [What vitsa do you belong to?] *Código*

Nombre de la vitsa [Name of the vitsa]

2.2 ¿A cuál kumpania pertenece . . . ? [What kumpania do you belong to?] *Código*

Nombre de la kumpania [Name of the Kumpania]

Continúe con la pregunta 38

[Black caribbean from archipelago of San Andrés, Providencia y Santa Catalina]

3. Raizal del Archipiélago de San Andrés, Providencia y Santa Catalina?

[Black marron from San Basilio]

4. Palenquero(a) de San Basilio?

Continúe con la pregunta 38

[Black, Mulatto, Afrodescendant, Afro-Colombian]

5. Negro(a), mulato(a), afrodescendiente, afrocolombiano(a)?

[Non-ethnic]

6. Ningún grupo étnico

Continúe con la pregunta 39

Figure B.2: Colombian Census 2018, Question for Collection of Mortality Information

26. ¿Cuántas personas que eran miembros de este hogar fallecieron en el 2017?
[How many members of the household died in the year 2017]

Total

(Si es 0, continúe con la pregunta 27)
(Si es 1 o más, relaciónelos en la siguiente tabla)

[Number] Número de orden	[Sex] Sexo		[Age of death] Edad al morir (para menores de 1 año escriba 0)	[Is there a certificate of the decease?] ¿Se expidió certificado de defunción?		
	1. Hombre [Male]	2. Mujer [Female]		1. Sí [Yes]	2. No [No]	3. No sabe [N/A]