UNIVERSITÀ DEGLI STUDI DI PAVIA

SCUOLA DI ALTA FORMAZIONE DOTTORALE

DOTTORATO DI RICERCA IN PSICOLOGIA, NEUROSCIENZE E STATISTICA MEDICA

Coordinatore: Chiar.ma Prof.ssa Gabriella Bottini

HEALTH IMPACT OF THE EMISSIONS FROM A REFINERY: CASE-CONTROL STUDY ON THE ADULT POPULATION LIVING IN TWO MUNICIPALITIES IN LOMELLINA, ITALY

Tutor:

Chiar.ma Prof.ssa Simona Villani

Candidato:

Marco Gnesi

ANNO ACCADEMICO 2017/2018

Compré varios botes de óleo y una tela de dos por dos, diciéndome que si fracasaba era mejor hacerlo a lo grande. Coloqué el lienzo en el suelo y con una precaución algo impostada marqué un gran punto negro en lo que podía ser su parte central: enseguida me pareció un gesto petulante. Pensé que debía "complejizar" aquel signo y extraerlo de las fauces de la boutade. Pinté otro punto del mismo color y tamaño a su lado: el cuadro se volvió insoportablemente "dialéctico". Además parecía que me miraba. Para destruir la simetría hice un nuevo punto lejos del centro. La tela se desordenó con tan sólo tres intervenciones.

Valentín Roma, *El enfermero de Lenin*
Editorial Periférica, 2017

INDEX

THESIS

APPENDIX

ABBREVIATIONS
(in alphabetical order)

A	AERMIC	AERMOD Improvement Committee
	AERMOD	AMS and US EPA Regulatory Model
	AIA	Integrated Environmental Authorisation (*Autorizzazione Integrata Ambientale*)
	AIC	Akaike Information Criterion
	AMS	American Meteorological Society
	ANOVA	Analysis of Variance
	ARPA	Regional Environmental Protection Agency (*Agenzia Regionale per la Protezione dell'Ambiente*)
	ATS	Health Protection Agency (*Agenzia per la Tutela della Salute*)
B	BIC	Bayesian Information Criterion
	BMI	Body Mass Index
	BPD	Barrels Per Day
C	CI	Confidence Interval
	CO	Carbon monoxide
	CONSAL	Knowledge and Health Project (*Progetto Conoscenza e Salute*)
	COPD	Chronic Obstructive Pulmonary Disease
	CRS	Coordinate Reference System
D	DALYs	Disability-Adjusted Life Years
	DF (df)	Degrees of Freedom
	DGR	Resolution of the Regional Government (*Deliberazione della Giunta Regionale*)
	D.L.	Decree (*Decreto Legge*)
E	EU	European Union

F	FE	Ferrera Erbognone
G	GEH	Global Environmental Health
H	HIA	Health Impact Assessment
	HRA	Health Risk Assessment
I	ICD-IX-CM	International Classification of Diseases – IX-CM
	ICS	Industrially Contaminated Site
	IQR	Inter-Quartile Range
	ISCST3	Industrial Source Complex Short-Term model
	ISPRA	Italian Environmental Protection Agency (*Istituto Superiore per la Protezione e la Ricerca Ambientale*)
	ISTAT	National Institute of Statistics (*Istituto Nazionale di Statistica*)
K	KW	Kruskal-Wallis
L	L.	Law (*Legge*)
	LN (ln)	Napierian Logarithm
	LR	Likelihood-Ratio
M	MONITER	Monitoring of Incinerators in the territory of Emilia-Romagna (*Monitoraggio Inceneritori nel Territorio dell'Emilia-Romagna*)
	MS	Mean Sum of Squares (variance)
	MW	Mann-Whitney
N	NOx	Nitrogen oxides
	NO2	Nitrogen dioxide
O	OR	Odds Ratio

P	PAH	Polycyclic Aromatic Hydrocarbons
	PBL	Planetary Boundary Layer
	PDF	Probability Density Function
	PHA	Public Health Assessment
	PM	Particulate Matter
	PM10	Particulate Matter, $\phi \leq 10$ μg/m^3
	PRIME	Plume Rise Model Enhancement
Q	Q	Question
R	ReNCaM	Registry of Causes of Mortality (*Registro Nominativo delle Cause di Morte*)
	RR	Relative Risk
	RV	Ratio of Variances
S	SBL	Stable Boundary Layer
	SD	Standard Deviation
	SdB	Sannazzaro de' Burgondi
	SDG	Sustainable Development Goal
	SDO	Hospital Discharge Record (*Schede di Dimissione Ospedaliera*)
	SE	Standard Error
	SENTIERI	National Epidemiological Study of Territories and Settlements Exposed to Risks from Pollution (*Studio Epidemiologico Nazionale dei Territori e degli Insediamenti Esposti a Rischio da Inquinamento*)
	SOx	Sulphur oxides
	SO2	Sulphur dioxide
U	US EPA	Environmental Protection Agency of the United States of America
	UTM	Universal Transverse Mercator
V	VIA	Environmental Impact Assessment (*Valutazione di Impatto Ambientale*)
	VIF	Variance Inflation Factor

	VIIAS	Integrated Environmental and Health Impact Assessment (*Valutazione Integrata dell'Impatto Ambientale e Sanitario*)
	VIS	Health Impact Assessment (*Valutazione di Impatto sulla Salute*)
	VOC	Volatile Organic Compunds
W	WGS84	World Geodetic System 1984
	WHO	World Health Organisation
	WT	Wald's test
#	95%CI	Confidence Interval, confidence level 95%

ABSTRACT

Air pollution is perhaps the most relevant environmental risk of our era: it is considered responsible of one ninth of all the deaths occurring worldwide, affecting every component of the society in any corner of the world. In fact, its reduction is an indicator of sustainable development [*WHO, 2016*].

Industrially contaminated sites, in particular, are a relevant issue for *environmental health* because they may be a harm for public health [*PRÜSS-USTÜN ET AL., 2016*]. At the same time, those sites are often located in socio-economically deprived districts [*MARSILI, 2016*], a fact that could strengthen their negative impacts by interacting with other health determinants: in other words, they are a concern also for *environmental justice* [*WHO, 2013*]. Investigation in environmental health is itself intricate, as it requires to integrate epidemiology, medicine and toxicology with environmental sciences, but its social dimension implies a greater complexity: robust scientific methodology, in this context, should go with consideration of the specific circumstances and urgencies expressed by stakeholders, like worries or needs to be properly informed.

In the district of Lomellina (Province of Pavia, Region of Lombardy, Italy), and specifically in the municipalities of Sannazzaro de' Burgondi and Ferrera Erbognone, an oil refinery is operating since 1963 and, as of 2018, it still represents a major player for the socio-economical and occupational standards of the area. Remarkably, Lomellina is in the Po Valley, one of the areas with the worse air quality standards in Europe [*EEA, 2016*]. In 2008, the private company running the plant (ENI S.p.A.) asked the competent authorisation bodies to set up a new facility ("EST"), which theoretically should increase the yield of the refining process and reduce the emission of pollutants (except for carbon dioxide). The authorisation decree, issued in 2010, was conditional to the implementation of surveillance activities and, notably, of an epidemiological study investigating public health before and after the commissioning of EST; the company was mandated to assume all the costs of these activities.

According to the authorisation decree, the Department of Public Health, Experimental and Forensic Medicine at the University of Pavia developed the CONSAL Project (*Conoscenza e Salute*, in English *Knowledge and Health*), aimed at investigating public health among the adult population living in Sannazzaro de' Burgondi and Ferrera Erbognone. The protocol was approved by the competent Ethical Committee. The Project included four epidemiological studies both *ante-operam* and *post-operam* and, as of 2018, it was not concluded.

The present thesis is focused on the *ante-operam* phase of CONSAL Study 1, which started in 2015 and was completed in the first months of 2018. Its specific aim was to investigate the health impacts of the emissions from the point sources pertaining to the refinery on the adult population living nearby, and to produce mutually adjusted estimates of the effects of environmental exposure and other additional information collected through a survey.

The study was designed as a case-control. Cases were defined as the subjects admitted to hospital between 2002 and 2014 due to acute conditions of respiratory, cardiovascular or gastrointestinal systems (ICD-IX-CM, Chapters 7-8-9 and codes 785-786); controls were selected among the subjects that were not hospitalised in the same timespan. Cases and controls, selected with a ratio of 1:3, had to be alive at time of enrolment, aged 20-64 years in the reference timespan, and were balanced for age, gender and municipality. Data were extracted from the databases of the local Health Protection Agency (Registry of insured citizens and Hospital Discharge Records); personal information were also checked with the Municipal Registries. After estimating the minimum sample size, 1046 subjects, of which 257 cases and 789 controls, were enrolled; all these subjects received a mailed survey. Fifteen subjects were excluded because they were reported dead or unavailable for any other reason. Respondents were 563 (54.6%), with a significant difference by municipality (49.1% in Sannazzaro, against 75.7% in Ferrera) and no substantial difference by age or gender. Moreover, 22 subjects declared to actually live elsewhere and, thus, only 541 were finally included in the analyses. Data management was made according to pre-specified procedures, and it was based on two databases, one containing personal information and respondence status and the other containing health data and survey data, linked with a pseudo-anonymous code; it also included an inspection regarding the consistency of survey data.

The fallout of the emissions from the refinery was predicted by the AERMOD model; individual environmental exposure was then assigned by linking the geocodes of home addresses to the modelled surface. Particulate matter (PM10) was chosen as a tracer, given that pairwise correlations with the ground-level air concentration of the other modelled contaminants (SO_2, NO_2) were extremely high. Exposures were then recoded in 2, 3 or 4 clusters by using K-means models. People living in the town of Ferrera came out to be relevantly less exposed to the emissions from the refinery than those living in Sannazzaro, and the 2-clustered PM10 exposure was roughly coinciding with the two municipalities.

The crude effect of the environmental exposure indicated an excess of health "risk" with the increase in PM10 level; however, none of the effect estimates was statistically significant. The Odds Ratios (ORs) were similar when Sannazzaro was contrasted with Ferrera (OR=1.60), or when 3- and 4-clustered exposures were used (with ORs between 1.40 and 1.50). Multivariate analyses, made by means of unconditional logistic regression, disclosed similar (and still non-significant) estimates while adjusting for age, gender, lifetime cigarette smoking and for being diagnosed or treated for other diseases that could be ascribed to the ICD codes used to define cases. Other variables were not included in multivariate models because they did not prove to be either informative or contributive. A comparative analysis of informativity looked at the three models – using as main exposure either municipality (as a *proxy* of 2-clustered PM10 concentrations), 3-clustered concentrations or 4-clustered concentrations – and showed that the best model was the first one.

A secondary analysis, evaluating the influence of several factors on self-perceived health, disclosed that living further from the refinery was reducing by 15% the "risk" of having a negative self-perception of health (OR=0.859), albeit the effect was not statistically significant. On the contrary, age and female gender (vs. male) were significantly worsening perceived health, and physical activity was significantly improving it.

The results described above are essentially consistent with previous epidemiological studies of populations living close to refineries and petrochemical plants and are toxicologically coherent. However, they might have been affected by various biases. Among others, the definition of cases and controls might have been influenced using databases that are not primarily conceived for epidemiological research; moreover, restricting to hospitalised cases might have excluded those with a less severe condition (even though this should have been corrected by a specific question in the survey). At the same time, restricting to those who were alive at the time of assessment might have excluded the most-severe ones because of death. In addition, the unexpectedly strong association between predicted PM10 concentrations and municipality, combined with the fact that cases and controls were originally balanced by municipality, may have determined an underestimation of the effects of environmental exposure (i.e. the ORs are biased towards a null effect). It should also be considered that those factors are measured at present times, while the outcome occurred in the past. Finally, albeit non-respondence bias and confounding bias were excluded, the response rate was lower than expected, so that study power was reduced.

Concluding, the results presented and discussed in this doctoral thesis indicate a possible excess of hospitalisation risk among people living in Sannazzaro de' Burgondi, in comparison with those from the Ferrera Erbognone. However, they might not be taken as a conclusive evidence, because a null effect cannot be excluded (perhaps because of the reduction in study power due to the lack of respondence) and because of the potential biases.

In the future, another study will be repeated *post-operam* and its results will be compared to the *ante-operam*; the results from Study 1 will also be integrated by other studies in the CONSAL Project. Moreover, it could be interesting to try different modelling strategies for exposure assessment (e.g. Lagrangian dispersion models) and to commit to the study of the effects that living near a contaminated site may exert on psychological wellbeing.

ACKNOWLEDGEMENTS

Alle mie radici e al mio cielo, dove ogni notte può sorgere il sole
A les meves arrels i al meu cel, on qualsevol nit pot sortir el sol
A mis raíces y a mi cielo, donde en cualquier noche puede salir el sol

I would like to thank my supervisor, Simona Villani, because in the three years spent as her postgrad student she gifted me with the two most generous things a supervisor could give: her trust, and the freedom to learn.

A special hug is owed to all the colleagues from the Section of Biostatistics and Clinical Epidemiology at the University of Pavia, and to all the people I met at the Barcelona Institute for Global Health. By making these times so special, you have made me grow as a professional and, even more important, as a decent human being.

My most heartful gratitude to my family, my roots, for unconditionally supporting me far more than I deserve; a particular mention to my uncle, who regularly provided me with food and chatting while I was in a writing rush and anyone else was on summer holidays. To the amazing friends that I've met, for the sparkles they bring. To you, my sky, for holding my hands.

To all these things, that made this walk so mesmerising, and to me, for walking 'til here.

THESIS

1. Introduction

Air pollution is perhaps the most relevant environmental risk factor of our era: it is considered responsible of one ninth of all the deaths occurring worldwide, with 90% of the global population breathing air that does not meet the criteria of the Air Quality Guidelines, issued by the World Health Organisation (WHO). As specifically regards outdoor pollution, it carries a burden of more than 3 million deaths per year and 85 million DALYs (Disability-Adjusted Life Years) [*WHO, 2016*]. It affects every component of the society, disregarding socioeconomic status or other relevant factors [*WHO, 2016*], even though not all those groups are affected to the same extent [*MARTUZZI ET AL., 2014*].

Moreover, air pollution can be seen also as a marker of sustainable development as it exerts other effects, albeit not as a direct consequence. For instance, it contributes to global climate change [*WHO, 2016*]: air pollution can be thus a threat for health in many different ways, given that global warming itself impacts on our health. For these (and more) reasons, keeping air pollution under control is included in the list of the United Nation's (UN) *Sustainable Development Goals*[1] (SDG indicator 11.6.2: population-weighted annual mean of PM2.5; SDG indicator 3.9.1: mortality attributable to indoor and outdoor air pollution) [*WHO, 2016*].

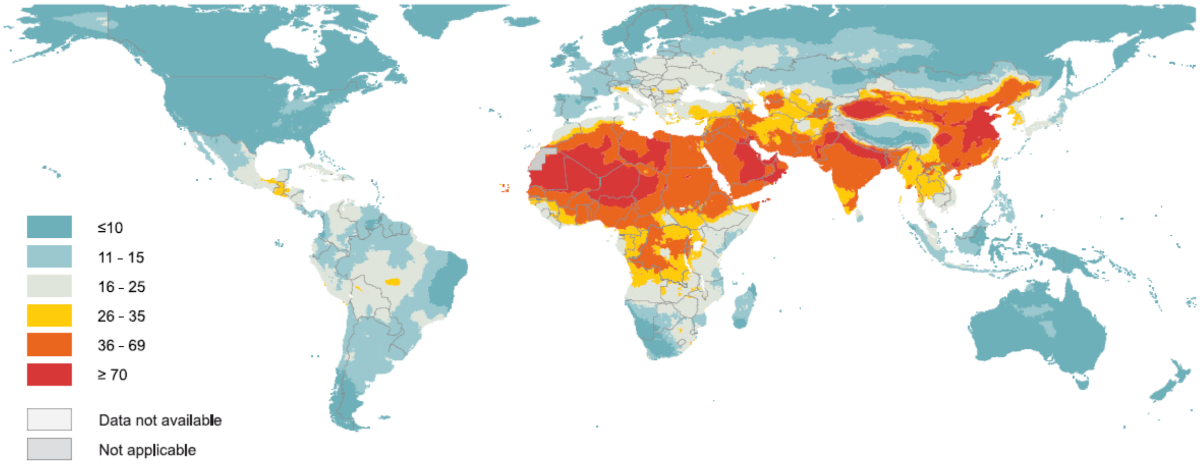

Figure 1.1 *– Modelled annual mean concentration of PM2.5 (in $\mu g/m^3$). With regards to Italy, the mean concentrations are higher in Po Valley. Reprinted from* WHO, 2016.

[1] The *2030 Agenda for Sustainable Development* was adopted by the UN in September 2015. It includes a set of 17 macro-SDGs accounting for the interaction between economic, social and environmental dimensions [*MARSILI, 2016*].

1.1. ENVIRONMENTAL HEALTH: FROM SCIENCE TO SOCIAL JUSTICE

Beyond air pollution, the dimension and the global scale of the environmental burden of disease turned it into a major concern. On this regard, it is useful to reflect about the "ethical background" of *epidemiology*. In fact, this discipline goes far beyond carrying out methodologically sound research in public health, and includes social justice in its core [*SOSKOLNE, 2016*]. An epidemiological study may be indeed the answer to any unusual exposure and/or outcome or to any need of monitoring, surveillance and improvement in the knowledge of a population, but it cannot be forgotten that epidemiological results are expected to improve the well-being of present and future generations by promoting informed policy-making and planning of interventions. Those objectives concerning public policies should be moved by fairness and beneficence – in other words, by pursuing public interest – and so should be the epidemiology behind them [*SOSKOLNE, 2016*].

The idea of environment as a determinant of health is recapped in the concept of *environmental health*, proposed by the WHO, which comprised in this definition all the physical, chemical and biological determinants that are extrinsic to a person but can have an effect on health, well-being or behaviour [*PRÜSS-USTÜN ET AL., 2016*]. The need to study how populations, environments and health interact emerged after awareness about the importance of environmental health and its impacts on mankind began to rise. The concept of environmental health stands behind the approach of the *Environmental Burden of Disease*, first elaborated in the Eighties by some pioneering researchers in the field of public health (see, for instance, the work by Doll & Peto, published in 1981). This approach was later structured by WHO with the reports about the *Global Burden of Disease*, periodically issued by the Organisation.

Environmental health is also strongly intercorrelated with the idea of *environmental justice*: the most-exposed populations for what regards environment factors, indeed, could be the most-deprived from a socio-economic standpoint. In a few words, by substantially contributing to environmental contamination, low-quality industrialisation spreads the costs on population health. This happened in the past century (for the first countries to undergo industrialisation) as well as at present times (for those where industrialisation is a recent process), and dislocation of industrial production to disadvantaged areas within or across countries is affecting communities with scarce instruments to know about industrial exposures (*awareness*) and to face it (*action*) [*Marsili, 2016*]. This fact may result in the amplification of the impacts of environmental stressors, because their effect would be combined with other health determinants, including the capacity of accessing prevention and socio-sanitary assistance [*WHO, 2013*]. To sum up, inequities in health are unfair, unnecessary and preventable, as well as exposure to unhealthy living and working conditions (especially when they result in health inequalities) [*Marsili, 2016*]. In this context, epidemiology is not only a discipline of health research (hopefully) carried out in methodologically robust ways, but it has also a lot to do with social justice issues [*Soskolne, 2016*].

Programmes by the WHO and by United States' federal agencies also promoted the framework of *Global Environmental Health* (GEH), defined as "*research, education, training, and research translation directed at health problems that are related to environmental exposures and transcend national boundaries, with a goal of improving health for all people by reducing the environmental exposures that lead to avoidable disease, disabilities and deaths*" [quoted in *Marsili, 2016*]. GEH broke national and sectorial boundaries by elevating environmental health to a global dimension for two main reasons: worldwide contamination from different sources determines health impacts on the global population as a whole, and industrial policies and practices (which effects often transcend national borders) are a critical factor for socio-economic development. GEH is made up of three mutually interacting blocks: scientific research, that can help improving our understanding of environmental health in all its biological and social components; international scientific cooperation; promotion of public and environmental health literacy, which is a support to affected communities aimed at making them more aware and more prepared, ultimately fostering informed choices [*Marsili, 2016*].

The vision of environmental risk reduction as a component of "social justice" stands at the basis of the *Ostrava Declaration on Environmental Health* [*WHO, 2017*], in which the signatories (the European Union among them) committed to "*shape future common actions to decrease the burden of diseases caused by environmental factors for current and future generations [...] achieving health and well-being objectives of the United Nations 2030 Agenda for Sustainable Development*". In the final document, concerns are raised regarding how inequalities are exacerbated by environmental degradation and pollution, climate change, harmful chemicals, affecting socially disadvantaged and vulnerable population groups more than others. It is emphasised also that "*every government and public authority [...] shares the common responsibility for safeguarding the global environment through intersectoral collaboration and citizens' participation*" and that everyone should be committed to "*coherent multisectoral strategies that emphasise system-wide and equitable preventive policies to improve environmental health conditions, and keep in mind the consequences for the social determinants of health, particularly amongst the least privileged in the Region*". Thus, it is pinpointed that actions should be undertaken with the aim of protecting and promoting health and well-being of all citizens, preventing premature deaths and diseases and reducing inequalities related to environmental stressors, and that this requires working in partnership with different sectors and stakeholders.

1.2. BASIC VOCABULARY

Before starting to explore the implications of air pollution and how these could be assessed, it is necessary to share a clear and consistent definition of several terms of common use in the field of environmental epidemiology.

First of all, an *environmental impact* can be defined as any qualitative and/or quantitative alteration of environment, either positive or negative, resulting from direct or indirect processes in the short or in the long term, temporary or not, due to the realisation of a plan, programme, intervention or project [*ISPRA, 2016*].

Pollution is specifically referred to the introduction by human activities of substances, vibrations, heat, noise or any other agent in air, soil, or water, potentially affecting environment (i.e. acting as an environmental risk factor). In both these definitions, *environment* should not be intended only as "natural environment" but rather as the system of anthropic, natural, chemical-physical and climatic factors, landscape, architecture, culture, economy [*ISPRA, 2016*]. Air pollutants are released by different types of sources, even though combustion of fossil fuel is the most relevant. They can be classified by source, size, physical and chemical characteristics. Another important distinction is between *primary pollutants*, i.e. those directly released into the environment, and *secondary pollutants*, which are chemical transformations of the primary pollutants after those are released in the atmosphere[2] [*BERNSTEIN ET AL., 2004*; *WHO, 2006*].

An *exposure* is the identity of a stressor (or multiple stressors) with which a *receptor* (for instance, a person) has a contact; the definition includes the location where the contact took place, the timing of the contact (both time and duration), and the *dose*, i.e. the amount of the receptor's exposure. If exposure to a given stressor from multiple sources is considered, then it is said to be an *aggregate exposure* [*NAS, 2017*]. From the definition of exposure, *exposure science* can be seen as "*the collection and analysis of quantitative and qualitative information needed to understand the nature of contact between receptors (such as people or ecosystems) and physical, chemical, or biologic stressors [striving to] create a narrative that captures the spatial and*

[2] A comprehensive and detailed classification is offered in *BERNSTEIN ET AL., 2004*.

temporal dimensions of exposure events with respect to acute and long-term effects on human populations and ecosystems" [National Research Council, 2012, quoted in *NAS, 2017*]. Exposure to air pollutants principally depends on the concentration of the pollutants in the places where the receptors spend their time and on the time they spend there: the most part of exposure occurs when a subject is in an indoor environment, because it is where people tend to spend most of their time. It is also important to say that indoor contamination may be the result of outdoor pollutants penetrating indoor, as well as of processes directly occurring indoor; often, outdoor measures are used as a *proxy* of indoor exposure, albeit the relation between indoor and outdoor contamination may not be so straightforward [*WHO, 2006*].

Exposure metrics can be classified in two main groups: *internal exposures* and *external exposures*. The former class includes all the measures of the type and amount of stressors directly at the receptor (e.g. biomonitoring with biomarkers), while the exposure levels in any matrix or media outside the receptor (e.g. air concentration of a pollutant) pertain to the latter class [*NAS, 2017*].

1.3. Exposure assessment for environment and health

An appropriate assessment of environmental contaminants is a crucial step towards the detection of its health effects, especially in studies involving residentially exposed populations. On the other hand, defining the exposure(s) is one of the toughest parts in those studies approaching the health impact of environmental pollution, and misclassifying the exposures could heavily undermine any finding because of an *information bias* [*Martuzzi et al., 2014*]. In any case, the choice of a methodology for exposure assessment cannot ignore the question to be answered [*WHO, 2006*].

1.3.1. Main issues of exposure assessment for epidemiological purposes

To face the task of exposure assessment, the first step is to develop a conceptual exposure model for the site under investigation, accounting for the complex pattern of contamination processes over space and time and combined presence of multiple sources (either point-emissions or not) [*WHO, 2013*]. The pattern of exposure is often made up of multiple pathways – including inhalation, ingestion, contact with contaminated soil or water, bioaccumulation in the food chain – and all these concurrent events should be carefully evaluated [*ISPRA, 2016*]. Moreover, similarly to what happens with the assessment of health outcomes, the time dimension is a major issue and a potential weakness of environmental epidemiological studies: exposure levels and exposure patterns may vary widely over time, and even the exposed population might not be stable in the timespan under investigation.

A comment regarding the assessment of long-term exposure to air pollution is worth, because of some specific issues that require careful consideration. Assessing exposure in the long term is needed when the aim is to investigate chronic effects, but such estimates of exposure should take into account both the trends in the sources of air pollution (for instance because this may change the pollutants or their relative importance) and in the lifestyles of the target population. Likely, this would not be an easy job, especially if the assessment is retrospective: past data may indeed be of different quality and completeness, and they might even have been obtained by means of different methods. Attitudes and lifestyle of the subjects in the target population towards the exposure (e.g. occupation, housing) may have changed as well [*WHO, 2006*].

1.3.2. Overview of the assessment methodologies

A possible way to assess exposure is biomonitoring, i.e. (quantitatively) assessing the presence of contaminants or their metabolites directly in human biological matrices[3]. Biomonitoring, when applied directly to humans, has the advantage of being a "real life assessment": the actual individual exposure due to any source in a certain moment is measured. It is particularly useful with well-known, specific contaminants. Another form of real-life monitoring is the assessment of environmental contamination (i.e. real-world diffusion models, indirectly indicating human uptake) [*WHO, 2013*].

It is clear that real-life monitoring might be preferable from several points of views, because it does not require any assumption but the "identity" of the contaminants that are measured. However, it is not always possible to perform such assessments due to cost issues or practical feasibility (e.g. populations of great size, retrospective studies with no available historical series of environmental monitoring) [*WHO, 2013*], and also real-life environmental measurements might not be possible due to the lack of monitoring systems in the area of interest [*WHO, 2016*] or because the costs of a measurement campaign cannot be sustained. In all those cases, *in silico* modelling might be a good alternative: despite several limitations, it is relatively cheap and can make a combined use of data from different sources. An introduction to dispersion modelling is presented in ***Appendix A***. If a modelling approach is adopted, ICSs are often considered as point sources – disregarding their possible space complexity – and only airborne emissions are modelled, even though contamination of other environmental matrixes might play a relevant role for human exposure [*WHO, 2013*].

In addition, the evolution of technology has recently come to help epidemiological sciences by providing new ways to replace or integrate the methods used until these days. Personal sensors, remote sensing, new computational tools, non-targeted analysis (i.e. performing of broad surveys allowing to assess the presence of any chemical in any matrix of interest) are now used, and the *Omics* approach is opening the new era of *exposomics*[4]; the development of exposure sciences also allowed a more

[3] The definition of biomonitoring could be applied also to the assessment performed on biological matrices from animals or other living organisms, or the use of animals as bioindicators [*WHO, 2013*].

[4] See, for instance, the *HELIX* Project and the *EXPOSOMICS* Project. The term "exposomics" was proposed by Wild in 2005 [*NAS, 2017*].

accurate site-specific assessment. Similarly, thanks to the advances in molecular techniques, bioinformatics, sensors technology, computational tools and analytical methods, innovation reached also the field of toxicology that stands behind causality assessment of the associations between exposure and outcomes [*NAS, 2017*].

1.4. INDUSTRIALLY CONTAMINATED SITES

An alteration of environment quality (including air quality) can be due to the presence of an *Industrially Contaminated Site* (ICS): one of its possible (yet not univocal) definitions is "[an area] *hosting or having hosted human activities which have produced or might produce environmental contamination of soil, surface or groundwater, air, food-chain, resulting or being able to result in human health impacts*" [*WHO, 2013*; also quoted in *PASETTO ET AL., 2016*]. Particularly in Europe, this problem is relevant because of an earlier industrialisation process and a delayed awareness of the environmental impact of industries, resulting in a complex environmental legacy of those sites [*WHO, 2013*], but the same actually happened in the most recently industrialised countries. Moreover, in Europe as in the rest of the world, ICSs are often located close to urban areas and socio-economically deprived neighbourhoods [*MARSILI, 2016*], thus increasing the strength of the potential impacts by synergic or supra-addictive interactions with other health determinants [*MARTUZZI ET AL., 2014*; *WHO, 2013*] and raising issues concerning environmental justice.

1.4.1. Industrial contamination and environmental health

ICSs are deemed to considerably affect public health. Being a threaten to environment, it is straightforward that ICSs can become threatens for the health of people living in the contaminated nearby environment [*PASETTO ET AL., 2016*] through several different ways and different exposure patterns [*WHO, 2013*]. Going deeper in this topic, the contamination of various environmental matrixes (i.e. air, superficial water, groundwater, soil) can reach high levels and can be due to different pollutants, sometimes coming from different sources; moreover, human exposure can be the result of several inter-related exposure pathways (e.g. occupational, residential or other environmental exposures) [*MARTUZZI ET AL., 2014*]. When an ICS is located in a mainly rural area, it can also affect farming land[5] and thus the basis of the food production chain [*MANCINI ET AL., 2016*], opening an additional exposure pathway (see ***1.6.2 Brought to the lunch table: pollutants and food chain***).

[5] A detailed analysis of this issue, including a *Health Impact Assessment* perspective, can be found in the work by *VANNI ET AL.* [*2016*].

In addition, all these exposure pathways strictly due to "environment" are likely to be characterised by complex interactions with other risk factors of different nature (socio-economic factors, lifestyle and habits, and so on), even though it is not yet clear if and how they interact [*Martuzzi et al., 2014*]. Unfortunately, despite evidence regarding interactions is lacking, their possible existence alone might be enough to undermine any step towards causality [*Pasetto et al., 2016*].

Beyond effects on physical health, exposure to contamination among people living near an ICS can affect the psychological and social components of wellbeing as well [*Granieri, 2015*]. The awareness of living in a potentially harmful environment can indeed foster internal conflicts, feeling of helplessness and hopelessness, social withdrawal, negative psychiatric conditions like depression or anxiety, and even impact on the sense of belonging to a place and a community [*Granieri, 2015*].

1.4.2. Assessing the impact of ICSs

The choice of the best methodology for the investigation of a certain hazard scenario related to an ICS, besides of the scenario itself, also depends on the aims of the study (including the needs of decision-makers and stakeholders, if the study is supposed to provide policymakers with data to help them in developing *evidence-based policies*) [*WHO, 2013*]. The available frameworks for investigation include both modelling paradigms and proper epidemiological approaches, as discussed below. A promising tool, based on integrated measures of population health, has been proposed with the approach of the *Environmental Burden of Disease* [*Pasetto et al., 2016*].

1.4.2.1. Assessment through modelling frameworks

In general, modelling approaches can be defined as the set of procedures aimed at "*characterising the nature and magnitude of health risks to human beings*" due to environmental stressors those humans are exposed to [*WHO, 2013*]. It is a very valuable and convenient tool to study the impacts of an ICS on public health, and in some cases the modelling way can be even more appropriate than a proper epidemiological approach [*Pasetto et al., 2016*; *Briggs, 2008*]. This said, models have some critical points that can substantially weaken their results: a heterogeneous conjunct of hazards might be present at a time, and exposure data might not be reliable; moreover, aetiological pathways might be extremely complex, involving also other determinants of health (e.g. socio-economical and occupational factors) [*WHO, 2013*].

Broadly speaking, two main frameworks can be identified: *Health Risk Assessment* (HRA), focused on hazardous exposures, and *Health Impact Assessment* (HIA), regarding more strictly the balance between positive and negative effects of exposures or, in other words, interventions (policies, authorisations to production activities, *et similia*) rather than agents [*Pasetto et al., 2016*; *Briggs, 2008*].

The HRA framework is essentially a toxicological approach, which combines toxicological evidence and exposure levels to produce theoretical estimates of the potential health *risks* due to a series of stressors, often leading to the identification of those exposures for which the tolerable dose could be exceeded. The process essentially consists of four steps: hazard identification, dose-response assessment, exposure assessment, and risk characterisation [*ISPRA, 2016*]. The HRA methodology requires to identify the sources of contamination and the pathways from the contaminant to a receptor under certain circumstances, this being for example a worst-case scenario or an average scenario. It can account for all the key exposure pathways and for different contaminants but, actually, each one of them is considered individually (one at a time) [*WHO, 2013*]. Under the Italian law, HRA should be part of the *Environmental Impact Assessment* or VIA (*Valutazione di Impatto Ambientale*)[6], even though public health aspects are lacking in a substantial share of the submitted VIAs [*ISPRA, 2016*].

The HIA (under the Italian law: VIS, *Valutazione di Impatto della Salute*), according to one of its most-accredited definitions (proposed in 1999 by the WHO's European Centre for Health Policy), is "*a combination of procedures, methods and tools by which a policy, programme or project may be judged as to its potential effects on the health of a population, and the distribution of those effects within the population*"[7]. It is essentially based on the same four steps of HRA, but the outcome of the procedure is different. Generally speaking, HIA methodologies are grounded on modelling the causal pathways as a web. They allow to gather together epidemiological evidence (or, seldomly, toxicological evidence) from different sources and regarding different risk factors and health outcomes, and then to make use of dose-response functions to

[6] Depending on the type of project for which the authorisation is required, the applicant might be required to comply with different procedures.

[7] See *http://www.who.int/hia/about/defin/en/* (opened on August 25th, 2018).

compute the excess of health risks due to the exposure(s) of interest [*ISPRA, 2016*]. This is done by means of health impact statistics measuring mortality and morbidity, like attributable cases or DALYs: in other words, HIA is a *burden-of-disease* approach (i.e. it leads to impact estimates, rather than the risk estimates obtained from a HRA) and, for this reason, it can be particularly useful when the research question regards the comparison of different scenarios (*Comparative Risk Assessment*). A flowchart of the HIA process is reported in ***Figure 1.2***.

An important difference between HRA and HIA is that the former only accounts for the effects of the source of interest, regardless of any pre-existing condition, while the latter takes into account the *ante-operam* conditions (e.g. background concentration of pollutants) [*ISPRA, 2016*].

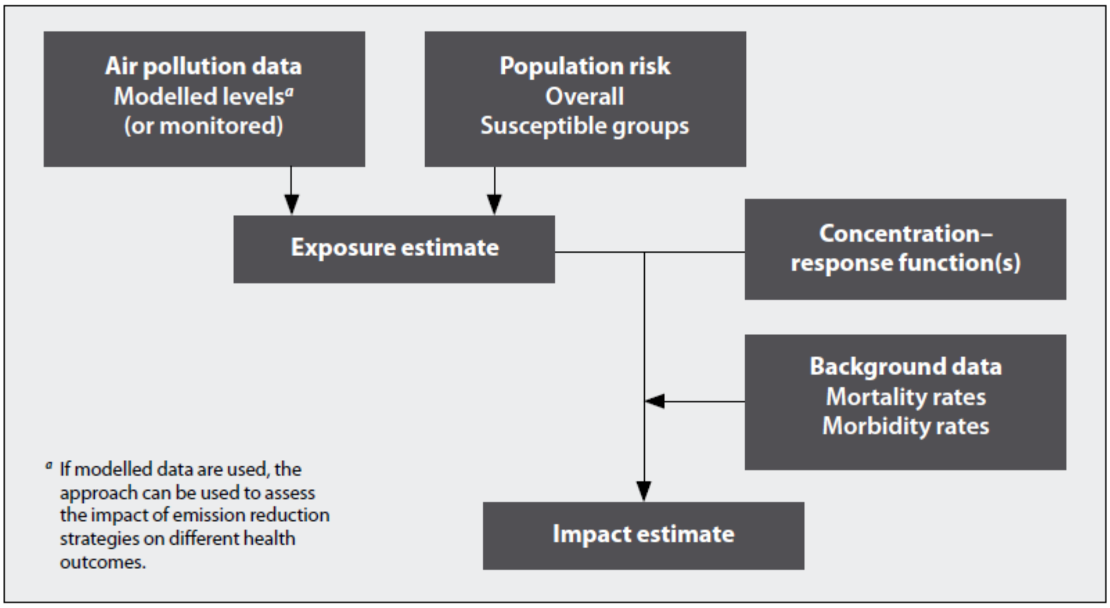
Air pollution data
Modelled levels[a]
(or monitored)
Population risk
Overall
Susceptible groups
Exposure estimate
Concentration–response function(s)
Background data
Mortality rates
Morbidity rates
Impact estimate
[a] If modelled data are used, the approach can be used to assess the impact of emission reduction strategies on different health outcomes.

Figure 1.2 – *General outline of the HIA process. Reprinted from* WHO, 2006.

More complete modelling paradigms have been elaborated by proposing the integration of environmental and health aspects. At the beginning of their history, these integrated frameworks were developed to study complex issues like atmospheric acidification or climate change, which clearly require any approach not to be limited in terms of space or scope [*Briggs, 2008*]. These integrated approaches can be defined as "*means of assessing health-related problems deriving from the environment, and* *health-related impacts of policies and other interventions that affect the environment, in ways that take account of the complexities, interdependencies and uncertainties of the real world*" [*BRIGGS, 2008*][8] and are "*multidisciplinary instrument that draws from many different disciplines, like public health, social and political sciences, environmental science, urban planning, epidemiology and statistics*" [*PASETTO ET AL., 2016*]. This kind of integrated approaches, according to *BRIGGS* [*2008*], recognise "*the systemic nature of risks to human health*" which may benefit from a transparent relationship with the different authorities involved in policy-making processes, and might also help those authorities working together. A graphical outline of this framework is represented in ***Figure 1.3***.

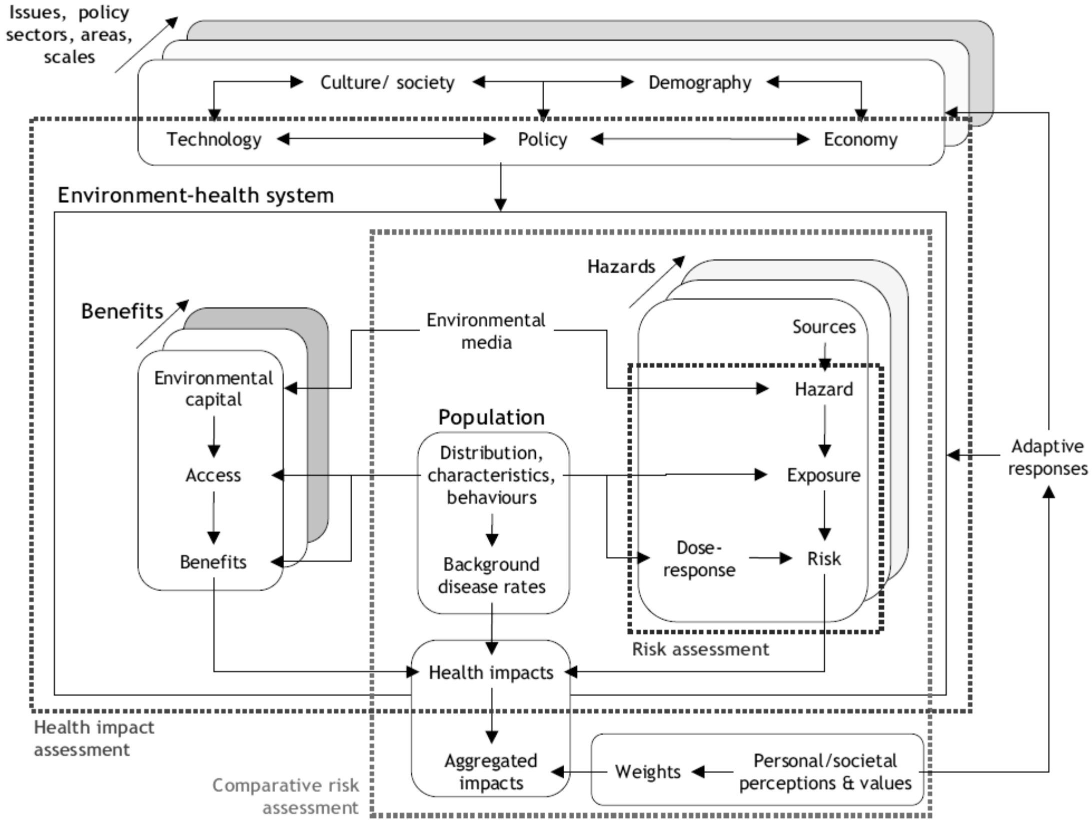

Figure 1.3 *– General outline of the integrated approaches to the investigation of environmental risks in public health. Diagram from* BRIGGS, 2008.

[8] Another nice definition, given by Rotmans and van Asselt, is that integrated approaches are "*an interdisciplinary and participatory process of combining, interpreting and communicating knowledge from diverse scientific disciplines to allow a better understanding of complex phenomena*" [quoted in *BRIGGS, 2008*].

1.4.2.2. Integrated approaches in the Italian experience: VIIAS

In Italy, obtaining the authorisations for the commissioning of an industrial plant (or to realise any other project of main relevance), either for a new facility or for substantial modification of an existing one, is conditional to the compliance of specific authorisation procedures that generally require an assessment of environmental and health impacts. According to the Italian legislation, the adoption of an integrated approach – called *Integrated Environmental and Health Impact Assessment* or VIIAS (standing for *Valutazione Integrata dell'Impatto Ambientale e Sanitario*) – is mandatory for several kinds of industrial activities (but only in the case they are considered as "strategic"), including oil refineries[9]. The VIIAS is a multidisciplinary tool that gathers various aspects pertaining different fields, from public health, epidemiology and medical statistics to environmental sciences, social sciences and urban planning [*Pasetto et al., 2016*]; it is a valuable instrument for risk management and particularly for the policymakers and the officers who are responsible for those interventions [*ISPRA, 2016*]. Being an integrated view of health and environment, it is considered as a practical application of the intrinsic features of sustainable development [*ISPRA, 2016*]. A detailed review of legislation regarding authorisation procedures requiring VIS, VIA, VAS and AIA is presented in a report by the *Italian Environmental Protection Agency* (ISPRA), together with a summary of the data sources available in Italy and of the main reference parameters [*ISPRA, 2016*]. It is interesting to notice that, in all the procedures mentioned above, the authorisation body can mandate *intra-operam* monitoring and surveillance to confirm that real risks or impacts are those that were predicted, an may even ask to perform further investigations – including epidemiological studies – also if those provisions cannot be fulfilled within the procedure. Surveillance and further investigations should go beyond a mere measurement of the indicators accounted for in the *ante-operam* phase, because the project under evaluation might have unpredicted consequences [*ISPRA, 2016*]. These provisions can be included in the authorisation decree, and the applicant must demonstrate how those are accomplished.

[9] In 2014, the Regional Government of Lombardy approved an act (D.G.R. 1266/2014) that required applicants submitting a VIA to include a VIS as well; at the national level, this was required by the D.L. 133/2014, which integrated the previous D.Lgs. 152/2006 [*ISPRA, 2016*]. Other relevant acts are L. 231/2012 and L. 164/2014.

1.4.2.3. Limitations of modelling

Whether the assessment is done by using an integrated approach or not, the modelling way will carry several potentially cumbersome caveats, tied to the already mentioned systemic nature of health risks. A first challenge is in the definition of the area affected by the source undergoing the assessment: this area should be defined regardless of anything but the source of interest, because the emissions of this source will add to the existing background. Particularly for emissions in the air matrix, it is generally easy to determine a gradient of contamination from the source, but it may not be as simple to define how far contamination due to the source could be assumed to be zero [*ISPRA, 2016*]. Nonetheless, this choice is of utmost importance because of its strong implications on the results of the assessment: several parameters are strictly related to the spatial definition of the area of interest, like population size (which in turn influence the number of attributable cases, DALYs or any other measure of impact), background incidence of the health outcomes considered in the assessment (which depends on the population), and even the types of receptors [*ISPRA, 2016*].

A second point is that a HIA requires previous quantitative epidemiological evidence about the risk associated to a certain exposure, just as HRA requires previous toxicological research usually carried out in animal models [*ISPRA, 2016*]. The lack of robust evidence would obviously undermine any finding. Moreover, even when such evidences are available, the coexistence of multiple causes may entail non-linear interactions in terms of health risks, making the system's behaviour harder to predict. In other words, the issue regards the very definition of "cause" in a complex system of interacting factors that change in space and time [*Briggs, 2008*; *WHO, 2006*]. Also, epidemiological studies can demonstrate associations of a certain strength, but it is far more difficult to demonstrate causality (even in the probabilistic meaning of *cause*). Yet, causality is implicitly assumed in HIA [*WHO, 2006*].

1.4.2.4. Assessment through the epidemiological approach

Epidemiological studies may be effective in getting rid of part of the limitations of the modelling approaches discussed above, for instance by controlling for confounders either while designing the study or in statistical analyses [*Briggs, 2008*]. In other words, the epidemiological way is important as far as it is able to depict what happens in the real world. Nonetheless, the characteristics of ICSs depicted at the beginning of the paragraph make epidemiological studies a demanding task, requiring the integration of geographical, geostatistical and analytical standpoints.

The epidemiological approach, which is for sure needed when no strong evidence has been reported previously, may be applied as a form of *post-operam* surveillance which follows the *a priori* integrated evaluation done by means of modelling (VIIAS in Italy). This surveillance is aimed at assessing if what happens is what was forecast through modelling. In order to do that, it must take into account a wider range of possible outcomes rather than only those highlighted by the models. In any case, health outcomes should be preferably chosen *a priori* (according to previous literature), on the basis of the specific kind of industrial site under investigation; such practice is required by the need of limiting data dredging, as proposed in the framework of the SENTIERI Project[10] [*WHO, 2013*].

Epidemiological studies of populations living in the surroundings of an ICS can be grouped in three main categories [*Pasetto et al., 2016*; *WHO, 2013*]:

1. *descriptive studies*, which are aimed at depicting a health profile of the resident population, highlighting associations between local risk factors and health outcomes (on an ecological basis, with all the implied limitations)[11];
2. *analytical/aetiological studies* aiming to investigate *a priori* hypotheses regarding those associations and (potentially) their causal nature. In this case, time (and in particular latency from exposure to outcome) is a relevant factor that must be taken into account;
3. *health surveillance studies*, which purpose is to follow a population during time to see how its health profile is changing, for instance to evaluate the effectiveness of an intervention or the evolution of risk patterns.

Being aware about the efforts required by an epidemiological approach, the question of what the aforementioned studies can "add" for evidence-based policy-making is more than legitimate. Studies in the first group are useful tools for *diagnostic assessment*, which is meant to identify the presence and, in case, the magnitude and causes of a problem, also suggesting its causes; within the third group, we may include

[10] The SENTIERI Project will be discussed in ***1.4.2.5 Examples of the epidemiological approach***.
[11] Again, an example for this group is the SENTIERI Project.

prognostic and *summative assessment*, the former intended to back up the choice among different possible new policies, the latter to evaluate the effectiveness of interventions after they are put in place [*Briggs, 2008*].

In the choice of what kind of epidemiological study a public health specialist should better develop to respond to the specific needs of the situation he or she is called to unravel, it is necessary to consider that the environmental and social contexts are often very complex. For instance, multiple sources of contamination and multiple contaminants might be present at the same time, potentially interacting one with another; the pathways leading humans to be exposed to those contaminants might be different, occurring in different contexts, and both contamination and human exposure might vary across time. In addition, the background might be different because of socio-economic factors, size of the target population and relative importance of the exposures, worries of the population at risk, media coverage, and also with regard to the principles of environmental justice [*WHO, 2013*].

Then, the choice of the type of approach strongly depends on the latency between the time when a subject is exposed and the moment when the outcome of interest becomes manifest. Effects occurring in the short term might be disclosed by time-series, case-crossover or panel studies, while long-term effects can be brought to light by ecological studies, cross-sectional studies (possibly with biomonitoring for exposure assessment), cohort studies or case-control studies [*WHO, 2013*].

Summing up, any choice regarding how to design such studies should not "forget" to take into account all the relevant specificities of a site (patterns and types of contamination, residential and occupational exposures, population and socioeconomic features, needs of the stakeholders, concerns among people...) and to decide carefully which timespan is appropriate for the sake of theinvestigation. The decision regarding the timespan should take into account the situation-specific complexity, purposes and needs [*WHO, 2013*].

To conclude, it has been suggested that a so-called *funnel approach* could be appropriate and more informative, as it moves "zooming" from an ecological to an individual perspective [*WHO, 2013*].

1.4.2.5. Examples of the epidemiological approach

In Italy, several projects based on epidemiological investigation have been implemented to study the effects of industrial contamination; in particular, SENTIERI (*National Epidemiological Study of Territories and Settlements Exposed to Risks from Pollution*, in Italian *Studio Epidemiologico Nazionale dei Territori e degli Insediamenti Esposti a Rischio da Inquinamento*) and MONITER (*Monitoring of Incinerators in the Territory of Emilia-Romagna*, in Italian *MONitoraggio Inceneritori nel Territorio dell'Emilia-Romagna*) have been particularly relevant because they proposed different integrated approaches, rising from different starting points.

The SENTIERI Project is a nation-wide Italian project of epidemiological surveillance, focused in particular on the *National Priority Contaminated Sites* (SIN, *Siti di Interesse Nazionale*[12]), that was concluded in 2010. Its general aim was to study mortality in the 5.5-million-people population living close to SINs[13] for events occurred from 1995 to 2002 with various causes (63 selected "general" causes, and other causes more specifically related to the type of exposures); mortality was chosen as the indicator because mortality data are systematically collected and elaborated (thus being available and consistent) across the entire Country, and those data were analysed by applying the same methodology to all the areas of interest [*SENTIERI, 2010*; *PIRATSU ET AL., 2013*]. The study concluded that populations exposed to a SIN are characterised by a less favourable health status than the general population [*COMBA ET AL., 2016*; *PIRATSU ET AL., 2013*]. The SENTIERI Project is considered as a "first-level approach", because it essentially describes the health profile of exposed populations basing on information that are routinely collected for other purposes – i.e. the *Registry*

[12] A SIN is a site where hazardous contaminants are present and are severely impacting (or can potentially impact) particular environments and/or health of people living nearby. SINs were established by the Legislative Decree 22/1997 and modified or integrated several times since then (e.g. D.Lgs. 97/2002, D.Lgs. 152/2006); the act 134/2012, in particular, modified the criteria that define a SIN, thus reducing the number of inventoried SINs from 57 to 40, as of June 2016. The Regional governments took charge of monitoring and remediation for these sites, were provided with extra funds for that, and were required to comply with specific prescriptions under the control of the Ministry of Environment. See, for instance, the following link: *www.isprambiente.gov.it/it/temi/siti-contaminati/siti-di-interesse-nazionale-sin* (opened on August 25th, 2018).

[13] Actually, the SENTIERI Project only included 44 SINs. The remaining ones were excluded because they were of scarce public health interest (due to the irrelevance of human exposure), or the exposed population was too far from the geographical units (municipalities) for which mortality data were available.

of Causes of Mortality (ReNCaM, *Registro Nominativo delle Cause di Morte*) – and therefore it is able to suggest possible associations with environmental stressors, requiring low economical and time efforts [*WHO, 2013*]. The ultimate aim is to provide policymakers with evidence for the identification of priorities in terms of remediation and preventive interventions [*SENTIERI, 2010*]. Beyond this, the SENTIERI working group made a tough reviewing work in order to evaluate how robust and strong the associations between exposures and outcomes were in the scientific literature.

Another important epidemiological project in Italy has been the MONITER Project [*MONITER, 2012*], developed by the Region of Emilia-Romagna to acquire evidence regarding air quality issues near solid waste incinerators (the Region, at the time of the project, had 7 such plants) and related health impacts on the population as well as among the workers, in the background of the Po Valley with its well-known problems of high pollution levels[14] [*ROSSI ET AL., 2012*]. The project, which was articulated in seven work packages[15], accompanied scientific research with participatory processes and target-oriented communication tools, and it has been defined as an example of Health Impact Assessment[16].

It is worth mentioning also the so-called *Public Health Assessment* (PHA), developed by the US Agency for Toxic Substances and Disease Registries, in which HRA (quantitative risk assessment) and a descriptive epidemiological approach (health outcomes data) are combined. This framework has been applied to the study of all the priority sites identified by the Environmental Protection Agency of the United States of America (US EPA), and it was based on the same environmental data used by that agency. However, the assessment was carried out site-by-site (thus being more capable of capturing site-specific conditions) and with the aim to assess if whether the population was exposed to excessive hazards that had to be controlled by reducing or even eradicating the exposure [*PASETTO ET AL., 2016*].

[14] The issue is discussed in ***1.5.1 Environmental background***.

[15] See *https://www.arpae.it/cms3/documenti/moniter/descrizione_sintetica_progetto.pdf* (opened on August 30th, 2018), where the MONITER project is outlined (in Italian).

[16] See the abstract of Lanzalone, N., & Siciliano, T. (2011). Progetto Moniter: un modello di VIS per gli impianti di incenerimento. *Epidemiologia & Prevenzione, 35*(2):136-138.

1.5. AN OIL REFINERY IN LOMELLINA

In the district of Lomellina (North-West Province of Pavia, Region of Lombardy, Italy), an oil refinery is operating since 1963, being one of the most important refineries in Northern Italy as regards production[17]. During its life, new facilities have been built, so the production rate has raised at least twofold and the production has been differentiated [18]. More specifically, this plant is located amid the rural areas surrounding the settlements of Sannazzaro de' Burgondi and Ferrera Erbognone, on lands pertaining to those municipalities. As of 2018, the plant is owned and run by the private company ENI S.p.A. and still represents a major player for the socio-economical and occupational standards of the area.

1.5.1. Environmental background

The district of Lomellina geographically belongs to the Po Valley, one of the most extended river flood plains in the European Union (EU). The Po Valley is a highly urbanised area, with more than 20 million inhabitants, where a great concentration of human activities (like industries, factories, trading centres) is hosted. At the same time, it is also an area of intensive farming thanks to a very fertile land[19]. The climate is typically continental, with rainy days mainly during spring and fall, a high thermal excursion and high humidity levels [*ARPA-L, 2015*].

From a geographical standpoint, the Po Valley is closed on 3 sides (North and West by the Alps, South by the Apennines) and winds are generally weak and directed from North-West to South-East; thus, air renewal is limited, naturally facilitating the persistence of contaminants in the air. For this reason, if the Italian territories in general are critical concerning air quality conditions, the Po Valley in particular is one of the most polluted areas in Italy and in the entire EU [*EEA, 2016*]. In fact, observed concentrations of PM10 recorded in 2014 across the EU are represented in ***Figure 1.4***, where it can be clearly seen how air quality in the area is unfavourable.

[17] Data from *Unione Petrolifera*, referred to 2007 [*ENI, 2008*].

[18] It is interesting to notice that, as of 2018, the realisation of a regional asbestos dump in Ferrera Erbognone is planned.

[19] Specifically, the Lomellina district has a mostly rural vocation.

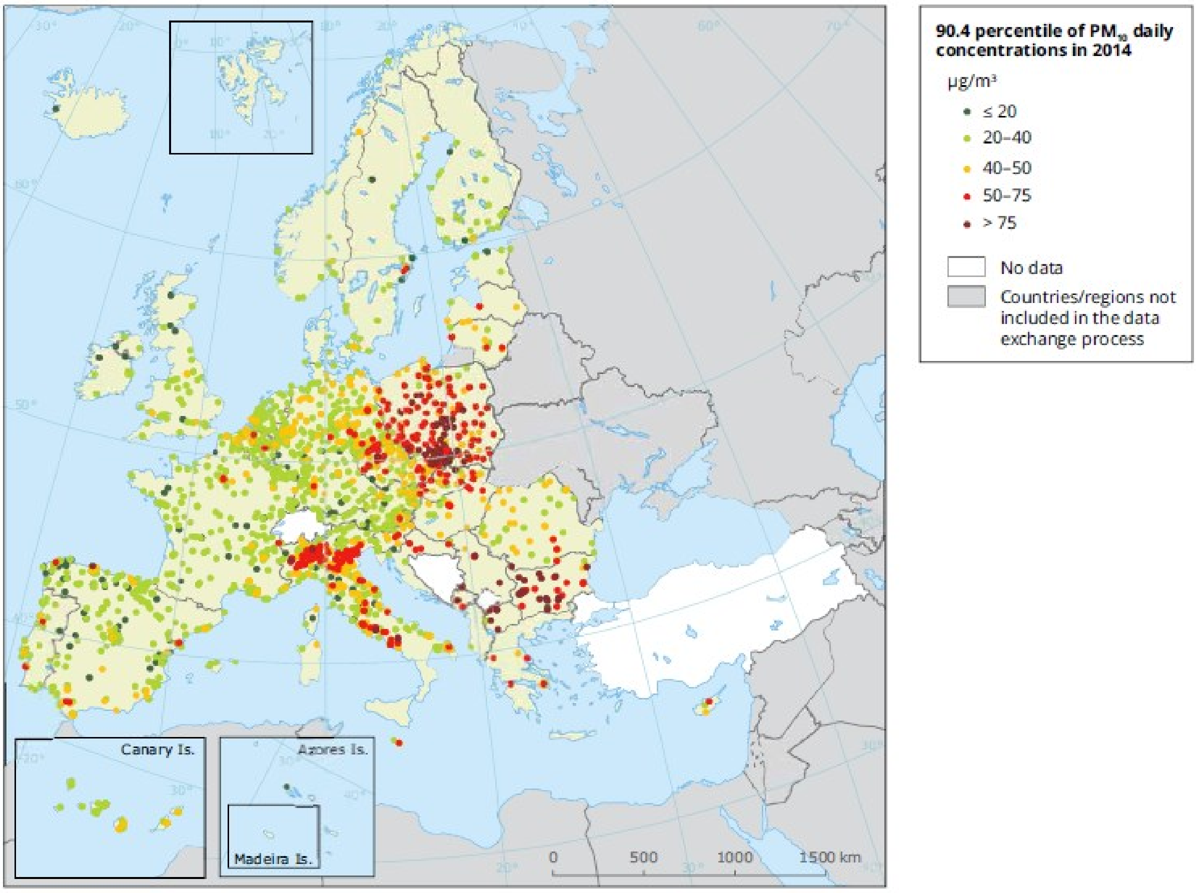
90.4 percentile of PM$_{10}$ daily concentrations in 2014
µg/m^3
● ≤ 20
● 20–40
● 40–50
● 50–75
● > 75
□ No data
▨ Countries/regions not included in the data exchange process
Canary Is.
Azores Is.
Madeira Is.
0
500
1000
1500 km

Figure 1.4 *– Daily concentrations of PM10 observed in 2014. Each dot represents the position of a monitoring station registered in the European network; a dot's colour indicates the concentration range of the 36th-highest value recorded by the monitoring station (it is worth recalling that, by law, 35 exceedances of the limit of 50 µg/m^3 are allowed). In particular, red and dark-red dots are those monitoring stations where the 36th value in the ranked series of records was above the limit of 50 µg/m^3. Reprinted from* EEA, 2016.

1.5.2. Main characteristics of the refinery

As of 2008, the plant was built on a surface of roughly 230ha and consisted of different units, among which topping distillation, vacuum distillation, catalytic cracking, visbreaking, production of hydrogen and hydrocracking, deasphalting, desulphurisation, GPL fractioning, extraction of solvents, and more. An exhaust system conveyed the waste from all the units to a flare stack (conveyed emissions, i.e. point sources), even though part of the emissions of pollutants from the refinery – particularly VOCs – was actually attributable to area sources. The refinery was authorised to transform approximately 10 million tons of crude oil per year, and the output of the refining processes covered a wide range of products, like propane and

GPL, gasoline and diesel fuel, kerosene, synthetic gas, liquid sulphur, propylene, bitumen, and others [*ENI, 2008*]. The plant also included various auxiliary facilities, like a water treatment plant and stocking areas. In addition, the synthetic gas was used by a co-generation power plant (EniPower) located in Ferrera Erbognone, next to the area of the refinery.

1.5.3. A new facility for the refinery

In the past years, the company developed a new technology – the so-called *ENI Slurry Technology* (EST), firstly experimented with a commercial demonstration 1200BPD-plant in Taranto – that allows to "grab the bottom of the barrel" and has the potential to increment the final distillate yield (in other words, incrementing the production of high-quality fuels being the amount of crude oil unchanged [*ENI, 2008*].

In order to implement the EST technology on industrial scale, the company built and commissioned a new full-scale pilot facility within the refinery of Sannazzaro. According to the AIA-VIA, that will be presented in the next section, this new part of the plant covers an area of 42 ha, next to the pre-existing part of the refinery (total area of the refinery: from 230 to 270 ha): a geographic detail of the refinery with the new facility is shown in ***Figure 1.5***. By design, its production rate should be around 23000 BPD[20]. The EST facility has three stacks (point sources) releasing mainly NO_X, CO, SO_X, VOC and PM into the air matrix (see paragraphs ***1.6.1.1-1.6.1.5***).

Actually, the EST facility was almost completely destroyed by a major fire occurred on December 1st, 2016. In any case, the emergency plan (which, among the rest, alerts people living in the nearby settlements to stay safe indoor and to keep windows closed) activated also the environmental and the public health surveillance protocols, respectively run by the Regional Environmental Protection Agency (ARPA) and the Health Protection Agency (ATS) of the Province of Pavia. Official statements released by the two authorities in the hours after the accident said that the fire did not determine particular risks for the population, and that no hospital admissions of residents in the

[20] The information reported in this paragraph, regarding the history of the refinery of Sannazzaro and the implementation of the ENI Slurry Technology, is based on materials from the company's website. A detailed *résumé* of the history of the refinery as a source of social conflict can be found in the Environmental Justice Atlas at the following link: *atlanteitaliano.cdca.it /conflitto/raffineria-eni-sannazzaro-de-burgondi* (opened on April 19th, 2017).

area were recorded as a consequence of the accident itself[21]. As of 2018, the facility is still under maintenance as a consequence of the accident.

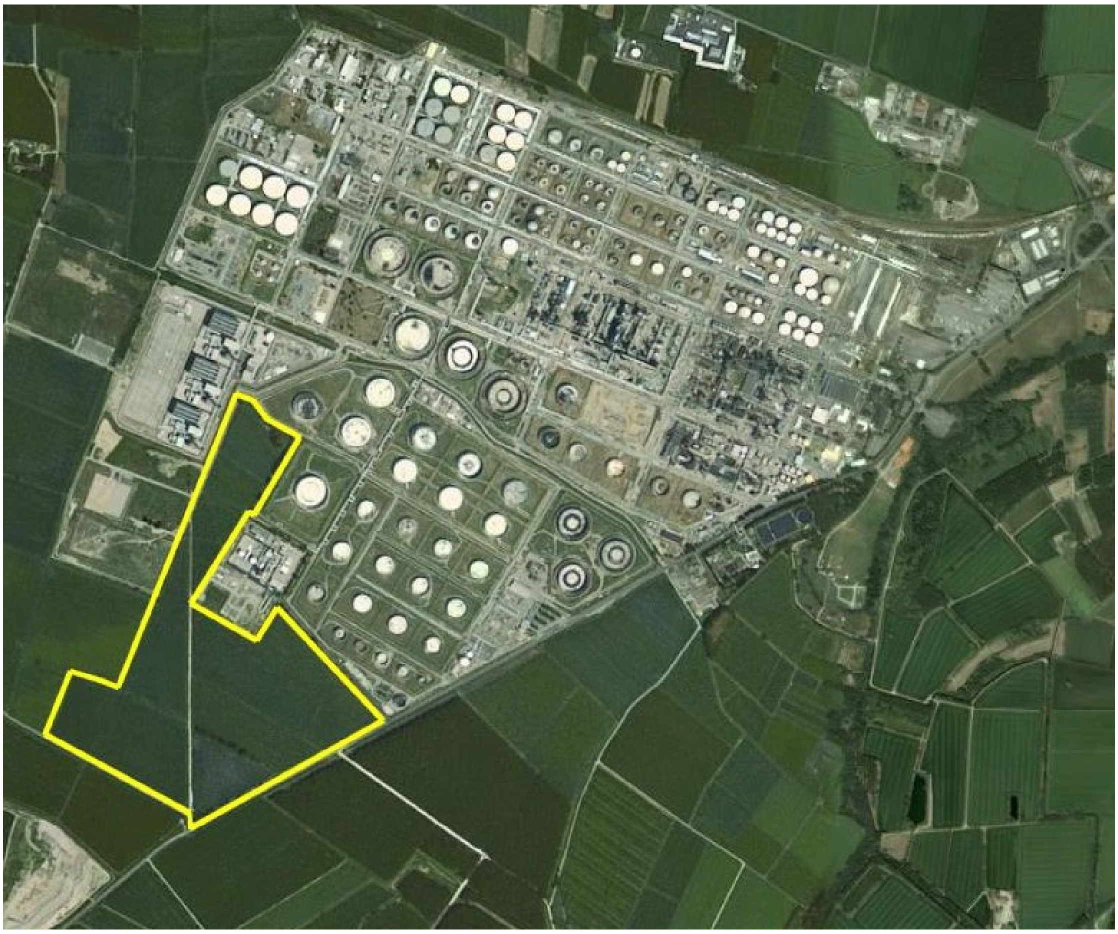

Figure 1.5 *- Map of the refinery (ante-operam). The area where the new facility was built is indicated by the yellow polygon. Reprinted from* ENI, 2008.

1.5.4. The authorisation procedure for the EST facility

According to the Italian legislation, depending on the type of industrial plant, the authorisation request for a refinery should comply with the AIA-VIA unified procedure. The AIA-VIA application for the new facility (which included also the request to increase the maximum amount of crude oil authorised for the refinery from 10 to 11.1 million tons/year) was sent by the company to the competent authorisation bodies in

[21] See *www.arpalombardia.it/Pages/new_02_12_20161141.aspx* (opened on August 27th, 2018).

December 2008[22], and was based on an integrated assessment (VIIAS). After the authorisation body entered a caveat in 2010, and after hearing the opinions expressed by the Regional government, the Ministry of Cultural Heritage, and the citizens, the authorisation decree was issued by the Minister of Environment in December 2010[23]. The pronouncement was “positive with provisions” or, in other words, the authorisation was conditional to the fulfilment of various requirements.

In particular, with the Provisions 3, 4 and 5, the Ministry set the maximum emission of pollutants allowed and, with Provisions 6 and 8, the company was compelled to carry out environmental monitoring campaigns[24]. In addition, albeit recognising that the levels of air pollution in the area cannot be entirely attributed to the refinery, in Provision 7 it was stated that the production rate had to be “adapted” in order to reduce the emissions of particulate matter by 20% in case the maximum of PM10 concentration (50 µg/m^3) was exceeded more than 35 days during a year.

Notably, Provision 21 compelled the company to start, on their own expenses and within one year from the commissioning of the EST facility, an *epidemiological investigation* of the population living in the two municipalities of Sannazzaro de' Burgondi and Ferrera Erbognone. This investigation had to be agreed upon with the local Health Protection Agency (ATS Pavia) and, as stated in the Article 2 of the authorisation decree, also with the Region of Lombardy and any other competent authority, and the protocol had to be sent to the Ministry of Environment for final approval. The authorities established that the epidemiological study had to be designed and carried out by a scientific third part, identified in the Department of Public Health, Experimental and Forensic Medicine of the University of Pavia. The project that was developed is summarised in another paragraph (see ***1.7 EPI-EST: the CONSAL Project***): a part of that project is the subject of the present thesis.

[22] Information regarding the course of the VIA procedure and pertinent documents are reported at the following link: *http://www.va.minambiente.it/it-IT/Oggetti/Info/280* (opened on August 28th, 2018).
[23] *Decreto di pronuncia di compatibilità ambientale*, no. 1014/2010 and its integration no. 592/2011. As indicated in the decree, the authorisation must be revised after 8 years from the date of release (i.e. December 31st, 2018).
[24] Similar provisions were issued also with regard to noise pollution and water pollution.

1.6. ENVIRONMENTAL RISKS DUE TO A REFINERY

1.6.1. Emissions of pollutants in the environment

According to the non-technical report submitted by the company during the AIA-VIA application [*ENI, 2008*], and referring to the *ante-operam* scenario, the main pollutants released in the atmosphere were identified in particulate matters (PM, 100 kg/h), sulphur oxides (and particularly SO_2, 680 kg/h), nitrogen oxides (NO_X, 780 kg/h) and carbonium monoxide (CO, 340 kg/g) and dioxide[25] (CO_2, 290000 kg/h)[26]. Those emissions were principally generated during the processes of power production (roughly 60%), catalytic cracking, sulphur recovery, and functioning of furnaces and flares. VOC were mainly produced by stocking and handling of the products.

In addition to atmospheric emissions, liquid waste from the refinery (either resulting from the production process or water) were released in open drains[27] at a rate of roughly 500 cubic metres per hour, after being collected by the refinery's sewer system and undergoing purification in an internal purification plant [*ENI, 2008*]. Other issues were noise pollution, light pollution (mainly due to the flare stacks), and smells.

The EST facility, again according to the *post-operam* scenario described in the AIA-VIA documentation submitted by the company [*ENI, 2008*], was predicted to generate new emissions of pollutants with regards to particulate matters (3 kg/h), carbonium monoxides (30 kg/h), sulphur oxides (60 kg/h) and nitrogen oxides (55 kg/h); a stronger increase was predicted for CO_2 (140000 kg/h); the release of sewage water would be slightly less than 70 tons/h. Anyway, because of process integration, the comprehensive emissions balance of the entire refinery would see a slight decrement for what concerns NO_X (-10 kg/h), SO_2 (-35 kg/h) and particulates (-1.5 kg/h), and an increment only for CO (+30 kg/h); also water release in open drains would actually be diminished by more than 5% (from 520 to 490 tons/h). Notably, a substantial exception would be represented by the new emissions of CO_2 which are not balanced

[25] Carbonium dioxide is mainly relevant to climate change, being a greenhouse gas, but it is of minor relevance in terms of direct health impacts [*ARPA-L, 2015*].

[26] The document reports all the emission parameters of the refinery, but their detailed analysis would be out of scope for the present thesis.

[27] Actually, a part of the liquid waste is dispersed in air because of evaporation.

by a gain in the rest of the processes, thus determining the *post-operam* emissions being increment by 46% respect to the *ante-operam*.

1.6.1.1. Sulphur oxides (SO_X)

Sulphur oxides (SO_X), and in particular the dioxide (SO_2), are essentially produced by oxidation during the combustion of fuels containing sulphur; anthropogenic sources outweigh natural ones[28]. In the atmosphere, and especially in presence of metal catalysts (e.g. when both are absorbed in particles), sulphur dioxide can be chemically transformed to sulphuric acid[29], which in turn can be neutralised by ammonia forming ammonium salts with a lower acidic power. The salts are able to form droplets by nucleation or to condense on fine particles [*WHO, 2006*].

As of 2018, regulatory limits are defined for SO_2 both for hourly and daily concentrations, and are integrated with alert thresholds and target-limits for environmental protection, as follows[30]: hourly concentration 350 µg/m^3, to be exceeded no more than 24 times per year; daily concentration 125 µg/m^3, to be exceeded no more than 3 days per year; alert threshold 500 µg/m^3 measured in more than 3 consecutive hours; flora protection 20 µg/m^3.

According to ARPA, in Sannazzaro de' Burgondi and Ferrera Erbognone the annual means in the years 2008-2014 have always been below 10 µg/m^3, except in 2011 (when, in the monitoring station of Ferrera, the annual mean was 20 µg/m^3) [*ARPA-L, 2015*].

In humans, the only relevant exposure pathway to SO_2 or its derivates is inhalation. Thanks to its hydrophilic properties, SO_2 is rapidly absorbed, and the process is so fast that it mainly occurs in the upper respiratory tract; however, the contaminant can interact with particulate matter, thus increasing the amount of SO_2 reaching the lower respiratory tract. It is highly irritant, and most of its health impacts regard the respiratory system, but it can also determine irritation of other mucosal tissues (e.g. eyes, skin). It can slow the motion rate of ciliary cells in the airways, thus reducing

[28] Starting from the Seventies, awareness of the serious consequences of this contaminant for human health and ecosystems determined a huge change also in the anthropogenic production of sulphate pollutants. Nowadays, most of the sulphur is removed from motor fuels with refining, and in industries the compound is removed from stack gases before they are released [*WHO, 2006*; *ARPA-L, 2015*].

[29] The reaction, in any case, is slow, and for this reason SO_2 is the reference pollutant.

[30] See *http://www.arpalombardia.it/sites/QAria/_layouts/15/QAria/Inquinanti.aspx* (in Italian, opened on August 30th, 2018). The limits are consistent with the European guidelines [*EEA, 2016*].

mucus clearance. Even though most of the effects are acute and tend to solve in the short term (e.g. irritation, inflammation, bronchoconstriction, or exacerbation of pre-existing conditions like asthma), chronic exposure can damage the epithelial tissues of the respiratory tract and increase the risk of diseases like asthma or bronchitis. Epidemiological researches showed that SO_2 contamination is associated with increased mortality and morbidity [*WHO, 2006*].

1.6.1.2. Nitrogen oxides (NO_X)

Nitrogen oxides (NO_X) is the general name for two molecular species that, in atmospheric conditions, are in gaseous state: nitrogen monoxide (NO) and nitrogen dioxide (NO_2). The anthropogenic formation of nitrogen oxides[31] is basically parallel to that of sulphur oxides and occurs during fuel combustion. This produces mainly NO as primary pollutant; in turn, NO spontaneously reacts with the oxygen normally present in the atmosphere and transforms to NO_2[32], which is produced almost totally as a secondary pollutant[33]. In addition, nitrogen oxides can be formed also by means of chemical reactions occurring between atmospheric nitrogen and oxygen, in presence of heat (e.g. during a high-temperature combustion). In the atmosphere, NO_2 extensively undergoes a photochemical transformation that forms radical species and vapours of nitric acid, which can be neutralised as ammonium salts (as for SO_2). For these reasons, NO_X concentration shows daily and seasonal variability and also depends on meteorological conditions. It has also been reported that contamination levels might be higher in indoor environments than outdoor [*WHO, 2006*].

For NO_2, the limits are currently established by law (as of 2018)[30] in a hourly concentration of 200 $\mu g/m^3$ (with 20 exceedances allowed per year) and an annual mean of 40 $\mu g/m^3$. An alert threshold is set at 400 $\mu g/m^3$, measured for at least three consecutive hours; the threshold for environmental protection is 30 $\mu g/m^3$ (annual mean). As reported by ARPA, in the area of Sannazzaro de' Burgondi and Ferrera Erbognone, the annual concentrations between 2008 and 2014 varied in the range 20-40 $\mu g/m^3$ [*ARPA-L, 2015*].

[31] Actually, most of the NO_X comes from natural sources, but those contribute to a low background concentration. Peaks are due to NO_X human activities.

[32] NO_2 can be found either in monomeric or dimeric form (N_2O_4).

[33] Of the NO_X emitted during a combustion process, 5-10% consists of NO_2 [*ARPA-L, 2015*].

With regards to the health effects of exposure to NO_X in humans, it occurs only by inhalation. A large share of ambient NO_2 (estimated around 70-90%) is absorbed in the respiratory tract; the amount of absorbed gas can even raise if the subject is doing physical exercise. Whilst NO is used also for therapeutic purposes (it has a vasodilatory effect), NO_2 is highly oxidant and irritant for mucosal tissues. It has negative effects on lung metabolism and function, and it determines an inflammatory state in the airways; it can also result in chronic bronchitis, asthma symptoms and emphysema. Moreover, chronic exposures undermine the receptor's capacity to defend himself or herself from respiratory tract infections, even at low doses [*WHO, 2006*]. Epidemiological evidence also identified associations with the risk of stroke and acute myocardial infarction [*BERNSTEIN ET AL., 2004*]. Nitrogen dioxide is also involved in the formation of photochemical smog (ozone), which may enhance its health impact [*BERNSTEIN ET AL., 2004*].

1.6.1.3. Particulate matter (PM)

Particulate matter (PM) includes both primary and secondary pollutants. It is a complex mixture in which the most relevant are sulphates, nitrates, sodium chloride, carbon, minerals and polycyclic aromatic hydrocarbons (PAH). PM characteristics strongly depend on particles' composition and on the chemical and physical properties of their components [*WHO, 2006*]. The biggest particles are mainly produced by mechanical break-ups like soil abrasion and dust from buildings and road (e.g. brakes and tyres abrasion), or they can be the result of aggregation processes of smaller particles [*BERNSTEIN ET AL., 2004*]; smaller particles can originate from combustion processes [*BERNSTEIN ET AL., 2004*] or from condensation [*WHO, 2006*]. Condensation of atmospheric gases with a low vapour pressure, like sulphur and nitrogen compounds, is particularly relevant, and explains why a change in their concentration can also determine a change in PM concentration [*WHO, 2006*].

PM is primarily classified after its aerodynamic characteristics, and specifically by its size (see ***Figure 1.6***, next page), because both the spread of PM in the atmosphere [*WHO, 2006*] and its capability of penetrating the human body [*ARPA-L, 2015*] largely depend on this parameter. *PM10* includes particles with an aerodynamic diameter below 10 µm. *Fine PM* or *PM2.5*, as well as *ultrafine PM* (*PM0.1*), are a sub-fraction of PM10 with aerodynamic diameters respectively below 2.5 µm and 0.1 µm: roughly two thirds of their concentration are attributed to human activities. The difference between PM10 and PM2.5 is referred to as "*(thoracic) coarse mass PM*"; the concentration of fine and ultrafine particles in PM10 is estimated to be 40-90% [*WHO, 2006*].

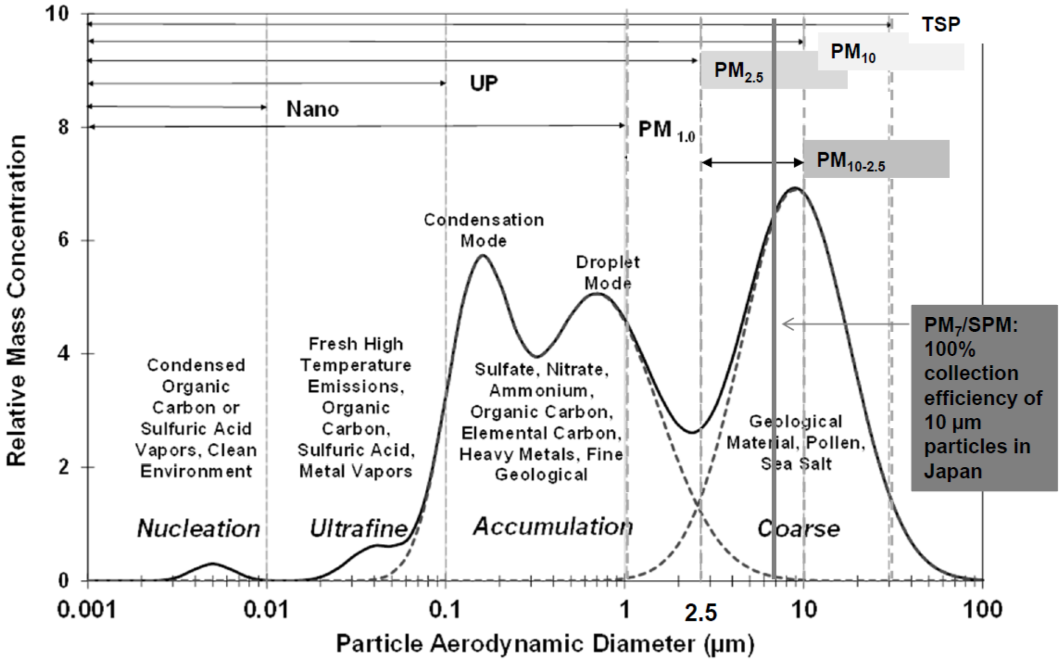

Figure 1.6 – *Concentration of particles in the total mass of PM by aerodynamic diameter, with formation processes and main components. Picture from* IARC, 2016.

As of 2018, the limits for PM10 are defined by law[30] to be 50 µg/m3 for daily concentration and 40 µg/m3 for the annual mean concentration; moreover, the daily limit cannot be exceeded more than 35 days per year. Since 2015, an annual limit has been imposed also for PM2.5 concentration[34] (25 µg/m3). Even though current laws established maximum thresholds on the atmospheric concentration of PM10 and recently PM2.5, following WHO's recommendations and guidelines, it should be clarified that the actual existence of a threshold for health risks has not been demonstrated. In the area of Sannazzaro and Ferrera, the annual mean concentrations of PM10 measured by ARPA in the years 2008-2014 were around 30 µg/m3, in line with those recorded in the rest of the Province[35] [*ARPA-L, 2015*].

[34] PM2.5 also contains black carbon, which is a major driver of climate change. See http://www.who.int/airpollution/ambient/pollutants/en/ (opened on August 30th, 2018).
[35] Measures of PM2.5 are not reported because limits were defined (and measured) only since 2015.

Regarding the health effects, their interpretation is made more difficult by the heterogeneity of PM, with particular reference to size (which determines the capability of the contaminant to reach different districts of the human body and to be cleared) and chemical composition [*WHO, 2006*]. Fine PM is likely to be deposited in smaller airways and alveoli [*WHO, 2006*]. Besides, ultrafine PM has the capability of penetrating the body via systemic circulation and even to reach the brain, thus being more toxic than PM10 [*Bernstein et al., 2004*]. Moreover, the biopersistence of particles in the body depends on solubility, which in turn depends on the specific characteristics of the PM to which the body has been exposed [*WHO, 2006*]. Concerning chemical composition, fine PM – as already mentioned – can be formed by condensation of other hazardous pollutants, but in general the particles have the capacity of absorbing other substances – e.g. PAH or heavy metals – which could significantly increase toxicity [*ARPA-L, 2015*].

Epidemiological and clinical investigation have linked exposure to PM with various consequences for human health, with plausible toxicological mechanisms. Several studies showed, among others, exacerbated symptoms in patients with asthma and COPD (with an increased risk of death or hospitalisation for the latter), lung and systemic inflammation, cancers, increased mortality and morbidity rates for patients with cardiovascular conditions, increased risk of acute myocardial infarction, and even a higher susceptibility to infectious diseases [*WHO, 2006*]. However, the role on respiratory chronic diseases was not confirmed [*Sunyer et al., 2006*]

1.6.1.4. Carbon monoxide (CO)

Carbon monoxide (CO) is a primary pollutant, coming from an incomplete combustion process of fuels containing organic compounds. In urban contexts, it is highly related to road traffic, and indeed the maximum concentrations of this contaminant tend to be observed during rush hours [*ARPA-L, 2015*].

The regulatory threshold for CO, according to the law in force in 2018[30], is set as a daily limit of 10 μg/m^3 (8-hours moving mean). Between 2008 and 2014, the annual mean concentration of CO observed by ARPA in the area of Sannazzaro de' Burgondi and Ferrera Erbognone was less than 1 mg/m^3 [*ARPA-L, 2015*].

Exposure to CO is extremely toxic for humans because this molecule can bind human haemoglobin with an affinity 220 times higher than CO_2. The resulting complex, called carboxyhaemoglobin, is physiologically inactive; thus, the capacity of blood red cells to oxygenate tissues in the body is reduced. Moreover, exposure (including chronic exposure to low doses) can worsen the clinical conditions of a subject, with particular reference to cardiovascular conditions.

1.6.1.5. Volatile Organic Compounds (VOC)

Volatile Organic Compounds (VOC) is the collective name for a wide variety of organic compounds that are present in the atmosphere in vapour phase, like benzene, toluene, xylenes, ethylbenzene, methyl-ethyl ketones, acetophenone, and trichloroethylene [*IARC, 2016*]. They can be the result of leakage of gaseous fuels, evaporation of liquid fuels, or they can be produced from incomplete burning during combustion processes and incineration [*WHO, 2006*]; in Europe, half of the emitted VOCs come from the use of solvents [*IARC, 2016*]. They also play a relevant role in the formation of secondary pollutants. Since their monitoring is not currently required by law, their levels were not monitored by ARPA.

Epidemiological research on the health effects of VOCs has not been as wide as for other air pollutants. A Canadian cohort study identified an excess in mortality associated to VOCs. However, some toxicological evidence has been made available, showing for instance mutagenic power [reported in *IARC, 2016*].

1.6.2. Brought to the lunch table: pollutants and food chain

The area around the refinery in Sannazzaro de' Burgondi and Ferrera Erbognone, except for the refinery itself, is a land of strong agricultural vocation. For this reason, a peculiar aspect of environmental pollution is worth spending a thought: the contamination of the food chain, occurring at any level of the production process, which can result in a dietary intake of contaminants that sums to other exposure routes [*Mancini et al., 2016*; *Vanni et al., 2016*]. An integrated assessment approach needs to consider the contaminants entering the food chain and their bioaccumulation [*Mancini et al., 2016*; *ISPRA, 2016*]. Obviously, for the sake of assessing the health impacts of a refinery on the resident population living nearby, the concern is relevant as long as the subjects in the population make use of locally produced ("km-zero") food, although this does not mean that they may not consume contaminated food produced elsewhere.

1.6.3. Health risks associated to a refinery: epidemiological perspectives

Broadly speaking, the effects of air pollutants depend on their atmospheric concentration, their persistence and their physical and chemical characteristics and, with regard to the receptors, their peculiarities and the time of exposure [*ARPA-L, 2015*]. The impact of pollutants on health can occur either in the short term (mainly because of high levels of contamination) or in the long term (with a prolonged, but not necessarily high-level, exposure). In particular, the effects of chronic exposure to mixtures of different contaminants at low concentrations are far from being ascertained [*ARPA-L, 2015*]. The risk of developing a certain health condition as a consequence of exposure to air pollution can also be moderated by pre-existing conditions or predispositions: different individuals may not respond to the same exposure in the same way because of underlying intrinsic and extrinsic factors (e.g. age, gender, genetic background, socio-economic status, nutrition, lifestyle) determining inter-subjects variability. This is commonly referred to as *susceptibility* [*WHO, 2006*].

The impact of air pollution on the general population is almost certainly dominated by subclinical or even asymptomatic conditions, rather than by severe events like hospital admissions and deaths (***Figure 1.7***, next page). Despite that, information regarding severe outcomes is generally more easily available, "visible" and robust, and therefore most epidemiological evidence is based on the study of severe health events [*WHO, 2006*]. Finally, it is important to look at the consequences of air pollution on health from the standpoint of public health: even though the risks associated to such exposures are low or very low, the proportion of exposed subjects in the population is extremely high and so, in terms of *impact*, these health effects represent a considerable burden [*WHO, 2006*].

Some epidemiological studies [*Piratsu et al., 2011*; *Cernigliaro et al., 2006*; *Fano et al., 2006/a*] found an excess of overall and cancer-specific mortality, as well as an increase in hospitalizations due to cardiovascular and respiratory diseases. In the short term, a metanalysis showed how the presence of petrochemical plants increases the risk of death from respiratory causes [*Biggeri et al., 2001*]. Among those who resides near an industrial area, in comparison with people living in non-industrial areas, an increment in the risk of developing lung cancer was observed [*Benedetti et al., 2001*]. These results are consistent with previous evidence and have been confirmed by subsequent researches on several Italian areas with strong industrial pollution like Civitavecchia, Gela, Porto Torres, Falconara, Cadeo, Priolo [*FANO ET AL., 2006/A*; *FANO ET AL., 2006/B*; *PIRATSU ET AL., 2011*; *CERNIGLIARO ET AL., 2006*]. Notably, another Italian study investigated the role of non-occupational exposures in the risk of lung cancer among workers of a petrochemical plant built in Gela in the Sixties; exposure was assessed as living closer or further from the refinery during the time spent working in the refinery in the timespan 1960-1993. A possible excess of mortality by lung cancer was found among those workers that used to live closer to the plant [*PASETTO ET AL., 2008*]. A multidisciplinary longitudinal project in the area was later implemented [*MUSMECI ET AL., 2009*].

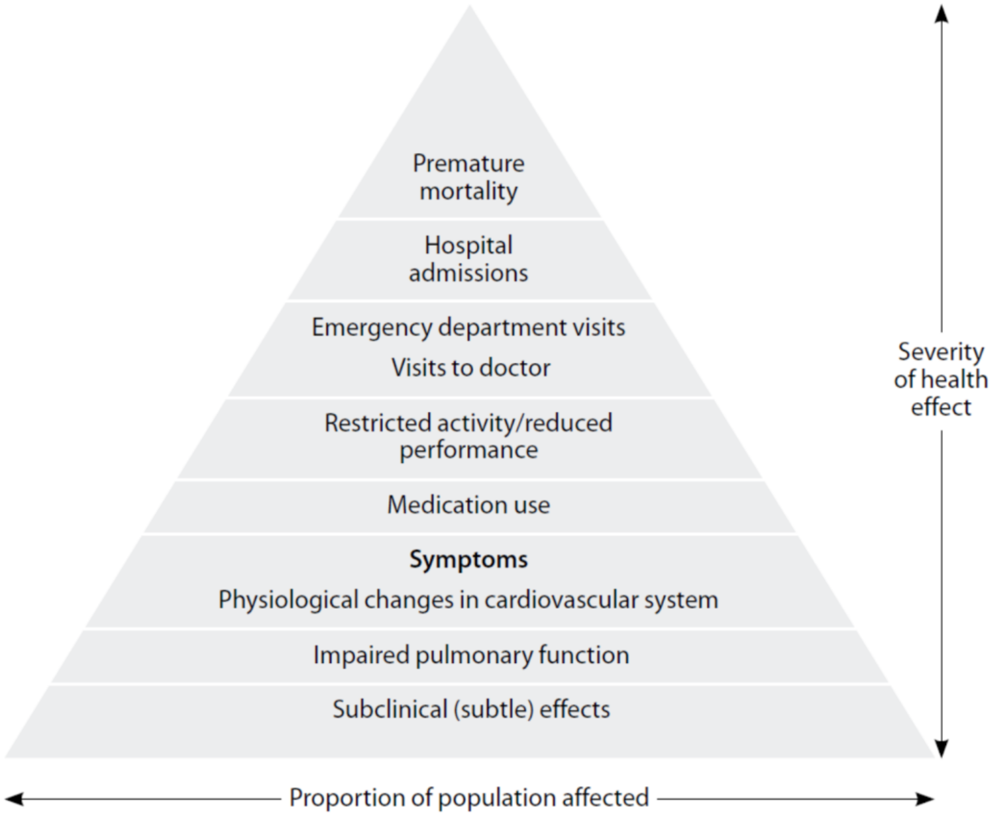
Premature mortality
Hospital admissions
Emergency department visits
Visits to doctor
Restricted activity/reduced performance
Medication use
Symptoms
Physiological changes in cardiovascular system
Impaired pulmonary function
Subclinical (subtle) effects
Severity of health effect
Proportion of population affected

***Figure 1.7** – Pyramid representing severity of air pollution effects and their relevance in the population. Reprinted from the American Thoracic Society, reported in* WHO, 2006.

In addition to the aforementioned effects, in some of these studies an increase in mortality was observed also for non-pulmonary oncological diseases (stomach, colon, larynx, bladder, non-Hodgkin's lymphoma). A study [*BUDRONI ET AL., 2010*] found mortality by any oncological cause to be increased among workers of the refinery in Porto Torres, and the association was stronger particularly for non-Hodgkin's lymphoma. Moreover, in those ICSs characterised by the presence of refineries and petrochemical industries (Falconara, Cadeo, Priolo), an increased risk of pleural mesothelioma was identified [*Fazzo et al., 2012*]. Other studies highlighted possible effects of air pollution also concerning the gastrointestinal system, either of inflammatory or neoplastic nature, occurring both as a direct consequence of environmental contamination and as indirect consequences of systemic effects, in the adult population [*Beamish et al., 2010*] and among children [*Orazzo et al., 2009*].

Outside Italy, other studies have disclosed associations between air pollutants released by petrochemical plants and the development of haemolymphopoietic tumours, either in children [*Weng et al., 2008*] or in adults [*Dahlgren et al. 2008*; *Kirkeleit et al., 2008*, *Barregard et al., 2009*]. In the age group 20-29 years, a case-control study conducted in China showed that residential exposure to emissions from this type of plants constituted a risk factor for the development of leukaemia [*Yu et al., 2006*]; similar findings came from Spain, regarding emissions from refineries and mortality from non-Hodgkin lymphoma [*Ramis et al., 2012*] and, in Nigeria, for non-Hodgkin lymphomas and petrochemical industries [*Omoti et al., 2006*]. A study conducted in Taiwan also reported that the risk of death from brain tumours was raised among those exposed to higher levels of pollution [*Liu et al., 2008*].

As regards respiratory function, several studies have found it to worsen in humans when, in the area where they live, the level of PM10 raises [*Ackermann-Liebrich et al., 1997*; *Schikowski et al., 2005*]; however, these findings were not confirmed for PM2.5 [*Götschi et al., 2008*]. A significant association was disclosed between outdoor levels of nitrogen dioxide and chronic bronchitis among women [*Sunyer et al., 2006*].

There are, however, also negative findings. A metanalysis about the relationship between the non-Hodgkin's lymphoma and employment in the production, distribution or use of gasoline reported no significant association [*Kane et al., 2010*]. Another study – conducted in the district of Pancevo (Serbia), where one of the greatest petrochemical plant of the nation is located – found no relationship between air pollution and cancer in the population [*Bulat et al., 2011*]. A study conducted in Louisiana (US) showed that cases of lung cancer had a higher probability to have lived near a petrochemical plant; however, controlling for other risk factors, the association tended to disappear [*SIMONSEN ET AL., 2010*]. A research conducted among workers of a Bulgarian petrochemical plant compared with 50 controls, using biomarkers for the sake of exposure assessment, disclosed no significant dose-response relationship for most of the haematological parameters under investigation [*PESATORI ET AL., 2009*].

The extensive review published in 2010 by the SENTIERI working group found that, for refineries and petrochemical plants, the epidemiological evidence regarding causal associations of several diseases with the exposure was substantially lacking [*SENTIERI, 2010*][36]. Specifically, with reference to the general population, the evidence supporting an association with mortality by any cause was deemed inadequate, and the same was for deaths by any oncological cause and several specific cancers; only for cancers of the lungs, bronchi and trachea, they found limited evidence. Epidemiological findings were inadequate also with regards to diseases of the cardiovascular system and chronic lung diseases, and limited for respiratory diseases, acute respiratory diseases, asthma, and congenital malformations. For this reason, it becomes necessary to provide new evidence regarding morbidity for cardiovascular and respiratory diseases. Epidemiological evidence was generally more robust with regards to "general" exposure to air pollution: in this case, it was considered to be sufficient for mortality by any cause, cancers of lungs, bronchi and trachea, cardiovascular diseases in general and also specific cardiovascular conditions (acute myocardial infarction, ischemic heart disease, brain circulatory diseases, acute respiratory diseases, as well as for the worsening of asthma, chronic respiratory diseases and all respiratory diseases.

Finally, as noted by WHO [*WHO, 2013*], accidents are one of the sources by which an industrial site can contaminate the nearby environment, with the consequent health risks for the population, but no review of studies regarding industrial accidents in Europe is available at the time.

[36] The review reported also epidemiological evidence regarding the effect of several confounders (including smoking, passive smoking, alcohol, socio-economic status and occupation).

1.7. EPI-EST: THE CONSAL PROJECT

When the authorisation for the new EST facility was issued, the authorisation body embraced the concerns of local stakeholders regarding the potential risks associated to the new facility and to the refinery as a whole, including as a condition the provision of an epidemiological study of the health status of people living in the area. The Department of Public Health, Experimental and Forensic Medicine of the University of Pavia, with its sections of Biostatistics and Clinical Epidemiology, of Occupational Medicine and of Hygiene, developed an epidemiological project according to these provisions. The Project is called EPI-EST or CONSAL (the latter, standing for *Knowledge and Health*, or in Italian *CONoscenza e SALute*, will be used in this thesis).

The EST facility was, in a certain sense, the tip of the iceberg, given that epidemiological evidence was basically lacking in the area since the refinery was commissioned in the Sixties[37]. For this reason, the general aim of the CONSAL Project was to study public health of the adult population[38] living in Sannazzaro de' Burgondi and Ferrera Erbognone in relation with the emissions from ENI's refinery, both before the commissioning of the new EST facility (*ante-operam* phase) and after commissioning (*post-operam* phase)[39].

The CONSAL Project included four epidemiological studies, and as of 2018 not all its parts have been concluded. The first study (S1) had the aim to investigate the association between exposure to the emissions from the refinery and various health outcomes, both *ante-operam* and *post-operam*, with a case-control design. The *ante-operam* phase of S1 is the subject of the present thesis, and therefore it will be described in detail in the next Chapters. The second study (S2), with a cross-sectional design, was aimed at assessing the respiratory health status of the population, with a specific focus on the prevalence of asthma and chronic obstructive pulmonary disease

[37] The only exception was represented by a previous analysis referred to the period 1995-2000, in which standardised mortality rates among residents in the area of Sannazzaro were found to be substantially comparable with those observed in the rest of the Province of Pavia [reported in *ENI, 2008*].

[38] The possibility of including also the younger population was excluded after meetings with the stakeholders, the local departments of the agencies of environmental and health protection, and the Ministry.

[39] As of 2018, because of the accident occurred at the end of 2016, the EST facility was not functioning and was undergoing a major maintenance. Therefore, the protocol of the *post-operam* phase will be adapted, at least for what concerns timing.

(COPD); the disease-free cohort of S2 was also designed to be the base for the third study (S3), a longitudinal cohort study aimed at estimating the prevalence rates of asthma and COPD in 7 years. Finally, the fourth study (S4) had the objective of capturing the health status – in terms of town-level crude, standardised and adjusted mortality and hospitalisation rates (by all causes and cause-specific) – of the entire population living in the Province of Pavia, with a specific focus on those from Sannazzaro and Ferrera, and to compare the rates observed in the two towns of interest with the rates in the rest of the Province and in the Region of Lombardy. While for S1, S2 and S3 it was planned to collect data directly from the subjects by means of questionnaires[40], for the sake of S4 only data from the administrative databases were used. The timeline of the CONSAL Project is reported in ***Figure 1.8***.

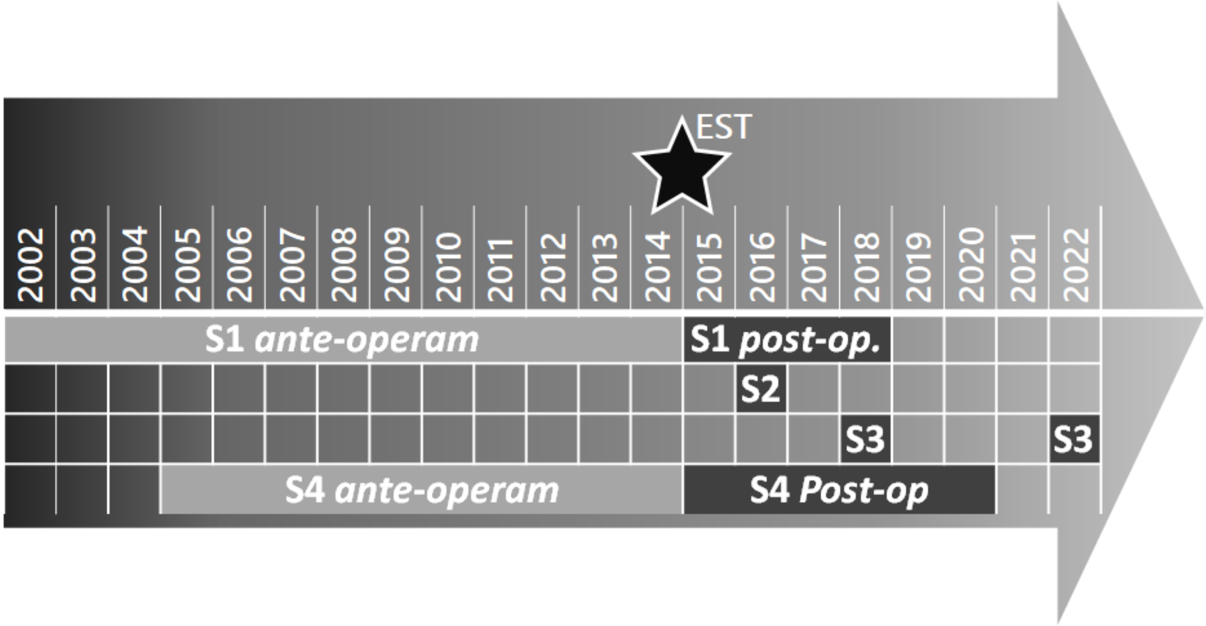

Figure 1.8 – *Timeline of the CONSAL Project. S1: Case-control study; S2: Prevalence study; S3: Incidence study; S4: Mortality & Hospitalisation study. The timeline here reported does not take into account variations due to the accident occurred at the EST facility in 2016, affecting the post-operam phase, because as of 2018 the facility was still under maintenance and, therefore, such variations still had to be defined.*

[40] Actually, in S1, data regarding the health outcomes were collected from the administrative databases, as will be explained in Chapter 3 (see ***3.2.3 Sample size estimation***).

2. Aims

The general aim of the CONSAL Project was defined in the ***Introduction*** to the present work. The present thesis is focused on the first part (*ante-operam assessment*) of Study I.

The specific aim of the study presented in this thesis was to investigate the health impacts of air pollution on a population living near a refinery (municipalities of Sannazzaro de' Burgondi and Ferrera Erbognone), with particular reference to the effects of contamination from the point sources pertaining to the refinery. This study was aimed at producing mutually adjusted estimates of the effects of the environmental exposure by collecting additional information through a survey.

3. Methods

3.1. STUDY DESIGN

The investigation was implemented as a ***case-control* study**. This is an *observational*, *analytical* (i.e. controlled) and *longitudinal* epidemiological design. By definition it is logically retrospective, as it looks first at the outcome and then at the exposure; this means, obviously, that it is also retrospective in terms of research implementation[41].

In a nutshell, the case-control approach is a way to make the observation of a population more efficient [*ROTHMAN & GREENLAND, 1998*]. Other types of longitudinal studies, like cohort studies, are aimed at estimating directly a *rate* of disease occurrence among exposed and unexposed people. Albeit it is a very elegant reasoning, and very close to one of the foundations of causality (temporal antecedence of the cause to the effect), a caveat is that all the population have to be enumerated and monitored within a certain period of time. A case-control design is convenient because, instead of assessing the whole population (i.e. the denominators of the rates for a cohort study), it starts from defining the current status of subjects with regards to the presence of a specified outcome ("*cases*") or its absence ("*controls*"). Cases are the same people that would be diseased in a cohort study, whereas controls should represent the distribution of the exposure in the general (target) population that originates cases [*ROTHMAN & GREENLAND, 1998*]. Exposure is then assessed in the same way both in cases and controls [*BACCHIERI & DELLA CIOPPA, 2004*].

3.1.1. Case definition

The status of case was defined as being admitted to hospital at least once for an acute condition (i.e. no planned admissions) for one of the following main causes (as coded under the *International Classification of Diseases*, version IX-CM[42], or ICD-IX-CM):

[41] It is worth recalling the difference with a cohort study, which is always logically prospective and can be either prospective or retrospective in its implementation.

[42] ICD-IX-CM was the ICD version used by ATS Pavia at the time of data collection (see ***3.2.3.2 Hospital Discharge Database***).

- Diseases of the respiratory system (*Chapter 8, codes 460-519*);
- Symptoms involving respiratory system and other chest symptoms (*Chapter 16, code 786*);
- Diseases of the circulatory system (*Chapter 7, codes 390-459*);
- Symptoms involving cardiovascular system (*Chapter 16, code 785*);
- Diseases of the digestive system (*Chapter 9, codes 520-579*)[43].

Hospitalisations were considered only if they occurred during the *ante-operam* timespan under investigation (2002-2014). In case of multiple admissions for the same subject, the first in chronological order was considered.

On the other hand, controls were subjects that had no unplanned hospitalisation recorded (for any cause) in the same timespan.

3.1.2. Advantages and limitations of the case-control design

The main advantage of a case-control design is that, among the possible choices of analytical observational studies using individual-level data, it usually require less resources than other approaches (e.g. cohort studies) and it can be less time-consuming, especially if compared to prospective studies [*BACCHIERI & DELLA CIOPPA, 2004*]. The choice of a retrospective study design, whilst potentially more vulnerable to information bias, was indeed aimed at quickly obtaining a first result regarding potential health effects due to air pollutants from the refinery referred in the *ante-operam* phase. However, it must be emphasised that, in the framework of the CONSAL Project, other studies (and in particular a prospective cohort study) were designed to later confirm or refute the findings from the case-control study. In addition, required sample sizes are generally lower in case-control than in cohort studies (this being particularly true for outcomes with low prevalence) [*BACCHIERI & DELLA CIOPPA, 2004*].

Nevertheless, a case-control design comes with some matters of concern. First, it is straightforward that the timing of exposure assessment is a critical aspect and should be defined carefully. In any case, although exposure is theoretically referred to the past, the fact that outcome and exposure are assessed at the same time cast doubts regarding

[43] This set of ICD-coded causes was added in the amended version of the protocol, after convincing evidence of a possible role exerted by air pollution was found in scientific literature (see ***1.6.2 Brought to the lunch table: pollutants and food chain***).

temporal antecedence of the exposure to the outcome and, thus, this type of study cannot give robust evidence of causality [*Bacchieri & Della Cioppa, 2004*].

It should be noted that, as will be discussed later in this Chapter, the case-control study was planned to collect also other relevant individual information (lifestyle, housing, other health issues) by sending questionnaires directly to enrolled subjects.

3.1.3. Implications for statistical analysis and interpretation

Exposure is measured in terms of the *odds* of being exposed or not in the group of cases and in the one of controls. To understand the effect of exposure on the outcome, the *odds* of exposure in the two groups (cases and controls) are compared with appropriate epidemiological measures (Odds Ratios, ORs; see ***Appendix B.4***). Additional information about study participants might be collected, with the aim to produce adjusted estimates of the effect. It is useful to remark that an immediate interpretation of ORs as "risk estimates" is not appropriate, because an OR represents how greater are the odds of being exposed in cases rather than in controls, and not the ratio of the probabilities (absolute risks) of developing the outcome in exposed and non-exposed people (Relative Risks, RRs[44]). That said, under certain conditions ORs from case-control studies may allow researchers to compute RRs and incidence rates [*Rothman & Greenland, 1998*].

Given its retrospective nature, it is straightforward why a case-control study is more vulnerable to *information bias* [*Bacchieri & Della Cioppa, 2004*]. Data are often extracted from databases that were not developed for that research (e.g. administrative databases of healthcare facilities or healthcare systems) and might be inaccurate or lacking. If missingness of outcome data occurs differentially between cases and controls, it may result in selection bias. Another possibility is to get data directly from study participants, but information might be affected by inaccuracies and *recall bias* (especially regarding the exposure). In addition, *selection bias* can occur also in this case if lack or inaccuracy of outcome data is differential between cases and controls.

[44] RRs are clearly unsuitable as effect measures for case-control studies, given that the number of cases is decided by the investigator and those measures depend on the number of cases.

3.2. POPULATION AND SAMPLE

3.2.1. Target population and eligibility

The target population of this study is the population of adults living in the municipalities of Sannazzaro de' Burgondi and Ferrera Erbognone (Lomellina district, Province of Pavia) during the *ante-operam* timespan (2002-2014).

According to the protocol, the eligible sub-population consisted of all the subjects aged between 20 and 64 in the period 2002-2014 (*ante-operam* timespan), who were residents in the municipalities of Sannazzaro de' Burgondi and Ferrera Erbognone for at least a part of that time and were currently alive and resident. Further criteria to be met were those to be classified as case or control (see ***3.1.1 Case definition***).

According to the National Institute of Statistics (ISTAT), on January 1st, 2015 there were 5512 registered residents in Sannazzaro de' Burgondi, of which 2980 in the age range 20-64 years (1554 males, 1426 females). In Ferrera Erbognone, the total number of residents was 1165, of which 668 in the age range 20-64 years (376 males, 292 females)[45]. By October 31st of the same year[46], again according to ISTAT, the total resident population slightly changed (5514 in Sannazzaro, 1183 in Ferrera)[47].

3.2.2. Sample size estimation

Sample size was estimated *a priori* by using the formula for unmatched case-control designs, as proposed by Fleiss (see ***Appendix B.1***). The unadjusted OR was assumed to be 1.6 from previous researches (see *BERTOLDI ET AL., 2012*, where environmental risk factors for respiratory diseases were studied), and exposure was considered to affect a share of 40% of the general population. The significance threshold was set at 5% (α=.05) and power at 80% (β=.20). The number of controls per each case was fixed at 3:1. This ratio was decided in order to enrol enough subjects to ensure a sufficient

[45] All the data reported here have been taken from ISTAT's demographic series and are available at the following link: *http://demo.istat.it/pop2015/index.html* (opened on July 26th, 2018).

[46] This calendar date was chosen as reference day because it is the end of the last month before data were collected.

[47] Data available at the following link: *http://demo.istat.it/bil2015/index.html* (opened on July 26th, 2018). The distribution by age is not available in the monthly demographic series.

statistical power in the eventuality of a stratified analysis, taking into account the number of cases in the population of interest, and the fact that increasing the number of controls might be less demanding (in terms of investigation costs) rather than enrolling more cases. Given these parameters, the required sample size resulted to be 762, of which 191 cases and 572 controls[48].

The numbers reported above were increased in order to face a share of non-available subjects of up to 10%, leading to a final sample size estimation of 210 cases and 630 controls (and a total of 840 subjects to be enrolled).

3.2.3. Data sources

3.2.3.1. Regional Registry of Health Insurance

Relevant personal information (name, gender, birth date, individual fiscal code, individual healthcare system code, declared residency address, name of the General Practitioner) were extracted from the regional database of health-insured citizens, maintained by the local ATS of the Province of Pavia (ATS Pavia) for the citizens of its district. Information in the list of health-insured citizens maintained by ATS is periodically updated with data regarding the demographic balance sent to ATS by the Register Officers of all municipalities. As a consequence, those lists might present small differences from the "real population", due to the time needed to process those updates and realign the lists. For this reason, the composition of the sampled population was further verified in November 2015 with the Municipal Registries of the two municipalities, in order to ensure it was up-to-date with migrations and deaths and that the addresses were correctly declared to ATS by the subjects.

3.2.3.2. Hospital Discharge Database

Data regarding the health status of the sampled population were extracted from the database of the Hospital Discharge Records (*Schede di Dimissione Ospedaliera*, SDO), maintained by ATS Pavia for administrative purposes related to the management of health services. Data from the SDOs database (and, in particular, those regarding time

[48] A number of on-line tools are available for the sake of sample size estimation in case-control studies. For instance, the calculation reported here can be reproduced by using the one at the following link: *http://www.openepi.com/SampleSize/SSCC.htm* (opened on July 25th, 2018).

and main cause of admission, coded under ICD-IX-CM specifications) were linked with personal information of the subjects with a deterministic linkage procedure, by means of the individuals' healthcare system codes and fiscal codes.

3.2.3.3. A focus on the Hospital Discharge Records

The collection of hospitalisation data by means of SDO records was introduced in 1991[49] in the intent of rationalising and improving the setting of healthcare system by setting up a structured data stream based on a standardised tool for use in health economics and management[50]. A SDO record contains all the relevant information about a hospitalisation event, like personal information of the patient, type of admission (urgent, planned, day-hospital), type of discharge, healthcare facility where the recovery took place, main diagnosis and concurrent diagnoses (coded under the ICD classification[51]), diagnostic and/or therapeutic procedures administered to the patient during the recovery[52]. The use of administrative databases is convenient because information is "immediately" available (meaning that data have already been collected, recorded in electronic form, and checked); however, it should be stressed that it comes with potential issues limiting their usefulness in epidemiological research. Those databases, indeed, are not primarily conceived for research purposes: relevant information might be lacking or clinically inaccurate. Thus, SDOs are not always reliable and suitable for scientific research.

3.2.4. Sample selection

The number of urgent hospitalisations in the years 2002-2014 among residents in Sannazzaro de' Burgondi and Ferrera Erbognone, as extracted from ATS Pavia's databases, was 1666; in case of multiple admissions referred to the same subject, only the first one complying with the outcome definition was considered. After filtering out those concerning subjects who not met the eligibility criteria, the number of eligible cases was found to be 266 (of which 222 from Sannazzaro de' Burgondi and 44 from

[49] Decree of the Ministry of Health, December 28th, 1991.

[50] A detailed description, on which the information reported here were based, can be found at the following link: *salute.gov.it/portale/temi/p2_6.jsp?id=1232&area=ricoveriOspedalieri&menu=vuoto* (opened on July 26th, 2018).

[51] The Minister of Health periodically issue a decree to update the adopted version of the ICD.

[52] It is important to remind that information regarding administered drugs and (if present) adverse drug reactions are not reported in SDO records, as they are collected by means of a different information flow.

Ferrera Erbognone). Numbers slightly changed when eligibility was verified with the Municipal Registries: only 9 cases had to be dropped from the list (8 from Sannazzaro, 1 from Ferrera). It was decided to enrol all the eligible cases (257 instead of the estimated number of at least 210, demonstrated in ***3.2.2 Sample size estimation***), thus potentially allowing for a higher share of non-respondents or unavailable subjects.

Eligible controls, as resulting in ATS Pavia's database, were 1240 (1063 and 173 in Sannazzaro and Ferrera, respectively). For what concerns controls, it is necessary to remind that the case-control ratio was previously mentioned to be fixed at 1:3 (i.e. roughly 771 controls, given the increase in the number of cases). Moreover, in the protocol it was stated that their frequencies regarding gender, age class (width: 5 years) and municipality of residency had to be balanced with those observed in cases, in order to reduce potential confounding effects.

Thus, controls to be enrolled were drawn from the sampled population with a random procedure, balancing the selection by gender, age class and municipality of residency. The number of eligible controls from Ferrera Erbognone left out from the enrolment was found to be low, so that the 176 eligible controls from Ferrera Erbognone were all included in the final sample; the case-control ratio among residents in that municipality was close to 1:4. In other words, only the 622 controls from Sannazzaro de' Burgondi underwent random selection. Also controls, like cases, were further verified with the Municipal Registries: 5 subjects from Sannazzaro and 4 from Ferrera were dropped from the list, so that enrolled cases were 617 in Sannazzaro (case-control ratio: 1:2.88) and 172 in Ferrera (case-control ratio: 1:4.00)

To sum up, at the end of the procedure of sample selection, the enrolled sample consisted of 1046 subject, of which 257 cases and 789 controls, 831 from Sannazzaro and 215 from Ferrera. The flowchart of sample selection is reported in ***Figure 3.1***.

For the sake of completeness, it should be mentioned that, at the beginning of study planning, the timespan for enrolment was shorter than the final 2002-2014, but the number of eligible cases was not sufficient to ensure the study would have been powerful enough. Thus, the timespan had to be extended to include a few more years in the past.

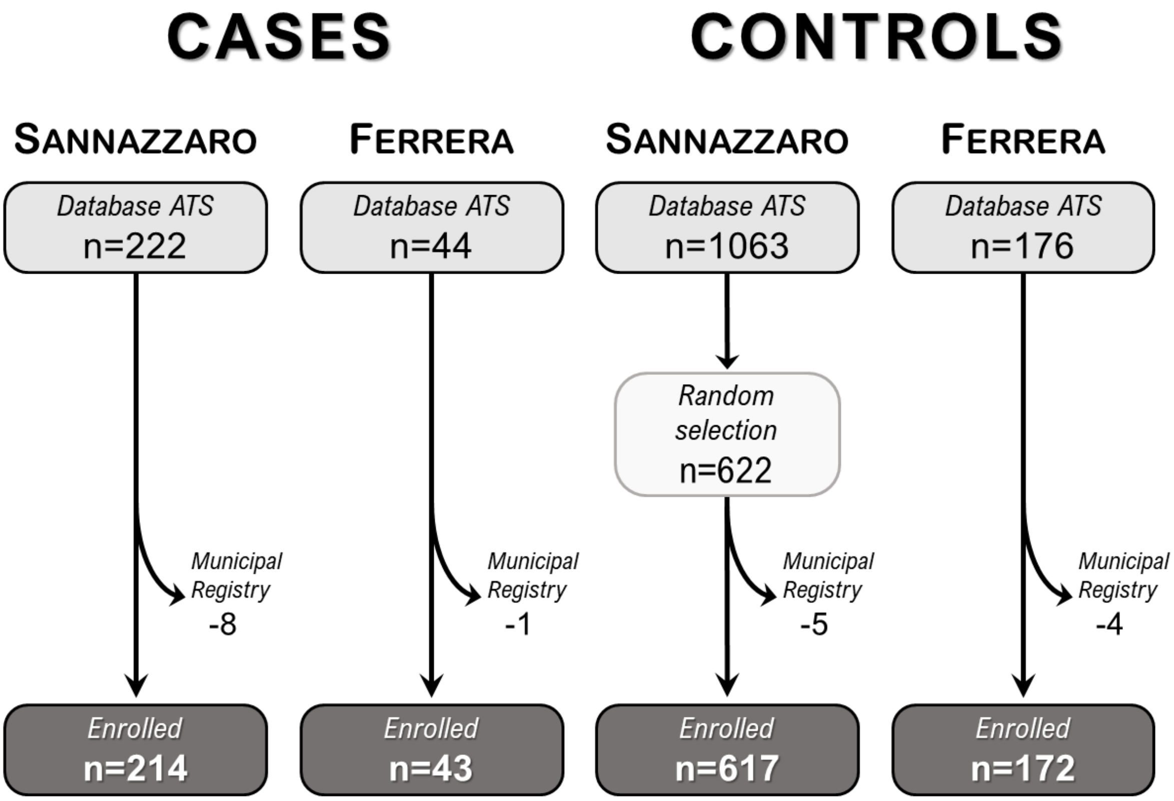

Figure 3.1 *– Flowchart of sample selection.*

3.3. STUDY IMPLEMENTATION

3.3.1. Data collection

Enrolled subjects were contacted various times ("waves") in order to ask them to take part in the study and fill in the survey questionnaire. In each wave of contacting, the subjects received the questionnaire packed together with a consent form and a cover letter in which the study was explained and instructions on how to return the questionnaire and the signed consent form were given (***Appendix C***). In all the contact waves but the first, a different accompanying letter highlighting the reasons for participation was sent. A flowchart of contact waves is represented in ***Figure 3.2*** (at the end of the section).

After preliminary meetings were hold in the attempt to engage pharmacists and the General Practitioners from the outpatient clinic of Sannazzaro de' Burgondi, the first wave of contacting started in January 2016. Parcels containing questionnaires and annexes were delivered by volunteers of the local Civil Protection group, asking people to fill in the questionnaire and return it in boxes placed in the local pharmacies.

In March 2016, enrolled subjects were reminded to return the questionnaire: a news was published on the websites of the town councils, and the municipality of Sannazzaro sent to its citizens an SMS and e-mailed a newsletter.

As of April 2016, 461 questionnaires were returned. Of the 585 non-respondents, 8 refused to participate, 4 were not available at their registered home address, 4 were subjects who emigrated, and one was found to have died in the meantime. A strong difference in the rate of respondents[53] was observed between the two municipalities (39.2% in Sannazzaro, 65.9% in Ferrera). However, those rates were not deemed sufficient, thus it was decided to contact again the non-respondents with different strategies for the two municipalities: in Ferrera, the Mayor committed to contact the subjects in person, while in Sannazzaro the General Practitioners were asked to do so. These strategies, in any case, contributed little in increasing the number of respondents, which was 474 in June 2016 when the first wave of contacts was closed.

[53] Respondence rates were computed after excluding deceased and unreachable people from the sample.

In the attempt to raise respondence, a new wave of contacts started in July 2016[54]. This time, the 559 questionnaires were delivered via a private postal service[55]; again, participants were asked to return them in boxes placed in the local pharmacies for the purpose, as explained in the cover letter. This second wave was closed at the beginning of October 2016, but the response rates[56] were still unsatisfactory, and particularly in Sannazzaro only 46.8% accepted to participate in the study (against 73.8% in Ferrera). As suggested also by the Mayors of the two municipalities involved in the study, this wave probably started too close to summer vacations.

After the conclusion of the second wave of contacting, it was attempted to involve local third-sector associations and groups, as they could be more effective in promoting participation thanks to their direct networks with residents. A first meeting was held in October 2016 with more than 30 representatives from different associations, initially willing to help with contacting non-respondents and delivering them the questionnaires. A second meeting lacked participation, so that this approach was abandoned.

A third wave, which relied on students[57] and on the outpatient clinic, produced scarce improvements as well. Data collection was declared definitively closed in October 2017. A total of 563 subjects returned the questionnaire, while non-respondents were 453 (after excluding 30, either deceased or unreachable subjects). The response rate[56] was 49.1% among subjects from Sannazzaro de' Burgondi, whereas in Ferrera Erbognone it amounted to 75.7%.

[54] The beginning of the second wave had to be delayed in order to avoid concurrence with the local elections in Sannazzaro de' Burgondi (June 2016).

[55] The second wave of Study I was carried out together with the first wave of Study II.

[56] Again, deceased or unreachable subjects were excluded from the computation of response rates.

[57] The students, attending the third grade in two High-schools in the Province (ITIS G. Cardano, Pavia, and IIS A. Maserati, Voghera) were involved thanks to a project for the school-work alternation. Beyond their field work, they also attended seminars for training. All the students were residents in Sannazzaro or Ferrera.

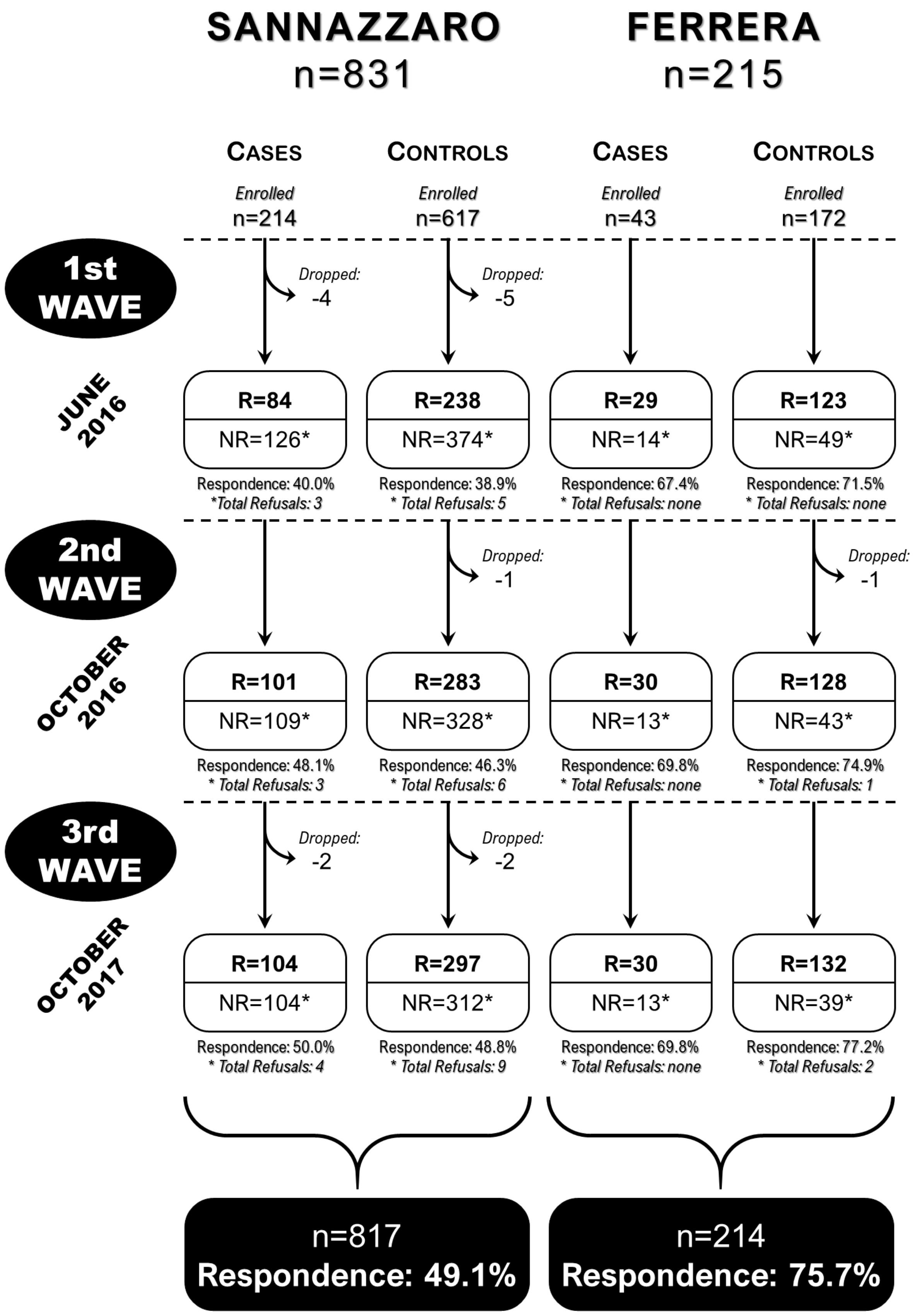

Figure 3.2 *– Flowchart of the "waves" of contacting of enrolled subjects.*

3.3.2. Data management

Two electronic databases were implemented in Microsoft Access. A first database – the ***Personal Information Database*** – stored all the information allowing the immediate identification of a subject (e.g. name, surname, fiscal code, healthcare system code), plus a ***pseudo-anonymous* code**[58] created to serve as the subject's identification in all the activities related to the study (when personal information was not strictly needed). This code was created so that it encrypted basic information about the subject (municipality, gender, year of birth, case/control status), thus making them available without any need to access the database of personal information[59]. The code was printed on every questionnaire in a graphic format (QR Code[60]) as well as in numbers (numeric string). The Personal Information Database was protected with a password, and only authorised personnel was allowed to access the table with personal information and identification codes.

A second, separate database, the ***Survey Database***, was used to record paper-collected questionnaire data in electronic form. On this regard, to reduce the chance of mistakes, data entry was made by trained personnel using a structured entry form. The pseudo-anonymous identification code was either inputted manually (as a numeric string) in the data entry form or (preferentially) read with a camera (as a QR Code), resulting in the ID field automatically completed in the entry form. The system automatically resolved the code to decrypt municipality, gender, year of birth and status, which were recorded in the entry. Besides, gender and year of birth were used to automatically check consistency with those self-reported by the participant in the questionnaire. In case of inconsistency, the system returned an error and the questionnaire was further checked in order to figure out if it was completed by someone different from the enrolled subject to whom it was addressed; in that case, the

[58] According to Article 4 of the General Data Protection Regulation (GDPR), which entered into force in all the Member States of the European Union on May 25th, 2018, pseudo-anonymisation is defined as "*the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person*". GDPR is available at *http://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1533822227828&uri=CELEX:32016R0679* (opened on August 8th, 2018).

[59] This system was useful to check consistency in survey data, as it will be explained below. The criterion used to generate the identification codes was protected.

[60] The QR (for *Quick Response*) Code is a bi-dimensional bar code.

questionnaire was discarded. Only for those questionnaires in which the respondent did not indicate gender and/or year of birth, the identity was verified by looking at the consent sheet. With regard to the other parts of the questionnaire, database's fields for close-ended questions were strictly encoded, and internal checks were imposed when possible (e.g. on calendar dates and nested questions).

The Survey Database was maintained protected and separated from the Personal Information Database. At the end of each "wave" of data entry, the identification codes of respondents were extracted and linked to the first database, where respondents were marked. The same database was also used to keep track of subjects explicitly refusing to participate, or those that deceased or emigrated after the sample was formed.

3.4. EXPOSURE ASSESSMENT

3.4.1. Dispersion modelling approach

In accordance with the aim of the study, which was to investigate the health effects specifically due to the refinery's emissions, only pollution from the refinery was considered. This was assessed by means of the AERMOD model, which allowed to estimate *in silico* how the point-source emissions from the refinery's stacks are dispersed over the study area and, thus, to predict the concentration of pollutants in the air matrix at ground level. A key vocabulary of dispersion modelling and a brief introduction to AERMOD are given in ***Appendix A.4***.

The dispersion model used in this study, which elaboration dates back to 2008, was implemented by the engineering consulting company *Snamprogetti SpA*. It was part of the procedures related to the Integrated Environmental Authorisation (*Autorizzazione Integrata Ambientale*, AIA) for the new EST facility, where it represented the *ante-operam* scenario (see ***1.5.4 The authorisation procedure for the EST facility***). Modelling assumed the emissions of the refinery at its maximum production capacity; the meteorological parameters needed for the implementation of the model (and, consequently, the predicted concentrations) were referred to the calendar year 2006. Concentrations predicted at the timepoints along the year were averaged to obtain the mean annual concentration of various contaminants (PM10, SO_2, NO_x[61]). The prudential choice of considering all particulate emissions as PM10 and all the nitrogen oxides emissions as NO_2 was adopted.

A preliminary analysis of the surface obtained by AERMOD showed that pairwise correlations between the predicted concentrations of different pollutants were extremely high both in Sannazzaro and Ferrera[62] (***Table 3.1***). For this reason, in order to avoid redundancies, particulate matter (PM10) was chosen as the representative pollutant.

[61] Actually, also CO concentrations were predicted, but the annual mean concentration was not computed for this pollutant.

[62] The surface included in AERMOD was wider than the area of the two municipalities involved in the present study.

Pollutant	PM10	SO_2	NO_x
PM10	1	-	-
SO_2	SdB:.992 FE: .995	1	-
NO_x	SdB: .994 Fe: .903	SdB: .976 FE: .933	1

Table 3.1 *– Pairwise correlations (Pearson's linear correlation coefficient) between couples of mean annual concentration of pollutants as predicted by the AERMOD model.* SdB*: Sannazzaro de' Burgondi.* FE*: Ferrera Erbognone.*

3.4.2. Individual exposure estimation

An individual estimate of PM10 exposure was assigned to each subject, depending on his or her home address, starting from the surface elaborated by AERMOD. An expert in environmental sciences collaborated in the related work.

The verified home addresses of all the enrolled subjects (see ***3.2.3.1 Regional Registry of Health Insurance***) were geocoded with an automated procedure based on the `R` package `ggmap`, in which map information was taken from Google Maps; coordinates (latitude and longitude) were determined in the Coordinate Reference System (CRS) EPSG:4326[63] (also known as the bi-dimensional geographic World Geodetic System 1984, or WGS84, because it is based on the homonymous ellipsoid[64]), or on its transformation EPSG:32632 (UTM[65] zone 32N, datum WGS84). All the geocodes were manually inspected to ensure for their consistency: automated geocoders can indeed return incorrect results, because they are usually developed to

[63] The basic geographical information for this CRS are reported at the following link: https://epsg.io/4326 (opened on July 26th, 2018). Official information about any CRS can be found at https://www.epsg-registry.org/ (opened on July 26th, 2018).

[64] An *ellipsoid* gives an approximated representation of the Earth by means of a regular rotation solid, characterized by its equatorial and polar semi-axes; specifically, the ellipsoids used to represent Earth are *oblate spheroids*, in which the equatorial semi-axes are of the same length and longer than the polar semi-axis. The orientation of the ellipsoid respect to the geoid (i.e. a representation of the Earth's profile, intended as the mean ocean level, based on gravitational fields) is called *datum*; WGS84 is a *global datum* because its ellipsoid is centred on the Earth's barycentre, so it can be used as a decent approximation of the entire Earth surface.

[65] *Universal Transverse Mercator* (UTM) is a conformal projection of WGS84 that uses Cartesian 2D coordinates (horizontal and vertical, in metres) to specify locations on Earth. Earth's surface is divided in 60 zones.

"interpret" the entry in order to correct input errors and accept different formats. The package ggmaps, anyway, includes also an automatic inspection system. Almost all the geocodes (1014, 97%) had a precision at the house-number level; the remaining were precise only at street level.

Individual exposure levels were assigned to each enrolled subject by interpolating the predicted PM10 concentration at each home address. Specifically, this was done by applying a bilinear interpolation (which gives the point estimate of concentration based on the 4 pixels located closest to the specified position) with the package raster in R.

Taking into account the spatial distribution of the enrolled subjects in the domain of modelled PM10 exposure, it was decided to consider clustered exposure levels as well. This allowed to reduce the selection bias due to the spatial distribution of exposure and to overcome the lack of linearity of PM10 when this variable had to be included in multivariate models. To do so, a non-hierarchical K-means model was applied in R (package cluster) on the full sample of all enrolled subjects (N=1046). The number of clusters in which the sample of enrolled subjects was subdivided was defined based on a sensitivity analysis, comparing different values of *k* (i.e. number of clusters): 3, 4, 5, 6, 7. These values were chosen to avoid categories with a very low number of subjects, thus allowing their use in multivariate models. The cluster of subjects with lower exposure was found to "fall" almost entirely in Ferrera Erbognone; the other clusters, representing increasingly high exposures, were in Sannazzaro de' Burgondi.

3.5. OTHER INDIVIDUAL-LEVEL VARIABLES

3.5.1. Questionnaire

A structured questionnaire, as already mentioned before in this Chapter, was sent to all the enrolled subjects as a part of the survey (see ***3.3.1 Data collection***). The questionnaire was designed to collect relevant information – including both factors that could influence the chance of being a case and potential confounders – at the individual level. Strictly personal information were avoided as much as possible, with the aim of reducing the chance of a participant being linked to the questionnaire he or she filled in[66]. The questionnaire was not meant to measure any latent trait, so it should be only intended as a tool for data collection.

The final version of the questionnaire is reproduced in ***Appendix C*** (in Italian). It consisted of 19 *Questions* (Q) (some including hierarchically nested sub-questions). Most of the answers were close-ended, either dichotomous, rating levels or multiple-choices; in some cases, the respondent was allowed to mark (and specify) an "*Other*" option. Open-ended questions were generally limited to indicating a number or to a very specific information (e.g. type of job in the table of job history). A summary of the Questions is reported below, grouped in different sub-parts (each one referred to a specific topic) as they were in the final layout chosen for the questionnaire.

A. *Introductory questions and checks:*
   - Day of compiling;
   - Gender (Male/Female);
   - Year of birth (4 digits);
   - Weight (in kgs);
   - Height (in cms).
B. *Sociodemographic information:*
   - *Q1* Marital status (either single, married, divorced, widowed, other);
   - *Q2* Education (maximum grade achieved);

[66] As explained with full detail in ***3.3.2 Data management***, only a pseudo-anonymous identification code was printed on the questionnaires as a way to prevent the identification of respondents.

- *Q3* Current job (employment status, main tasks, workplace, hours per day) and past jobs;
- *Q4* Actual home (yes/no for difference from the registered home address; if yes, same town or different town);
- *Q5* Daily time not spent at home for non-work-related reason.

C. *Housing:*

- *Q6* Type of building (multiple choice: flat, detached, etc.) and floor(s);
- *Q7* Type of nearby streets (multiple choice: main, lateral, park, etc.);
- *Q8* Traffic due to cars (4-levels frequency);
- *Q9* Traffic due to heavy vehicles (4-levels frequency).

D. *Lifestyle:*

- *Q10* Physical activities (5-levels frequency rating for different activities);
- *Q11* Alcohol consumption (if yes, 5-levels frequency rating for different alcoholic beverages;
- *Q12* Cigarette smoking (if yes in lifetime, age of beginning, number of cigarettes, current status of smoker and, if not current smoker, age at cessation);
- *Q13* Passive smoke (yes/no);
- *Q14* People smoking inside the house (if yes, number of cigarettes per each person);
- *Q15* Consumption of selected categories of locally produced food[67] (yes/no).

E. *Health:*

- *Q16* General self-perceived health status (5-levels rating);
- *Q17* Diagnosis for selected diseases and conditions (yes/no);
- *Q18* Therapeutic regimen for selected diseases and conditions (yes/no).

The last Question (Q19) asked the subject to self-rate on a 4-levels scale his or her accuracy and reliability in filling in the questionnaire with the required information. It is worth noticing that Q4 served as a check to exclude subjects living in a different place (especially if they were living in a different municipality): this was crucial because the assessment of individual exposure was based on registered home addresses (see ***3.4.2 Individual exposure estimation***).

[67] The reason for including this question were already discussed in ***1.6.2 Brought to the lunch table: pollutants and food chain***.

3.5.2. Survey variables used in the study

In the analyses of the present study, apart for demographic variables (*age* and *gender*), some additional variables were selected to be included in the analyses together with the main exposure of interest (PM10 concentration). The reason for that was to take into account relevant variables that could play a role with regard to the outcomes or could act as confounders of the relationship between main exposure and outcome. Those variables were obtained from the questionnaire (see ***3.5.1 Questionnaire*** and ***Appendix C***).

The main criterion for inclusion of a variable was an *a priori* reasoning about its meaning (i.e. semantic relevance) and, then, quality criteria based on a critical assessment of consistency and reliability of the information collected through the questionnaires (see ***3.6.2 Quality check of survey data***).

The following nominal (*) or dichotomous (#) variables were obtained from the questionnaire without any further recoding (reference categories are underlined):

- *Marital status* (Q1) * (single, married, divorced, widowed, other);
- *Education* (Q2) * (primary school, secondary school, high school, university, other)[68];
- *Type of house* (Q6) (detached house – isolated, detached house – close to other, semi-detached, flat, terraced house, other);
- *Alcohol consumption* (Q11) # (yes/no);
- *Cigarette smoking*, lifetime (Q12) # (yes/no);
- *Cigarette smoking*, current (Q12.2) # (yes/no)[69].

Other questionnaire items were recoded before statistical analyses. This step was needed mainly because recoding allowed to maximise information from the respondents, by reducing missingness and grouping categories with very low frequencies. However, all the variables were recoded only if the proposed recoding was reasonable, i.e. the information were complementary in the way they were put together.

[68] This variable could be considered ordinal, but in multivariate analysis it was included as nominal in order to contrast each level to the reference category (see ***3.6.5 Main multivariate analyses***).

[69] Given that Q12.2 was nested in Q12, those who declared they never smoked in their life (i.e. "No" in Q12) were considered as "No" also for current smoking (Q12.2).

In detail:

- *Traffic in the streets near subjects' houses*, defined as a dichotomous variable in which "high" (i.e. exposure) was assigned if the subject answered that traffic was "constantly" or "frequently" to either Q8 or Q9, "low" otherwise, and "missing" only if both Q8 and Q9 were missing (reference category: "<u>low</u>");
- *Physical activity*, defined as a dichotomous variable in which "at least one" was assigned if the subject declared to practise at least one of the activities indicated in Q10 and at least once or twice per week, "none" otherwise, and "missing" only if all the activities in Q10 were missing (reference category: "<u>none</u>");
- *Other diseases*, defined from Q17 and Q18 as a dichotomous variable in which "yes" was assigned if the subject declared to be affected by a condition that could determine the status of "case" if it resulted in a hospital admission[70], i.e. arrhythmia, hypertension, asthma, COPD, diseases of the digestive system ("yes" in at least one out of Q17.1, Q17.2, Q17.4, Q17.5, Q17.6), or to be regularly administered a treatment for such conditions ("yes" in at least one out of Q18.1, Q18.2, Q18.4, Q18.5, Q18.6), "no" otherwise, and "missing" only if all the answers in Q17 and Q18 were missing (reference category: "<u>no</u>").

Gender and year of birth (included in all analyses as *age*, referred to 2014 because this was the end of the timespan in which the health outcomes were observed and recorded) were taken from the database of personal information, so that they were available for all subjects, including those who did not indicate these data while filling in the questionnaire. Body Mass Index (BMI) was calculated from height and weight indicated in the questionnaire, according to the formula $BMI = weight\ in\ kgs/(height\ in\ mts)^2$.

[70] This was defined on the basis of ICD-IX-CM codes (see ***3.1.1 Case definition***) of such conditions.

3.6. STATISTICAL ANALYSES

The statistical analyses of data from the study presented in this doctoral thesis consisted of different steps, each one with a different purpose (preliminary analyses, description of data, univariate and multivariate analyses) and regarding different populations (either all the enrolled sample or only the sub-sample of respondents). For this reason, each step is detailed separately. All the analyses were performed in Stata 13 [*STATACORP, 2013*] and, in inferential testing, the threshold for statistical significance was fixed at 5% ($\alpha = .05$). In case of multiple testing, the significance level was adjusted by using Bonferroni's method, i.e. dividing the probability of a Type-I error for the k comparisons ($\alpha' = \alpha/k$).

3.6.1. Evaluation of non-respondence bias

In surveys, a *non-respondence bias* occurs when respondents are systematically different from non-respondents with regards to certain characteristics. In an epidemiological context, this results in a *selection bias* due to non-comparability of groups under comparison, which might affect the estimates of interest in a study. This could become particularly critical if the difference between respondents and non-respondents depends on factors that might be related to the effects under investigation.

To make sure that the survey was not affected by non-respondence bias, a preliminary statistical analysis was carried out. This analysis considered only data from the Personal Information Database (see ***3.3.2 Data management***), as they were the only data available for both respondents and non-respondents.

At first, descriptive statistics were produced. Nominal variables (i.e. *respondence*, *municipality*, *status* of case or control, *gender*) were described as absolute and percentage frequencies. There was only a quantitative variable (*age*, referred to year 2014) in this preliminary analysis, and it was represented with its mean and standard deviation (SD) and median and inter-quartile range (IQR). The distribution of the quantitative variable was also inspected graphically.

Association between *respondence* and *municipality* was investigated by using a Pearson's Chi-squared test for independence or Fisher's exact test, as appropriate (***Appendix B.3***). Then, Pearson's test (or Fisher's test) was applied also to test the association of *respondence* with *status* and with *gender*. To test for differences regarding *age* between respondents and non-respondents, Student's t test for unmatched groups or Mann-Whitney's (MW) U test were applied as appropriate (***Appendix B.2***). Differences in *age* and *gender* were also tested within *municipality* and *status*. Among cases only, association of *respondence* with the ICD Chapters was also tested by means of Pearson's test (or Fisher test).

3.6.2. Quality check of survey data

In order to assess reliability of the information provided by study participants, all the data collected in the survey underwent a quality check.

The first step of quality assessment was counting the number of missing answers per each question or sub-question; those with a high rate of missingness (over 20% of respondents[71]) were carefully evaluated. Then, for hierarchically-ordered questions, consistency between higher-level questions and their dependencies was assessed. Inconsistencies were evaluated one by one and, when possible (i.e. if the compilation mistake was straightforward), the answer to the higher-level question was manually corrected: for instance, if someone stated that he or she was not drinking any alcohol (yes/no question), but declared to drink wine 2-3 times per week, then the first answer was corrected from “No” to “Yes”.

Particular attention was posed to those situations in which an answer to a sub-question was given, whilst no answer was expected based on the higher-level answer, or on the opposite case (answer to the sub-question was missing while it was expected).

When a question required a number as answer, and the number was unconfutably expected to fall within a certain range (e.g. hours spent outside in a day), consistency of the answer with general limits was evaluated (e.g. answer not above 24 hours). In any case, questionable answers (e.g. more than 16 hours spent outside daily) were identified and, in “suspicious” cases, their correct recording in the electronic database was checked by looking back at the paper questionnaires. Answers regarding the hours

[71] It is worth recalling that some questions were hierarchically nested; in that case, the base for calculating the rate of missingness in sub-questions was the number of expected answer based on the higher-level question.

typically spent out of home for work-related reasons (sub-question Q3.1) and for non-work-related reasons (Q5) were also summed, and the same checks were carried out on the total hours spent out of home.

3.6.3. Descriptive analysis

Appropriate statistical measures were used to describe outcome data and exposure[72] (regarding both the sample of all the enrolled subjects and the sub-sample of respondents), as well as survey data once quality checking upheld for their consistency.

Nominal variables were represented by their distributions of absolute and percentage frequencies; for ordinal variables, also cumulative distributions were computed. Quantitative variables were described in terms of mean value and SD and by their median value and IQR.

When a graphical representation of data was needed, qualitative variables were represented in bar charts or pie charts and quantitative continuous variables by histograms or box plots.

3.6.4. Univariate analyses

Univariate analyses were aimed at estimating crude associations between the *status* of case or control (dichotomous variable) and the other variables of interest. For nominal variables, Pearson's Chi-squared test or Fisher's exact test, as appropriate, were used (see ***Appendix B.3***), while association against continuous variables were tested by means of Student's t test or Mann-Whitney's U test[73] (see ***Appendix B.2***).

Variables from survey data were used as detailed previously (see **3.5.2 *Survey variables used in the study***); in any case, also the remaining Questions from the survey were described. Environmental exposure (*PM10 concentration*) was used both as a continuous variable and, clustered, as a nominal variable (albeit it could be considered an ordinal one, in multivariate analysis it was treated as a nominal variable, thus we applied the same criterion in univariate analyses). For environmental exposure,

[72] In descriptive statistics, PM10 concentration was used both as a continuous variable and as clusters of concentration (treated as ordinal variable).

[73] As already stated, p-values for Mann-Whitney's test were computed by using the standard approximation.

given that it is the main exposure of interest, crude ORs (see ***Appendix B.4***) were calculated, together with their confidence interval at the 95% confidence level (95%CI) and the test for homogeneity of odds (i.e. testing as null hypothesis the absence of effect, OR=1).

The association of the main exposure (PM10 concentration, clustered) with the variables obtained from survey data was also tested (Pearson's Chi-squared test or Fisher's exact test, as appropriate; only for age: Student's t unpaired test or Mann-Whitney's U test and one-way analysis of variance for unmatched measures or Kruskal-Wallis test, KW). This had the aim of finding potential confounders.

3.6.5. Main multivariate analyses

Broadly speaking, multivariate statistical modelling is aimed at estimating the net effect of a series of exposures (independent variables) on an outcome (dependent variable)[74]. In the case of the present study, unconditional logistic regression (see ***Appendix B.5***) was applied for the sake of estimating the effect of the environmental exposure on the odds of being a case or a control, controlling for other relevant covariates. The model also allowed to estimate the net effect of the covariates on the outcome (again, for a definition of the variables see ***3.5.2 Survey variables used in the study***).

The environmental exposure, i.e. the main exposure of interest, was included in the model with its clustered variable[75]. It was treated as a merely nominal variable, because this choice made it possible to estimate the effect of each high-exposure cluster against the lowest-exposure one, taken as reference.

The effect of the main exposure was always adjusted for *age*, *gender*, and *lifetime cigarette smoking*[76] (Q12), as previously defined. These variables were forced to be

[74] The correct name would be "*multivariable* analyses" but, in spite of that, this kind of analyses are broadly reported as "multivariate analysis".

[75] As will become clearer in ***4.3.3 Environmental exposure***, using PM10 concentration as a continuous variable would have been critical because of issues regarding the variable's distribution.

[76] *Lifetime cigarette smoking* was preferred against *current cigarette smoking* because the exposure is referred to the past (outcomes were indeed considered until 2014, while current smoking refers to 2016 or 2017). Using *current smoking* would have required to take into account the age of cessation, in order to understand if past smokers had to be considered "smokers" as of 2014. As will become clear in ***4.2.3 Lifestyle***, age at cessation did not prove to be reliable enough. Moreover, *lifetime smoking* ensures that all the timespan under investigation is implicitly accounted for.

retained in the model, depending on *a priori* decisions[77]. *Age* was introduced as a continuous variable (unit increment: 1 year). *Gender* and *current cigarette smoking* were both dichotomous variables; as reference categories, male gender and no smoking were chosen.

The possible inclusion of other variables as covariates in the model was evaluated:

- *Marital status* (nominal variable, reference category: single);
- *Education* (nominal variable, reference category: primary schooling);
- *Type of house* (nominal variable, reference category: flat);
- *Traffic in the streets near subject's house* (dichotomous variable, reference category: low traffic);
- *Physical activity* (dichotomous variable, reference category: no physical activity);
- *Alcohol consumption* (dichotomous variable, reference category: no drinking);
- *Other diseases* (dichotomous variable, reference category: no);
- *BMI* (continuous variable, unit increment: 1 kg/m^2).

Art first, a simple model, accounting for the environmental exposure alone, was estimated[78]; then, *age*, *gender* and *cigarette smoking* were added. The other potential covariates were added one at a time to the latter model, and their role was carefully assessed. In detail, various criteria were taken into account to decide if whether a covariate had to be retained in the model or not. Firstly, informativity was evaluated with the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) (see ***Appendix B.5.2***). Then, it was evaluated if the variable significantly contributed to the model by applying a Likelihood-Ratio (LR) test (see ***Appendix B.5.3***). Briefly, this test compares the model with the independent variable(s) to a model without the independent variable(s); thus, it tells if when a variable is added to the model it gives a contribution by explaining more of the observed variability. The net effect of the covariate was investigated by using Wald's test (WT) (see ***Appendix B.5.4***), which says if the estimated effect is significantly different from a null effect.

[77] *Municipality* was not included because, as it will be detailed in ***4.3.3 Environmental exposure***, it was disclosed to be strongly associated with the main exposure.

[78] The effect estimates (ORs) from this model were obviously expected to be equal to crude ORs.

Finally, a model with the retained covariates was run, and those covariates were dropped one by one to confirm their fit in the model with the criteria reported above.

Although the evidence supporting the relation between diseases of the digestive system and air pollution is mounting, it was deemed to be somehow less strongly ascertained. For this reason, a sensitivity analysis was performed by excluding those subjects that became cases because of causes related to the digestive system (ICD-IX-CM Chapter 9). Multivariate models were estimated on this sub-sample, and the results were compared with those obtained from the full sample of respondents.

3.6.6. Secondary multivariate analyses

A secondary multivariate model was developed to investigate the effect of several variables on the self-perceived general health status (Q16). The variable, which had originally a Likert-like structure with 5 levels ("poor", "passable", "good", "very good", "excellent"), was dichotomised by grouping together the answers described with a negative adjective ("poor" and "passable") with the other answers, described by more positive terms (from "good" to "excellent")[79]; the "positive" category was taken as reference for the computation of ORs. A logistic model was applied, and the same criteria already described for the main analyses were adopted.

To assess if environmental exposure had a role on perceived health, the *distance* between subjects' houses and the refinery's centroid was used. The choice of this variable over exposure to PM10 was made because it was believed to be closer to residents' "immediate" perception. The effect of environmental exposure was always controlled for *age* and *gender* and also for *case/control status*, because suffering from a condition that led to a hospital admission could clearly influence one's perception about his or her health. *Traffic* (Q8 and Q9, combined), *current cigarette smoking* (Q12.2)[80] and *physical activity* (Q10, recoded) were tested as additional variables.

[79] As will become clear in the Results (see ***4.4.4 Health***), recoding the variable was necessary because the distribution of frequency of the original levels was too unbalanced, with almost 80% of the answers concentrated in only two levels ("passable" and "good").

[80] In this case, *current cigarette smoking* (Q12.2) was used, coherently with the fact that the outcome was referred to the time of assessment.

3.7. Ethics and conflicts of interest

A first version of the research protocol of the CONSAL Project, including the part regarding Study I, was approved by the Ethical Committee of the University of Pavia on June 13th, 2013. An amended version was later approved in June 2016 (after a reorganisation, the competence for the approval was moved to the Ethical Committee of *IRCCS Policlinico San Matteo*). In this Chapter, the various issues related to study planning have been discussed, and their practical implementation has been commented.

As stated in the protocol, the CONSAL Project was sponsored by ENI S.p.A.[81], but it is worth saying that the scientific responsibility of the research and its results uniquely belongs to the Department of Public Health, Experimental and Forensic Medicine of the University of Pavia.

[81] It is worth recalling that, as previously explained (see ***1.5.4 The authorisation procedure for the EST facility***), the company was compelled to bear the costs of the epidemiological investigation. A nice comment concerning ethical perspective on funding and competing interests in epidemiological research projects can be found in the paper by *Soskolne* [*2016*].

4. Results

4.1. Analysis of respondence

There were 563 survey respondents out of the 1046 enrolled subjects, as defined previously in ***3.2.4 Sample selection***. Out of the 483 non-respondents, 15 were actually found to not meet eligibility criteria and had to be dropped: 2 (1 case, 1 control) died after enrolment, 3 (2 cases, 1 controls) were not available at their registered home address, 9 (3 cases, 6 controls) emigrated, and 1 (a control) because of homelessness. Thus, the "real" final enrolled sample counted 1031 of the total 1046 enrolled[82]. The overall response rate[83] was 54.6%. Among the 468 who not participated in the study, 15 (3.2%) explicitly refused to give their consent. Respondence data were analysed as defined in ***3.6.1 Evaluation of non-respondence bias***.

The response rate was significantly different by the municipality in which the subjects were registered as residents, being 75.7% in Ferrera Erbognone and 49.1% in Sannazzaro de' Burgondi (χ^2=48.48, p<.001) as represented in ***Figure 4.1***.

However, overall respondence was comparable between cases and controls (χ^2=0.20, p=.66), as reported in ***Figure 4.2***; the rate of respondents by case/control status was not different either in Sannazzaro (χ^2=0.09, p=.76) or in Ferrera (χ^2=1.03, p=.31). Analogously, gender was not associated with respondence (χ^2=1.29, p=.26), which was found to be 53.1% among males and 56.6% among females (***Figure 4.3***). No difference was observed within municipality: in Sannazzaro, respondents were 48.7% among males and 49.7% among females (χ^2=0.09, p=.77), and in Ferrera the shares were respectively 73.8% and 77.5% (χ^2=0.40, p=.53). Mean age[84] was significantly different between respondents and non-respondents (t=-3.44, p=.0006) but this was of scarce practical relevance, with respondents (57.1 years) being roughly 3 years older than non-respondents (54.4 years). The difference was greater in the sub-sample of Ferrera (t=-2.17, p=0.03) than it was in Sannazzaro (t=-3.50, p=.0005), as shown in ***Figure 4.4***.

[82] The corrected total number (N=1031) will be used throughout this section

[83] This rate (as well as the followings) is computed on N=1031, unless differently specified.

[84] As stated in ***3.5.2 Survey variables used in the study***, age was calculated as of 2014.

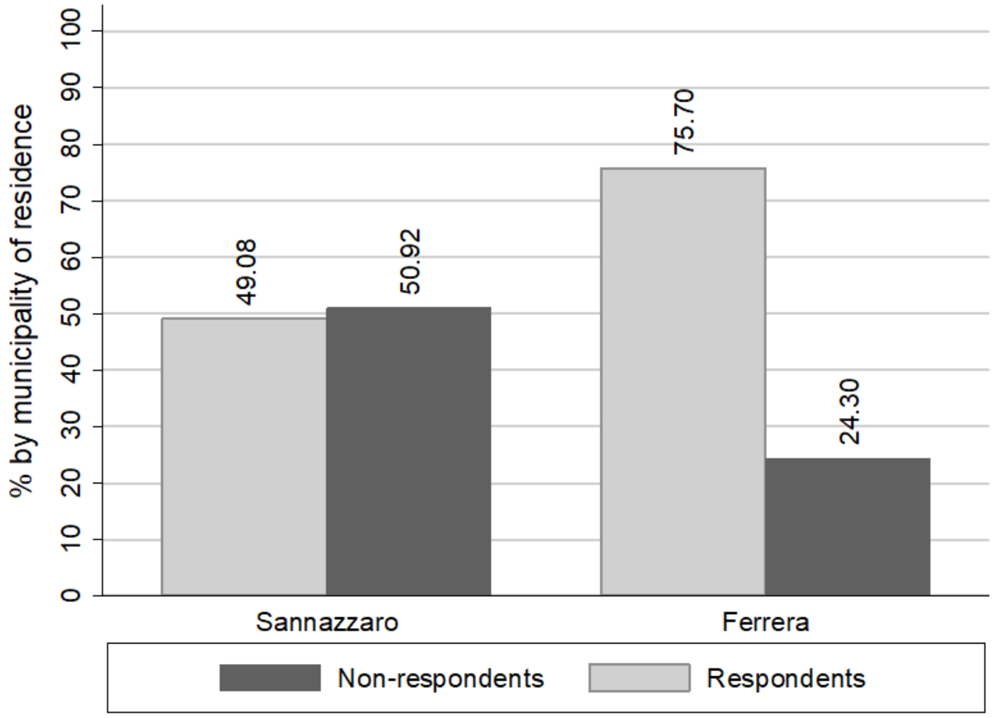

***Figure 4.1** – Response rates among residents in Sannazzaro de' Burgondi and Ferrera Erbognone.*

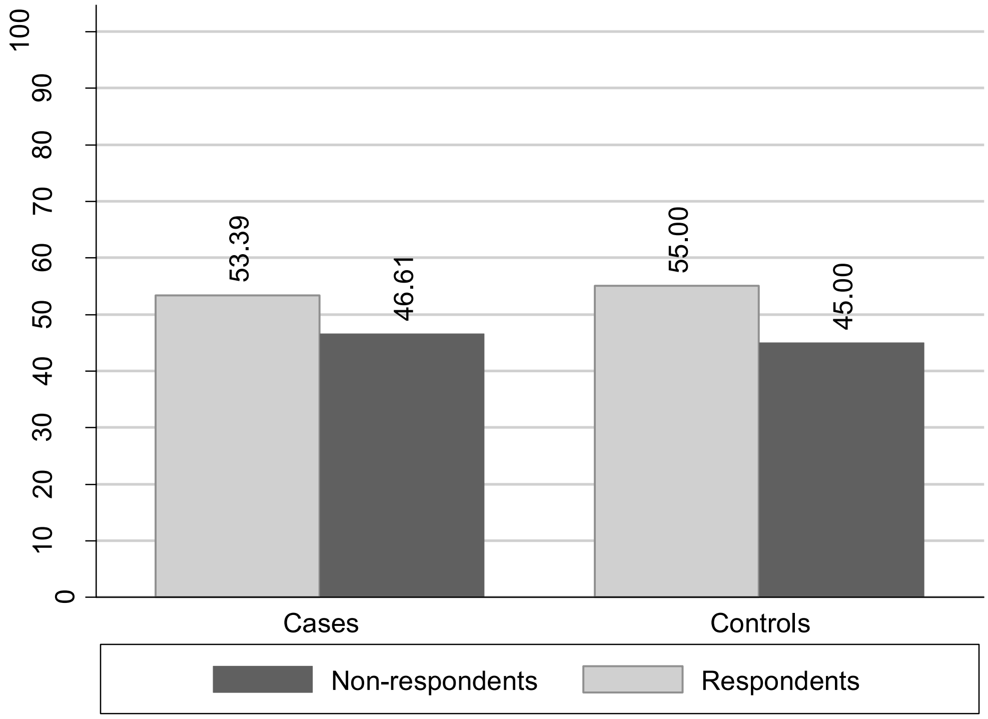

***Figure 4.2** – Response rates among cases and controls.*

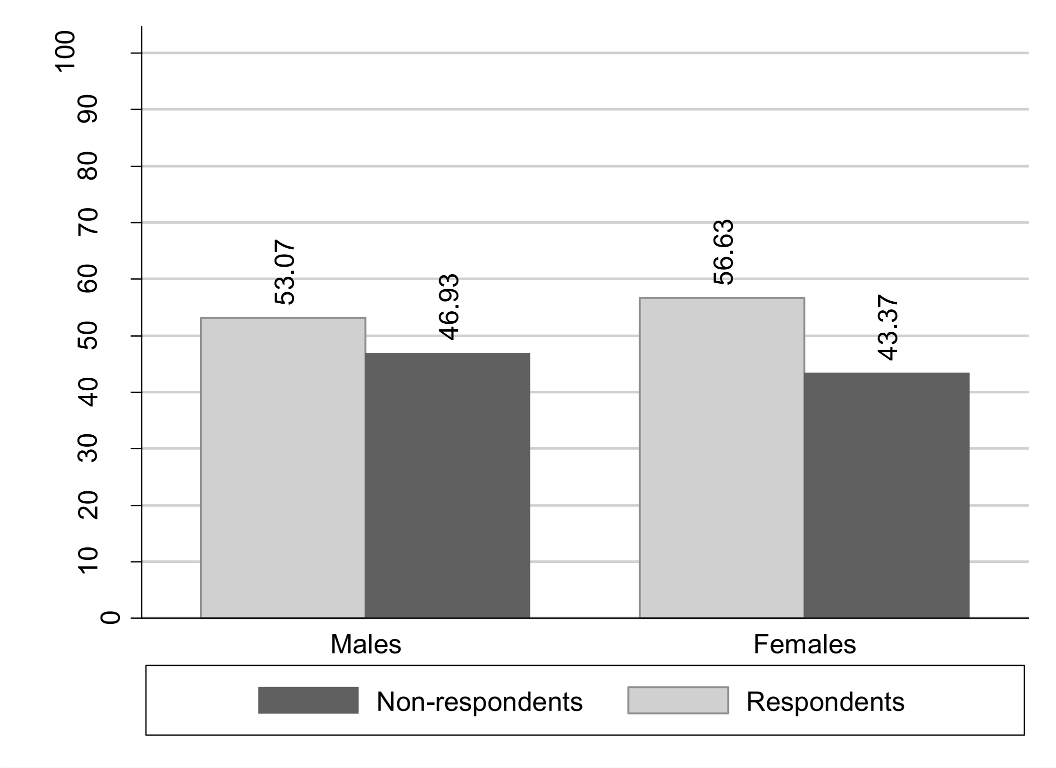

***Figure 4.3** – Response rates among males and females.*

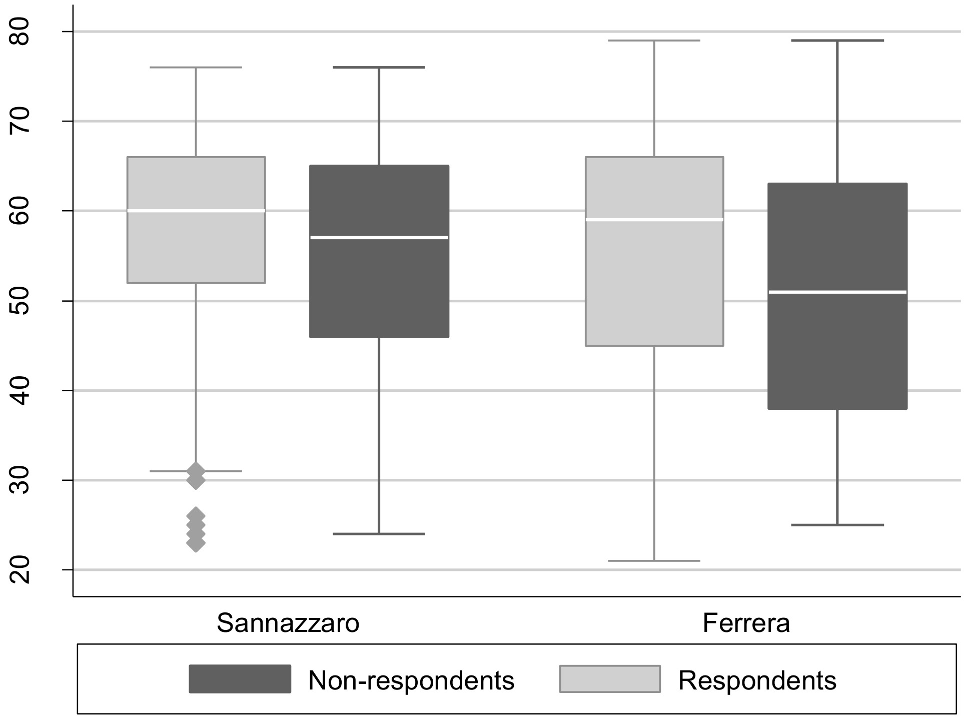

***Figure 4.4** – Age of respondents and non-respondents in Sannazzaro and Ferrera (line: median; box: 25th-75th centiles).*

Among cases only, no difference was found between respondents and non-respondents with regards to the distribution of ICD-IX-CM-coded Chapters of the main cause of hospitalisation, as recorded in the SDO database (p=.66). Data are reported in ***Table 4.1***.

Respondence	**ICD-IX-CM Chapter** *n (%)*			
	Chap. 7	***Chap. 8***	***Chap. 9***	***Chap. 16***
Respondents	49 (44.1%)	20 (48.8%)	46 (48.4%)	6 (60.0%)
Non-respondents	62 (55.9%)	21 (51.2%)	49 (51.6%)	4 (40.0%)
Total	111 (100%)	41 (100%)	95 (100%)	10 (100%)

Table 4.1 *– Distribution of respondence by main diagnosis (ICD-IX-CM chapters) of the first hospitalisation of cases.*

4.2. CONSISTENCY OF SURVEY DATA

Among the 563 survey respondents, the preliminary analysis of the consistency of data collected though the questionnaire was assessed as detailed previously in ***3.6.2 Quality check of survey data***.

First of all, the question in which respondents were asked to self-rate the quality and consistency of their questionnaire (Q19) was answered by 548 participants (97.3%). Essentially all of them rated their compiling as either "Good" (204, 37.2%) or "Very Good" (342, 62.4%), except for two who self-rated their accuracy as "Poor".

4.2.1. Socio-demographic characteristics

Information regarding marital status (Q1), education (Q2) and employment status (Q3) were provided by all or almost all respondents. Question Q3 had some nested sub-questions, asking about how many hours the participant used to spent out of home for his or her work (Q3.1), the description of the tasks (Q3.2), the town (Q3.3a) and address (Q3.3b) of the workplace. After excluding those who were identified as non-workers in Q3 (unable to work, unemployed, housewife, retired), 224 were expected to answer to these sub-questions; in Q3.1 and Q3.2, 205 (91.5%) did it, while the town of the workplace was provided by 206 (92.0%) and the exact address by 173 only (77.2%). It should be noted that, with regards to the hours spent out of home (Q3.1), some unrealistic answers were observed: 1 subject declared 16 hours, one 24 hours, and one even exceeded the number of hours in a day (40 hours). Job history (Q3.4), albeit the question was potentially applicable to all respondents, was reported by a minority of subjects: roughly 150, out of 563 who indicated at least one past profession.

The participants answering if their actual address was different from their registered home address (Q4) were 539 (95.7%); of them, 34 were expected to answer to the nested sub-question (Q4.1), asking if their actual address was in the same municipality where they were registered. A few discrepancies were observed: 2 subjects answered to the latter question without answering to the former one; one other subject, on the contrary, was expected to answer to the sub-question while no answer was given. These discrepancies were deemed to be of minor relevance, and the sub-question was considered to prevail on the higher-level question.

In Q5, i.e. the number of hours spent outside during an average day for non-work-related reasons, 427 answers (75.8%) were given. Like in Q3.1, some answers were unrealistic, again with one subject declaring 16 hours, one other 24 hours, and another even 40 hours. Consistency of Q5 was also inspected in combination with Q3.1, by looking at their sum – i.e. the total hours spent out of home for any reason – and it was found to be 16 hours or above in 35 participants, among which 4 exceeded the limit of 24 hours per day.

4.2.2. Housing

Almost all the respondents (555, 98.6%) indicated the type of house they used to live in (Q6). Regarding the sub-questions nested in Q6, 112 subjects declared their house to be a flat in a block-building, being consequently expected to answer Q6.1 (floor at which the flat is located); actually, an answer to this sub-question was provided by 117. The 443 who declared to live in any other type of house but *flat* were expected to answer to Q6.2 (total number of floors of the building), but only 321 of them (72.5%) did it.

The question about the type of street(s) near the house (Q7) was answered by 537 participants (95.4%); missingness was slightly higher in the specific questions about traffic levels due to car (Q8) and trucks (Q9), where respondents were 91.5% and 89.9% respectively.

4.2.3. Lifestyle

In the Questions regarding physical activities (Q10.1 – Q10.5[85]), a high number of missing answers was recorded, with each sub-question's missing rate falling in the range 19-47%. It is worth noticing that the number of subjects not answering to any of Q10's sub-questions was as low as 51 (9.1%), but only 280 (49.7%) compiled all the sub-questions.

For what concerns alcohol consumption, this information (Q11) was provided by almost all the participants: it was missing only in 6. Anyway, a certain number of subjects answered "No" in the general question regarding the drinking of alcoholic

[85] There was actually one more sub-question (Q10.6), namely "Other activities" for which the respondent was asked to specify the activity and rate the frequency, but only a few completed properly this sub-question.

beverages, but then compiled at least one of the nested sub-questions, making it clear that, at least occasionally, they actually consumed a certain type of drink (e.g. beer or wine). In such cases, where a mistake made by the respondent in Q11 was unconfutable, the answer to the higher-level question was manually reassigned for the sake of coherence. Regarding the sub-questions (Q11.1-Q11.4[86]), 252 subjects were expected to fill in the answers; the share of missing data lied in the range 6-48%, and roughly half of those subjects (128, 50.8%) provided all the answers.

A similar case was for the question regarding lifetime cigarette smoking (Q12): the number of missing answers was negligible (5) but, as already seen in Q11, several inconsistencies were observed in the nested questions. The answer to Q12 was manually reassigned when it was straightforward how to solve the inconsistency; for instance, if the participant answered "No" to Q12 (i.e. never smoked in their life) but then indicated either the age when they began to smoke, or the number of cigarettes smoked. The age when the subject started to smoke (sub-question Q12.1) was provided by almost all the subjects previously stating they have been smokers during at least part of their life (only 6 missing answers among smokers). Among those who stated to have been smokers in their life, 193 declared they quit smoking (sub-question Q12.2); this number is slightly less than the 201 participants who indicated the age when they ceased this habit (sub-question Q12.3), thus highlighting a potential inconsistency in the answers; specifically, 11 respondents gave an age at cessation without answering to Q12.2, and 1 gave an age at cessation despite identifying as a current smoker in Q12.2. The age at cessation was not considered for this last subject. The number of smokers providing the information about the number of cigarettes smoked (sub-question Q12.4) was 301, i.e. 95.6% of the expected answers.

The question regarding exposure to passive smoking (Q13) was answered by 476 participants (84.6%). The information on the number of people smoking inside the house (Q14) was given by 415 (73.7%); of them, 119 were expected to fill in the sub-question about the number of cigarettes smoked inside the house by each person (Q14.1-Q14.4), and 110 indicated the requested information for at least one. Anyway, it should be mentioned that, for 18 of the 119 respondents, the number of persons

[86] Also in this case, there was actually one more sub-question (Q11.5) for "Other": only 12 subjects gave a frequency rating for other types of beverages, but none of them indicated the type of drink.

indicated in Q14 was not consistent with the number of persons filled in in Q14.1-Q14.4.

4.2.4. Health

Almost all the respondents indicated their self-perceived general health status (Q16): only 14 answers were missing. Concerning the questions about diagnoses (Q17) or therapies (Q18) for specific conditions, the missingness rate is moderate or high, falling in the range 12-27%. Thirty-five subjects (6.2%) did not give answer to any of the sub-questions regarding diagnosed conditions (Q17.1-Q17.7), while 394 (70.0%) answered all the sub-questions. For the sub-questions regarding the administration of a therapeutic regimen (Q18.1-Q18.7), 58 (10.3%) did not answered any of the sub-questions and 381 (67.7%) answered all. Besides, 121 participants filled in the "Other" option regarding diagnoses (Q17.8) and 84 the one regarding therapies (Q18.8).

4.3. CHARACTERISTICS OF RESPONDENTS

The analyses presented here were carried out following the methodologies detailed in the sections ***3.6.3 Descriptive analysis*** and ***3.6.4 Univariate analyses*** of the previous Chapter. The 22 subjects who declared to live in a different town than the one where they were registered (Q4.1) were excluded from the analyses. Thus, a sub-sample of 541 enrolled subjects who participated in the survey was finally included in the analyses.

4.3.1. Health outcome

Among the 541 enrolled subjects who accepted to participate in the survey, 416 were controls and 125 cases. The distribution of residents in the two municipalities among cases and controls is reported in ***Table 4.2*** and ***Figure 4.5***: among cases, the share of subjects from Sannazzaro de' Burgondi seems higher than in controls, but the difference is statistically borderline (χ^2=3.76, p=.052).

With regards to cases only, a description of the causes of hospitalisation, as coded under ICD-IX-CM in the SDO records, is reported in ***Appendix D***.

Municipality	**Cases** *n (%)*	**Controls** *n (%)*	**Overall** *n (%)*
Sannazzaro	98 (78.4%)	289 (69.5%)	387 (71.5%)
Ferrera	27 (21.6%)	127 (30.5%)	154 (28.5%)
Total	125 (100%)	416 (100%)	541 (100%)

Table 4.2 *– Distribution of municipality by case/control status. The distribution in the overall sample is also reported.*

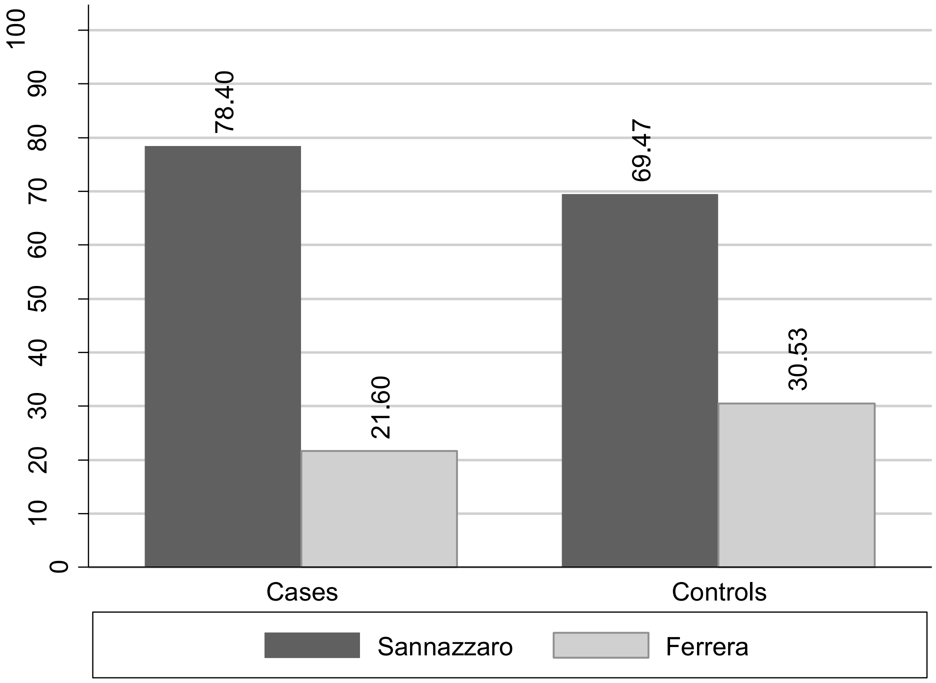

Figure 4.5 *– Graphical representation of the distribution of municipality by status.*

4.3.2. Demographics

The distribution of gender (***Table 4.3*** and ***Figure 4.6***) was similar in cases and controls (χ^2=0.92, p=.34), even though the percentage of males was slightly higher among cases. Analogously, age (***Figure 4.7***) was not significantly different between groups (t=-1.08, p=.28): on average, cases were 58.1±1.1 years old while controls were 56.8±0.6 years old.

Gender	**Cases** *n (%)*	**Controls** *n (%)*	**Overall** *n (%)*
Males	74 (59.2%)	226 (54.3%)	300 (55.5%)
Females	51 (40.8%)	190 (45.7%)	241 (44.5%)
Total	125 (100%)	416 (100%)	541 (100%)

Table 4.3 *– Distribution of gender by case/control status. The distribution in the overall sample is also reported.*

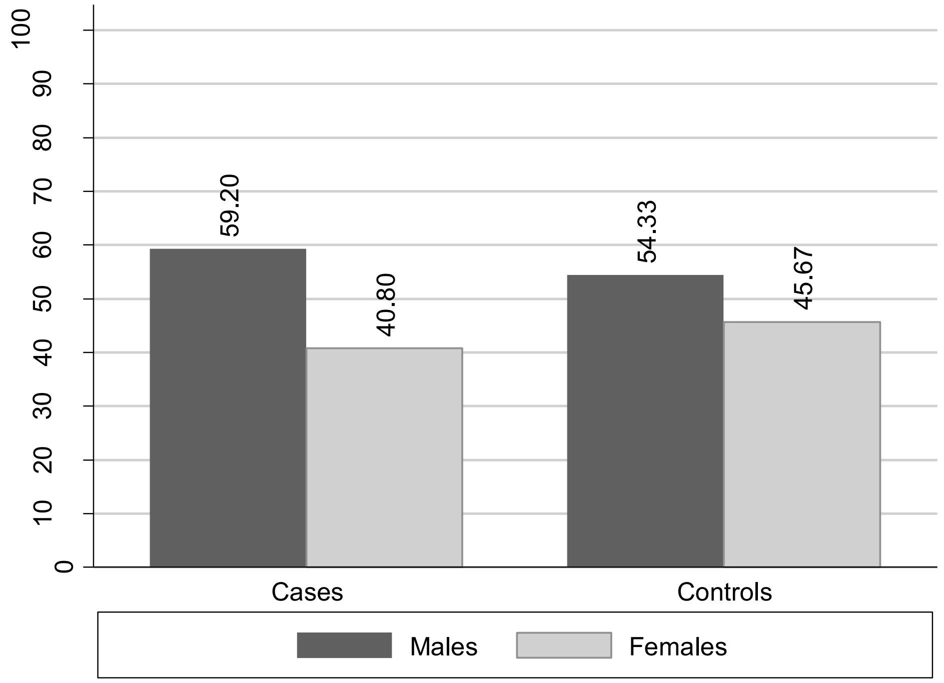

***Figure 4.6** – Distribution of gender, separately among cases and controls.*

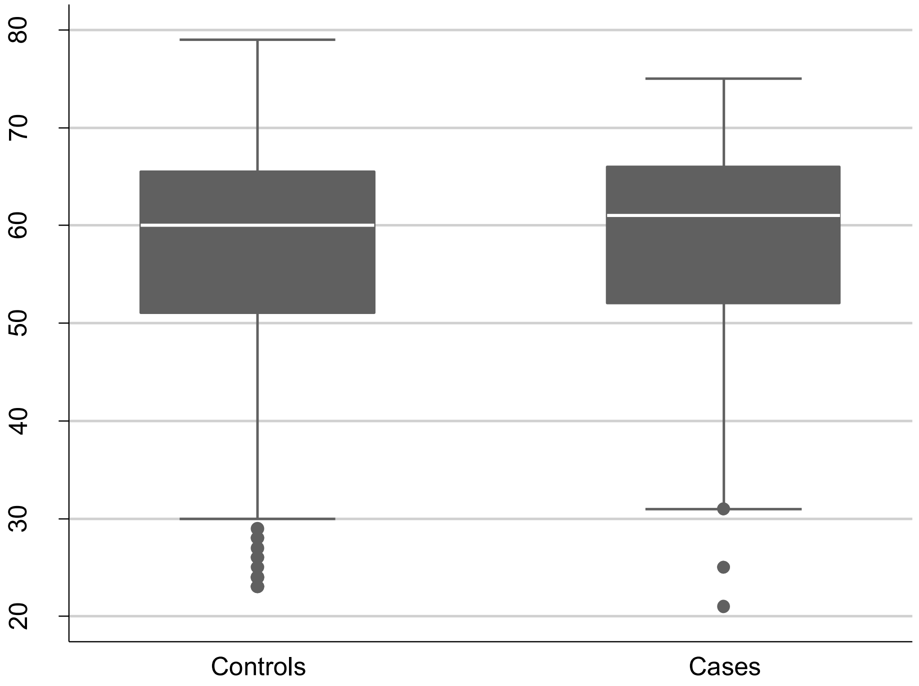

***Figure 4.7** – Description of age, referred to the calendar year 2014, in cases and controls. The white line represents the median; the box goes from the 25th to the 75th centile.*

4.3.3. Environmental exposure

The home addresses of all enrolled subjects were geocoded as detailed previously (see ***3.4.2 Individual exposure estimation***); those of the 541 who participated in the survey (and were not domiciled outside the municipality where they were registered) are reported in ***Figure 4.8***.

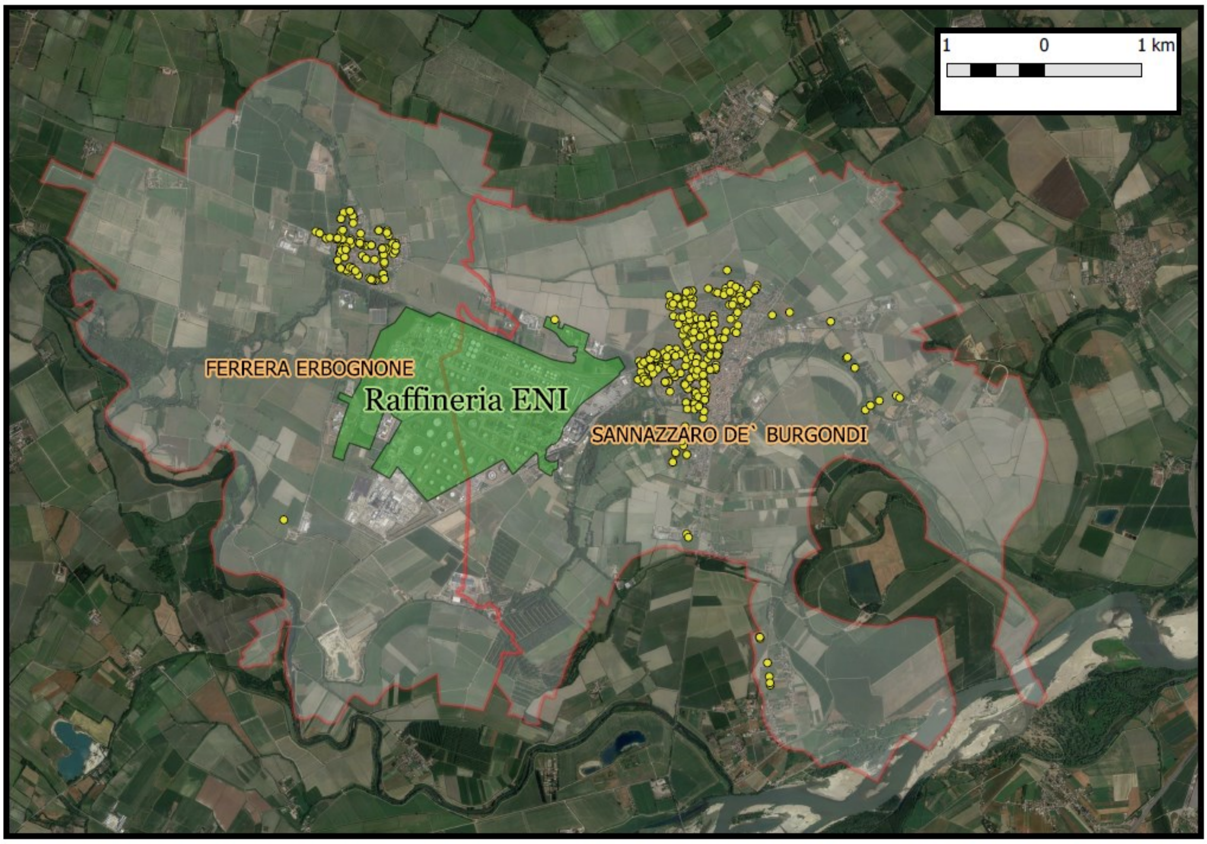

Figure 4.8 *– Map of the registered home addresses (yellow circles) of the 541 study participants considered in the sub-sample. The borders of the two municipalities of Sannazzaro de' Burgondi and Ferrera Erbognone are represented, together with the area of the refinery. Background: Google Maps, satellite view.*

4.3.3.1. Distance from the refinery

On average, distance of subjects' homes from the centroid of the refinery was 2593±662 metres (median 2707 metres, IQR 967 metres); the distribution of the variable is showed in ***Figure 4.9*** and it appears bimodal. The average distance is different in cases and controls, being greater among cases, and also by municipality (higher among residents in Sannazzaro), as reported in ***Table 4.4***.

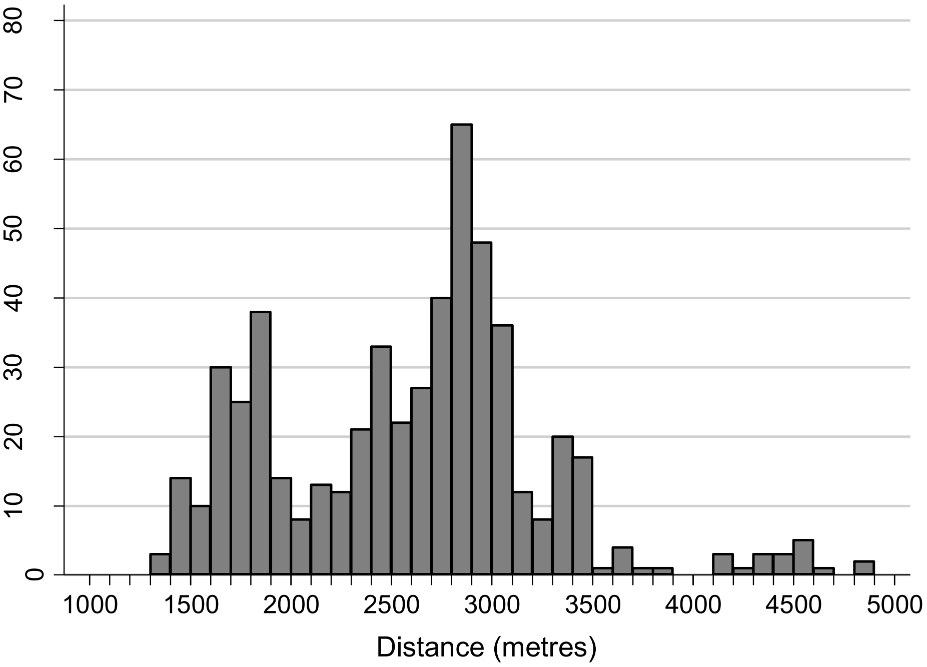

Figure 4.9 *– Distribution of distances calculated from the geocoded home addresses of survey participants to the centroid of the refinery.*

Group	**N**	**Summary statistics** *(metres)*			**Test p-value**
		Mean (SD)	*Median (IQR)*	*Min - Max*	
Cases	125	2738 (709)	2838 (698)	1383 – 4884	MW=-2.537 p=.011
Controls	416	2550 (642)	2689 (1004)	1383 - 4850	
Sannazzaro	387	2920 (471)	2878 (412)	1628 - 4884	MW= 18.09 p<.0001
Ferrera	154	1772 (199)	1790 (235)	1383 - 2160	

Table 4.4 *– Description of distances from the geocoded home address to the centroid of the refinery by case/control status and by municipality.*

4.3.3.2. Continuous PM10 concentration

The concentrations of PM10[87] attributed to the refinery, as predicted by the AERMOD model, are distributed in a bimodal fashion as can be clearly seen in the distributional plot (***Figure 4.10***). Summary statistics of this variable, by status and municipality, are reported in ***Table 4.5***. Whilst exposure to PM10 seems slightly higher in controls

[87] PM10 concentration, continuous variable.

rather than in cases in both municipalities, no significant difference is detected between cases and controls, neither stratifying by municipality nor when an overall test is performed (MW=-0.82, p=.41).

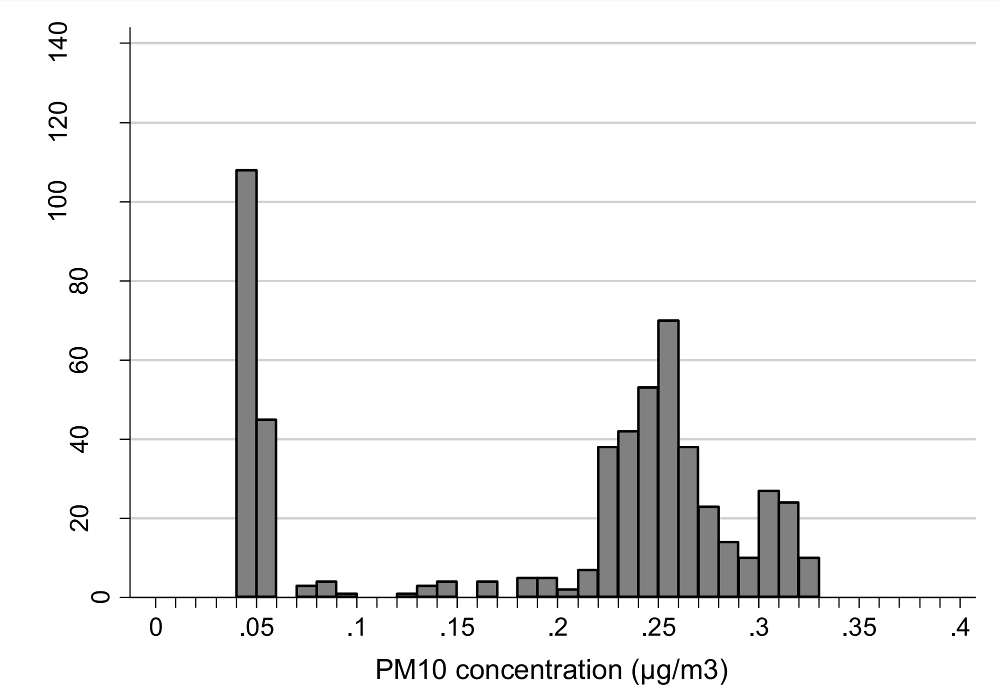

Figure 4.10 *– Distribution of PM10 concentration predicted by the AERMOD model at participants' home addresses. Each bin's width is 0.01 µg/m^3.*

Group		**N**	**Summary statistics** *(metres)*		
			Mean (SD)	*Median (IQR)*	*Min - Max*
SdB	*Cases*	98	0.2492 (0.0475)	0.2494 (0.0439)	0.0821 – 0.3272
	Controls	289	0.2541 (0.0410)	0.2547 (0.0387)	0.0780 – 0.3290
FE	*Cases*	27	0.0470 (0.0035)	0.0468 (0.0058)	0.0412 – 0.0529
	Controls	127	0.0474 (0.0052)	0.0465 (0.0061)	0.0408 (0.0924)

Table 4.5 *– PM10 concentrations among cases and controls, separately in the two municipalities of Sannazzaro de' Burgondi (SdB) and Ferrera Erbognone (FE).*

The existence of two peaks in the distributional plot of PM10 is essentially related to the two municipalities: concentrations are significantly higher in the area of Sannazzaro de' Burgondi than in Ferrera Erbognone (MW=18.16, p<.0001). This can be confirmed also by looking at the map in ***Figure 4.11***.

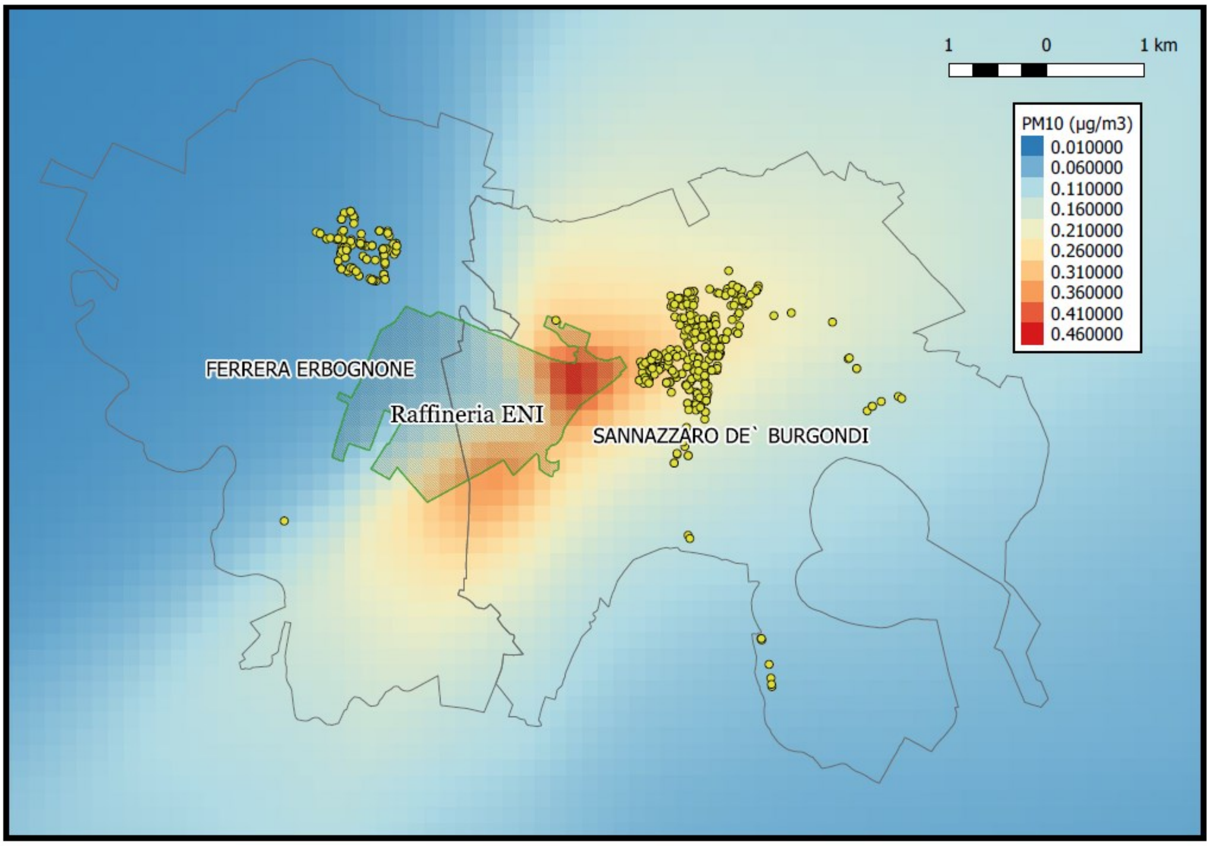

Figure 4.11 *– Map of the registered home addresses (yellow circles) for the 541 study participants, plotted over the grid of PM10 concentrations attributable to gaseous emissions from the refinery (AERMOD estimates). The borders of the municipalities of Sannazzaro de' Burgondi and Ferrera Erbognone are represented, together with the area of the refinery.*

4.3.3.3. Clustered PM10 concentration

A summary of individual PM10 exposure, evaluated on the basis of a previous cluster analysis (which included all the enrolled subjects; see ***3.4.2 Individual exposure estimation***), is reported in ***Tables 4.6 – 4.10***; the PM10 exposure level increased with the cluster ranking. Briefly, several possible ways of clustering (3, 4, 5, 6 and 7 groups) were identified by this analysis. Clustering in two levels essentially coincided with the two municipalities, with a few exceptions, so the 2-clustered PM10 was replaced by municipality. In any case, none of the clustered PM10 concentration variables was significantly associated with being a case or a control, apart for the variable with 5 clusters for which the association was borderline significant. Besides, since clustering in 5, 6 and 7 groups presented categories with very low frequencies (***Tables 4.8 – 4.10***), those were not used as the main exposure variable in multivariate analyses. As can be seen in ***Figures 4.12 – 4.13***, the lowest-PM10-

exposure cluster is almost coinciding with Ferrera Erbognone when PM10 concentrations are clustered in 3 or 4 groups; indeed, these clustered variables are associated with municipality (Fisher's exact tests: 3-clustered PM10, p<.001; 4-clustered PM10, p<.001).

Group		**n (%)**	**Test p-value**	**PM10** *($\mu g/m^3$)*		
				Mean (SD)	*Median (IQR)*	*Min - Max*
Cases *(100%)*	*Cluster 1*	32 (25.6%)	χ^2=2.41 p=.30	0.0565 (0.0250)	0.0473 (0.0073)	0.0412 – 0.1422
	Cluster 2	66 (52.8%)		0.2388 (0.0214)	0.2442 (0.0228)	0.1622 – 0.2676
	Cluster 3	27 (21.6%)		0.3008 (0.0155)	0.3061 (0.0232)	0.2728 – 0.3272
Controls *(100%)*	*Cluster 1*	137 (32.9%)		0.0523 (0.0199)	0.0467 (0.0062)	0.0408 – 0.1422
	Cluster 2	198 (47.6%)		0.2430 (0.0196)	0.2472 (0.2555)	0.1622 – 0.2696
	Cluster 3	81 (19.5%)		0.2983 (0.0176)	0.3061 (0.0300)	0.2701 – 0.3290

Table 4.6 *– Distribution of PM10 concentrations clustered in 3 groups (k=3) by case/control status, with summary statistics of PM10 concentration per each one.*

Group		**n (%)**	**Test p-value**	**PM10** *($\mu g/m^3$)*		
				Mean (SD)	*Median (IQR)*	*Min - Max*
Cases *(100%)*	*Cluster 1*	30 (24.0%)	χ^2=2.58 p=.46	0.0509 (0.0124)	0.0470 (0.0064)	0.0412 – 0.0879
	Cluster 2	6 (4.8%)		0.1614 (0.0188)	0.1657 (0.0272)	0.1376 – 0.1880
	Cluster 3	64 (51.2%)		0.2441 (0.0139)	0.2456 (0.0195)	0.2116 – 0.2728
	Cluster 4	25 (20.0%)		0.3030 (0.0137)	0.3066 (0.0163)	0.2785 – 0.3272
Controls *(100%)*	*Cluster 1*	131 (31.5%)		0.0484 (0.0078)	0.0466 (0.0062)	0.0408 – 0.0924
	Cluster 2	18 (4.3%)		0.1725 (0.0274)	0.1835 (0.0510)	0.1294 – 0.2073
	Cluster 3	193 (46.4%)		0.2473 (0.0149)	0.2450 (0.0230)	0.2105 – 0.2732
	Cluster 4	74 (17.8%)		0.3009 (0.0163)	0.3063 (0.0281)	0.2739 – 0.3290

Table 4.7 *– Distribution of PM10 concentrations clustered in 4 groups (k=4) by case/control status, with summary statistics of PM10 concentration per each one.*

Group		**n (%)**	**Test* p-value**	**PM10** *(µg/m^3)*		
				Mean (SD)	*Median (IQR)*	*Min - Max*
***Cases** (100%)*	*Cluster 1*	30 (24.0%)	p=.082	0.0509 (0.0124)	0.0470 (0.0064)	0.0412 – 0.0879
	Cluster 2	6 (4.8%)		0.1614 (0.0188)	0.1657 (0.0272)	0.1376 – 0.1880
	Cluster 3	43 (34.4%)		0.2369 (0.0105)	0.2403 (0.0182)	0.2116 – 0.2500
	Cluster 4	26 (20.8%)		0.2631 (0.0110)	0.2597 (0.0185)	0.2511 – 0.2867
	Cluster 5	20 (16.0%)		0.3083 (0.0093)	0.3072 (0.0099)	0.2915 – 0.3272
***Controls** (100%)*	*Cluster 1*	131 (31.5%)		0.0484 (0.0078)	0.0466 (0.0062)	0.0408 – 0.0924
	Cluster 2	16 (3.9%)		0.1683 (0.0260)	0.1830 (0.0490)	0.1294 – 0.1983
	Cluster 3	102 (24.5%)		0.2349 (0.0105)	0.2353 (0.0167)	0.2054 – 0.2502
	Cluster 4	115 (27.6%)		0.2638 (0.0097)	0.2617 (0.0149)	0.2507 – 0.2840
	Cluster 5	52 (12.5%)		0.3099 (0.0098)	0.3100 (0.0080)	0.2881 – 0.3290

Table 4.8 *– Distribution of PM10 concentrations clustered in 5 groups (k=5) by case/control status, with summary statistics of PM10 concentration per each one. (*) Fisher's exact test.*

Group		**n (%)**	**Test* p-value**	**PM10** *($\mu g/m^3$)*		
				Mean (SD)	*Median (IQR)*	*Min - Max*
Cases *(100%)*	*Cluster 1*	27 (21.6%)	p=.32	0.0470 (0.0035)	0.0468 (0.0058)	0.0412 – 0.0529
	Cluster 2	3 (2.4%)		0.0860 (0.0033)	0.0879 (0.0058)	0.0821 – 0.0879
	Cluster 3	6 (4.8%)		0.1614 (0.0188)	0.1657 (0.0272)	0.1376 (0.1880)
	Cluster 4	25 (20.0%)		0.2300 (0.0084)	0.2300 (0.0145)	0.2116 – 0.2423
	Cluster 5	41 (32.8%)		0.2543 (0.0095)	0.2528 (0.0132)	0.2433 – 0.2787
	Cluster 6	23 (18.4%)		0.3051 (0.0121)	0.3071 (0.1663)	0.2814 (0.3272)
Controls *(100%)*	*Cluster 1*	126 (30.3%)		0.0471 (0.0034)	0.0465 (0.0060)	0.0408 – 0.0529
	Cluster 2	5 (1.2%)		0.0832 (0.0065)	0.0795 (0.0092)	0.0780 – 0.0924
	Cluster 3	14 (3.4%)		0.1640 (0.0250)	0.1726 – 0.0423	0.1294 – 0.1927
	Cluster 4	72 (17.3%)		0.2286 (0.0095)	0.2297 (0.0104)	0.1982 – 0.2423
	Cluster 5	137 (32.9%)		0.2584 (0.0097)	0.2567 (0.0150)	0.2433 – 0.2787
	Cluster 6	62 (14.9%)		0.3055 (0.0135)	0.3090 (0.0196)	0.2814 (0.3290)

***Table 4.9** – Distribution of PM10 concentrations clustered in 6 groups (k=6) by case/control status, with summary statistics of PM10 concentration per each one. (*) Fisher's exact test.*

Group		n (%)	Test* p-value	PM10 *(µg/m^3)*		
				Mean (SD)	*Median (IQR)*	*Min - Max*
***Cases** (100%)*	*Cluster 1*	16 (12.8%)	p=.47	0.0446 (0.0020)	0.0448 (0.0038)	0.0412 – 0.0470
	Cluster 2	11 (8.8%)		0.0505 (0.0019)	0.0506 (0.0038)	0.0476 – 0.0529
	Cluster 3	3 (2.4%)		0.0860 (0.0033)	0.0879 (0.0058)	0.0821 – 0.0879
	Cluster 4	6 (4.8%)		0.1614 (0.0188)	0.1657 (0.0272)	0.1376 – 0.1880
	Cluster 5	25 (20.0%)		0.2300 (0.0084)	0.2300 (0.0145)	0.2116 – 0.2423
	Cluster 6	41 (32.8%)		0.2543 (0.0095)	0.2528 (0.0132)	0.2433 – 0.2787
	Cluster 7	23 (18.4%)		0.3051 (0.0121)	0.3071 (0.0166)	0.2814 – 0.3272
***Controls** (100%)*	*Cluster 1*	74 (17.8%)		0.0446 (0.0015)	0.0447 (0.0025)	0.0408 – 0.0475
	Cluster 2	52 (12.5%)		0.0506 (0.0016)	0.0506 (0.0022)	0.0476 – 0.0529
	Cluster 3	5 (1.2%)		0.0832 (0.0065)	0.0795 (0.0092)	0.0780 – 0.0924
	Cluster 4	14 (3.4%)		0.1640 (0.0250)	0.1726 (0.0423)	0.1294 – 0.1927
	Cluster 5	72 (17.3%)		0.2286 (0.0095)	0.2297 (0.0104)	0.1982 – 0.2423
	Cluster 6	137 (32.9%)		0.2594 (0.0097)	0.2567 (0.0150)	0.2433 – 0.2787
	Cluster 7	62 (14.9%)		0.3055 (0.0135)	0.3090 (0.0196)	0.2814 – 0.3290

***Table 4.10** – Distribution of PM10 concentrations clustered in 7 groups (k=7) by case/control status, with summary statistics of PM10 concentration per each one. (*) Fisher's exact test.*

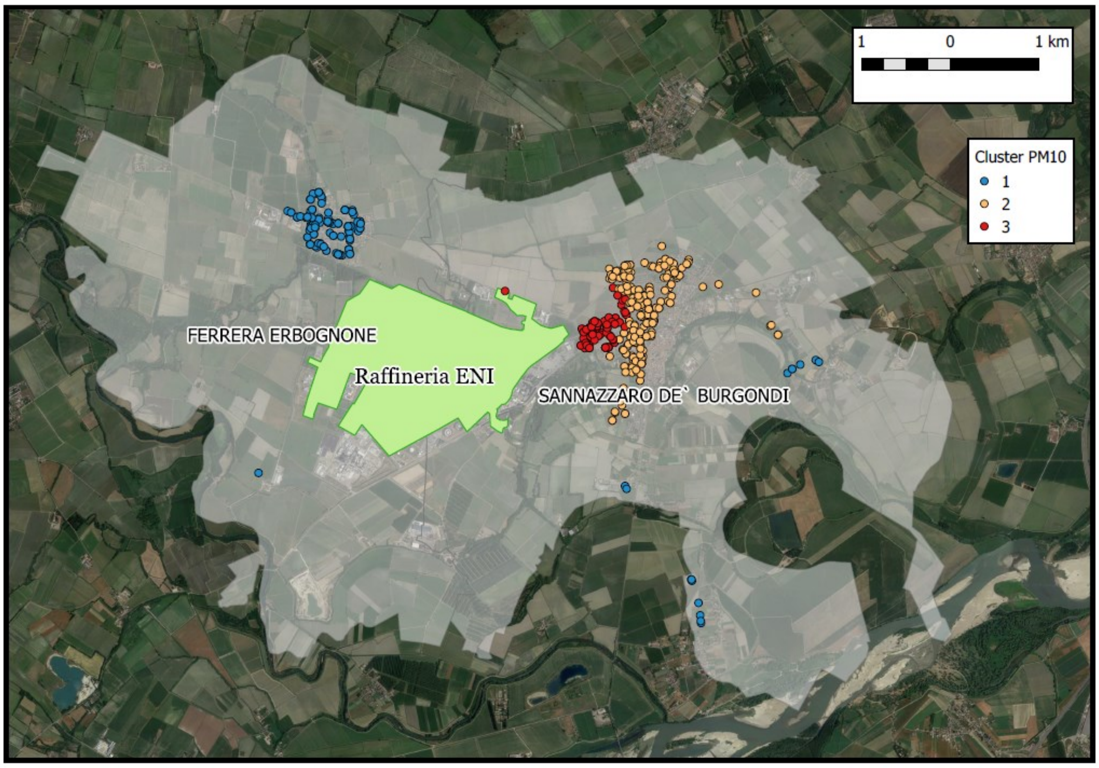

Figure 4.12 *– Spatial distribution of 3-clustered (k=3) individual exposure to PM10. The circles represent geocoded home addresses of study participants.*

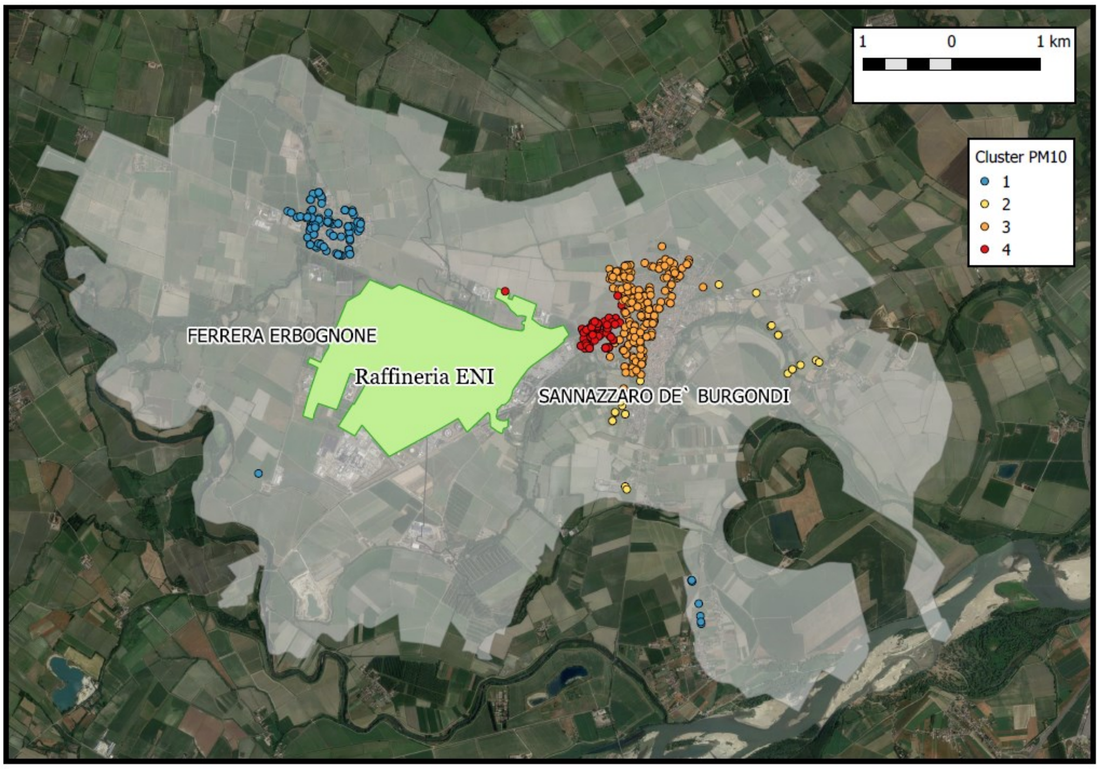

Figure 4.13 *– Spatial distribution of 4-clustered (k=4) individual exposure to PM10. The circles represent geocoded home addresses of study participants.*

4.4. DESCRIPTION OF SURVEY DATA

4.4.1. Socio-demographic characteristics

Regarding marital status[88] (Q1), roughly two-thirds of the participants declared to be married at the time of assessment, while one-sixth was single; 10 subjects marked the "Other" option in the questionnaire, but only one of them specified to be in a domestic partnership). No difference between cases and controls was observed (χ^2=2.32, p=.68) as reported in ***Figure 4.14***.

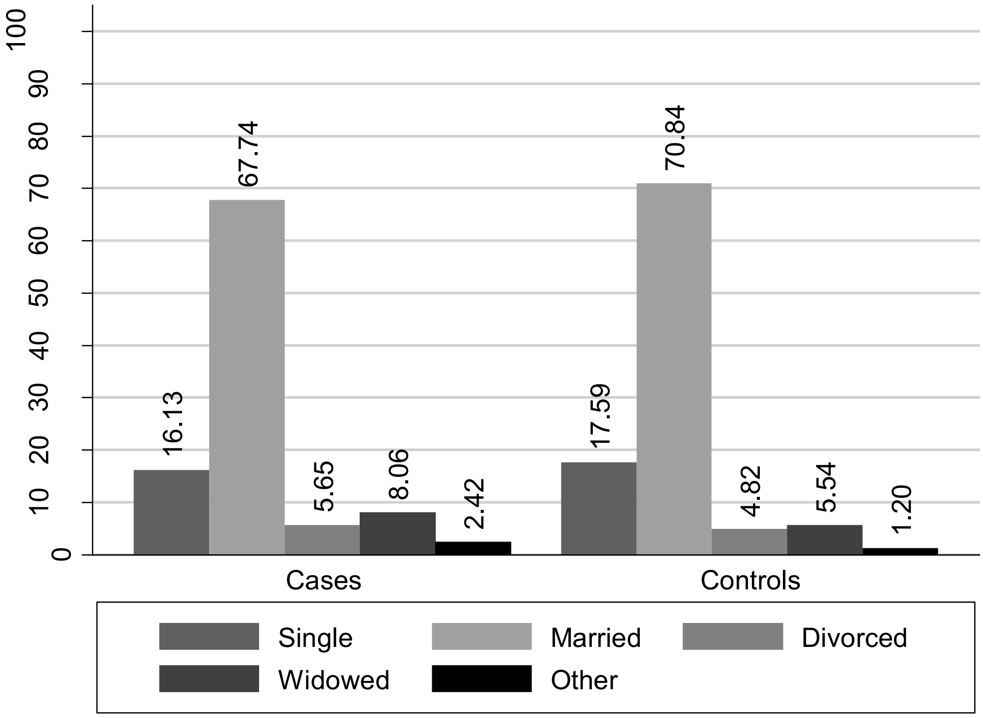

Figure 4.14 *– Distribution of marital status (Q1) by case/control status.*

Concerning the maximum educational grade achieved (Q2), a relevant share of subjects (54%) declared a low-level grade (primary or secondary school), while slightly more than one third of the respondents (34.8%) declared to have completed high school and only one in ten (9.4%) completed academic studies (bachelor's or master's

[88] Question Q1 had 2 missing answers.

degree). Some differences between cases and controls were found, as could be seen in ***Figure 4.15***; in particular, the most frequent educational grade among cases is primary school (31.2%), followed by secondary school (27.6%), whereas among controls the most frequent grade is high school (36.8%) followed by secondary school (31%). However, the association of education and case/control status reached no more than a borderline statistical significance (χ^2=7.95, p=.094).

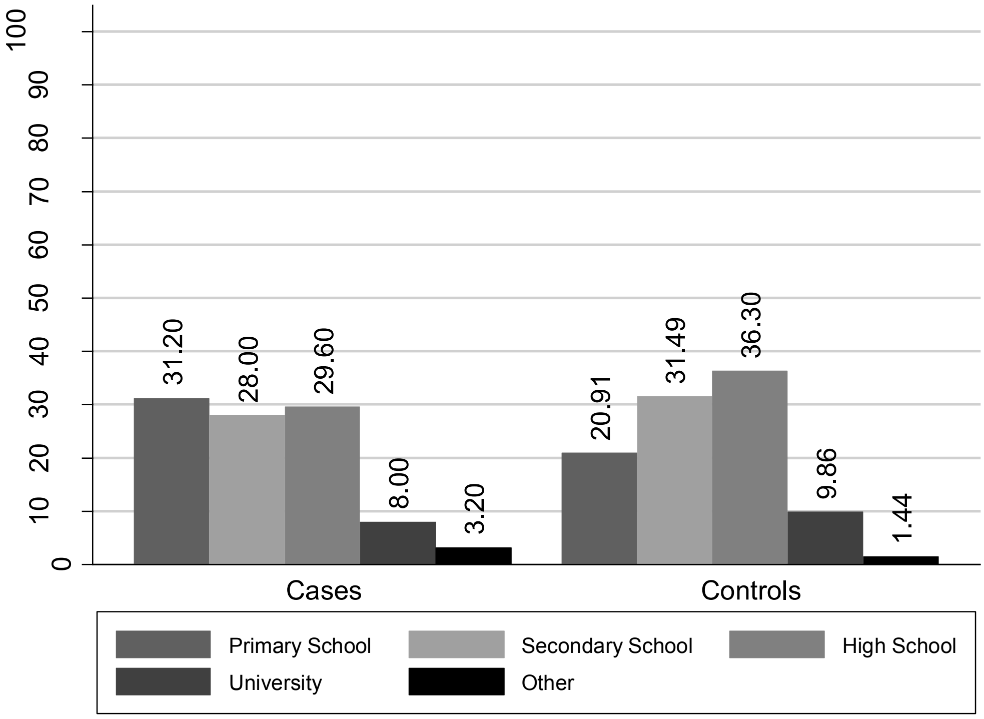

Figure 4.15 *– Distribution of educational level (Q2) by case/control status.*

Significant differences were found between cases and controls in relation to employment status (Q3) (χ^2=17.12, p=.009): although in both groups the most frequent answer was "Retired", about one third of controls (31.5%) declared to have a permanent position, while this was indicated by less than 20% of cases. The distribution of answers to Q3 is reported in ***Figure 4.16***. If educational level is taken into account when looking at the employment status (***Table 4.11***), retired people are more represented among those who achieved a primary or secondary school grade, and less represented among the university-educated.

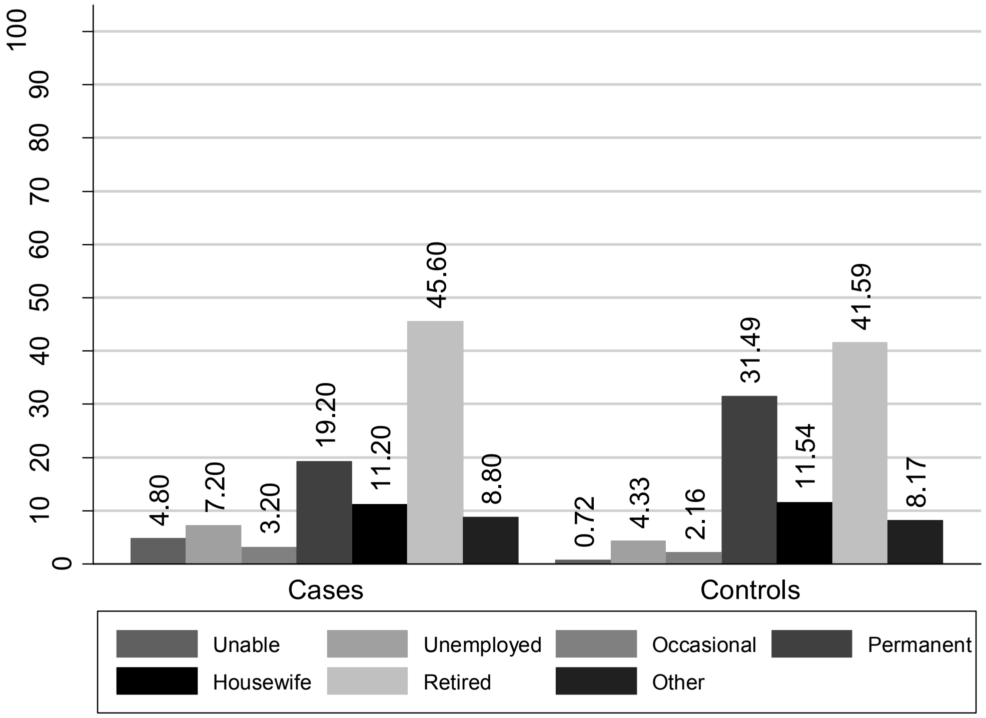

Figure 4.16 *– Distribution of employment status (Q3) by case/control status.*

Employment status	**Educational level**, *n (%)*					**Overall** *n (%)*
	Primary School	***Secondary School***	***High School***	***University***	***Other***	
Unable	1 (0.8%)	5 (3.0%)	3 (1.6%)	0 (0.0%)	0 (0.0%)	9 (1.7%)
Unemployed	6 (4.8%)	11 (6.6%)	9 (4.8%)	1 (2.0%)	0 (0.0%)	27 (5.0%)
Occasional	1 (0.8%)	4 (2.4%)	5 (2.7%)	2 (3.9%)	1 (10.0%)	13 (2.4%)
Permanent	6 (4.8%)	39 (23.5%)	82 (43.6%)	26 (51.0%)	2 (20.0%)	155 (28.7%)
Housewife	22 (17.5%)	21 (12.7%)	17 (9.0%)	1 (2.0%)	1 (10.0%)	62 (11.5%)
Retired	85 (67.5%)	71 (42.8%)	55 (29.3%)	14 (27.5%)	5 (50.0%)	230 (42.5%)
Other	5 (4.0%)	15 (9.0%)	17 (9.0%)	7 (13.7%)	1 (10.0%)	45 (8.3%)
Total	126 (100%)	166 (100%)	188 (100%)	51 (100%)	10 (100%)	541 (100%)

Table 4.11 *– Distribution of employment status within maximum educational grade achieved.*

4.4.2. Housing

Half of the respondents declared to live in a detached house, either isolated or close to other houses. The type of house (***Table 4.12***) was not related to case/control status (χ^2=4.74, p=.45). Taking into account the lack of association, the fact that several categories were scarcely represented, and that more than 5% of respondents marked the option "Other" without specifying, finally this variable was not included in multivariate analyses.

Type of house	**Cases** *n (%)*	**Controls** *n (%)*	**Overall** *n (%)*
Detached house (isolated)	9 (7.3%)	36 (8.8%)	45 (8.4%)
Detached house (close to others)	55 (44.4%)	166 (40.6%)	221 (41.5%)
Semi-detached	17 (13.7%)	60 (14.7%)	77 (14.5%)
Flat	26 (21.0%)	79 (19.3%)	105 (19.7%)
Terraced house	8 (6.5%)	49 (12.0%)	57 (10.7%)
Other	9 (7.3%)	19 (4.7%)	28 (5.3%)
Total	124 (100%)	409 (100%)	533 (100%)

Table 4.12 *– Distribution of type of house (Q6) by case/control status and over the whole sub-sample included in the analyses.*

According to participants' answers in the survey (Q7), most of the houses were close to side roads, while one out of four had a main road nearby. Less than 10% of the subjects lived close to a garden or a pedestrian area. The distribution of type of road near the house (***Figure 4.17***) was similar among cases and controls (χ^2=3.94, p=.41). It is necessary to mention that answer option "Expressway" was seldom used by subjects, and for the sake of multivariate analyses it was grouped together with "Main road".

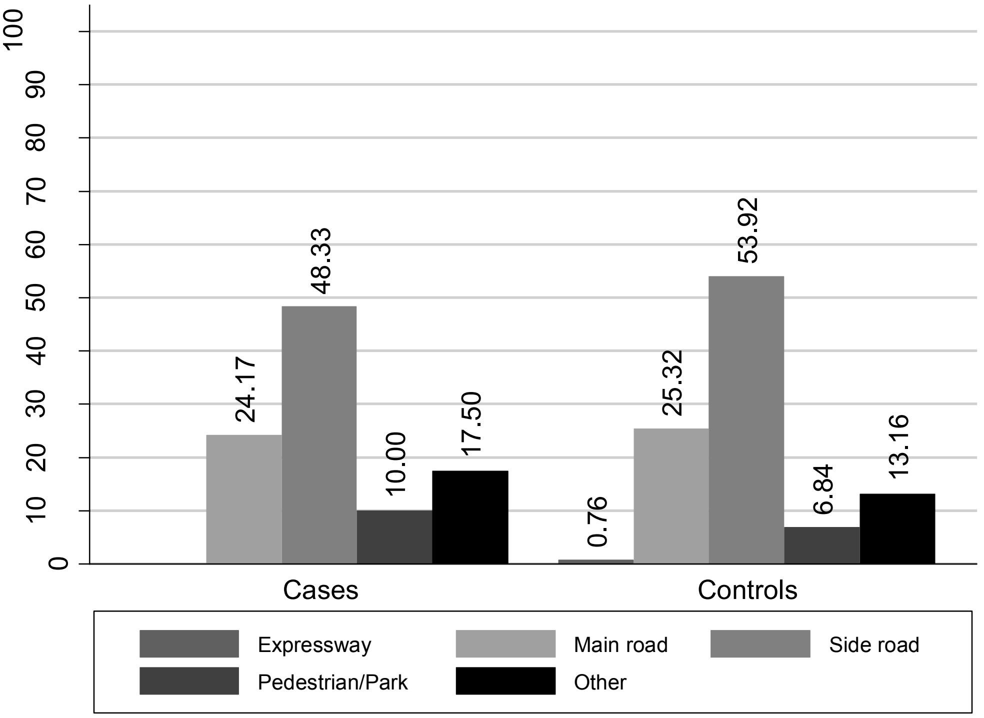

***Figure 4.17** – Distribution of type of road near the house (Q7) by case/control status.*

Even though less than a third of those who answered to Q7 declared to live close to heavily trafficked roads (main roads or expressways), more than half declared that cars were passing frequently or constantly near their house (Q8). A look at ***Figure 4.18*** suggests that cases referred a "Constant" car traffic more often than controls; in the latter group, "Frequent" was more common than in the former. Anyway, no significant association was disclosed between traffic due to cars and being a case or a control (χ^2=1.69, p=.64). On the other hand, traffic due to heavy vehicles (Q9) was defined "Sporadic" by 47% of respondents and "Frequent" by another 12%, again with no statistically significant difference by case/control status (χ^2=1.41, p=.70) (***Figure 4.19***).

When traffic either due to cars or heavy vehicles was recoded in a unique variable, as detailed previously in ***3.5.2 Survey variables used in the study***, then 63% of respondents turned out to be exposed to pollution from road traffic, with no difference between cases and controls (χ^2=0.0004, p=.99). The recoded variable is represented in ***Figure 4.20***.

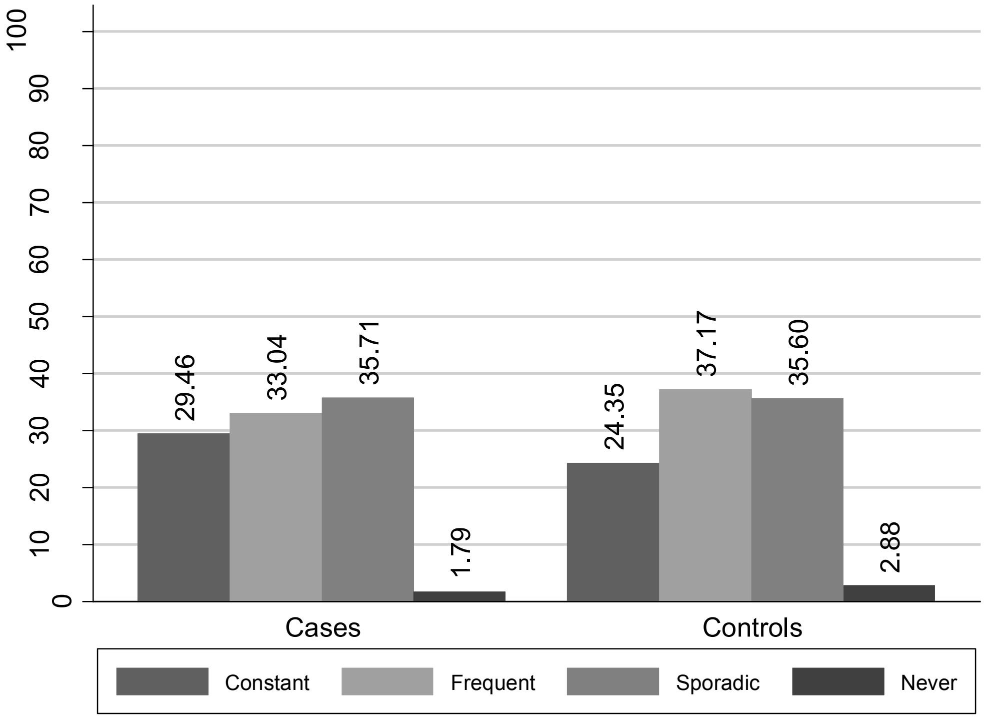

Figure 4.18 *– Distribution of traffic levels due to cars in the streets near the house (Q8) by case/control status.*

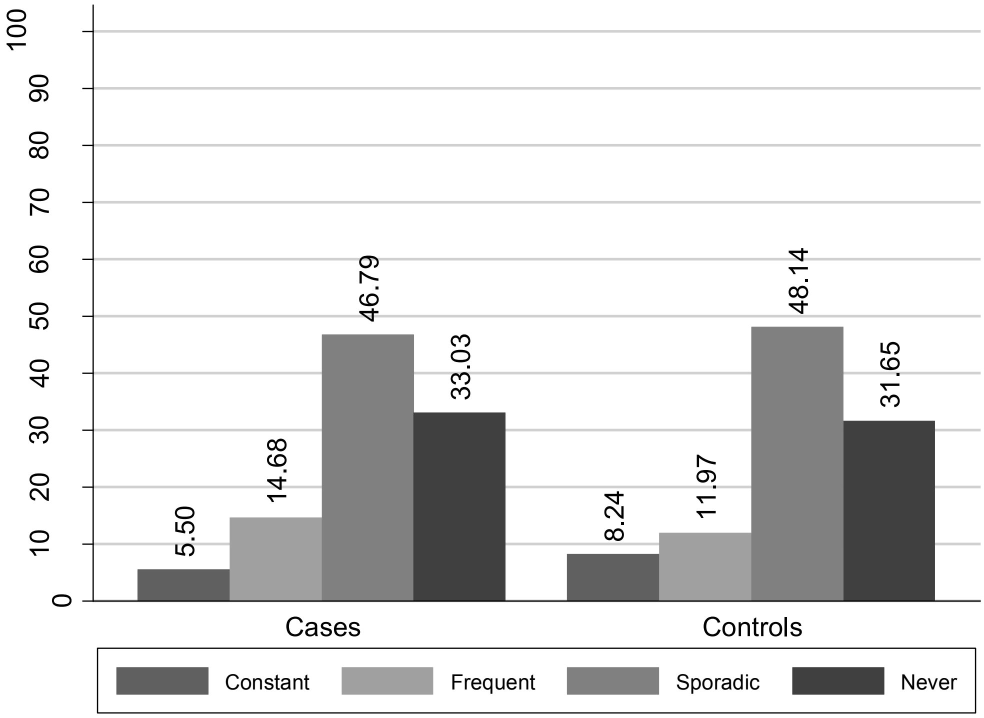

Figure 4.19 *– Distribution of traffic levels due to heavy vehicles in the streets near the house (Q9) by case/control status.*

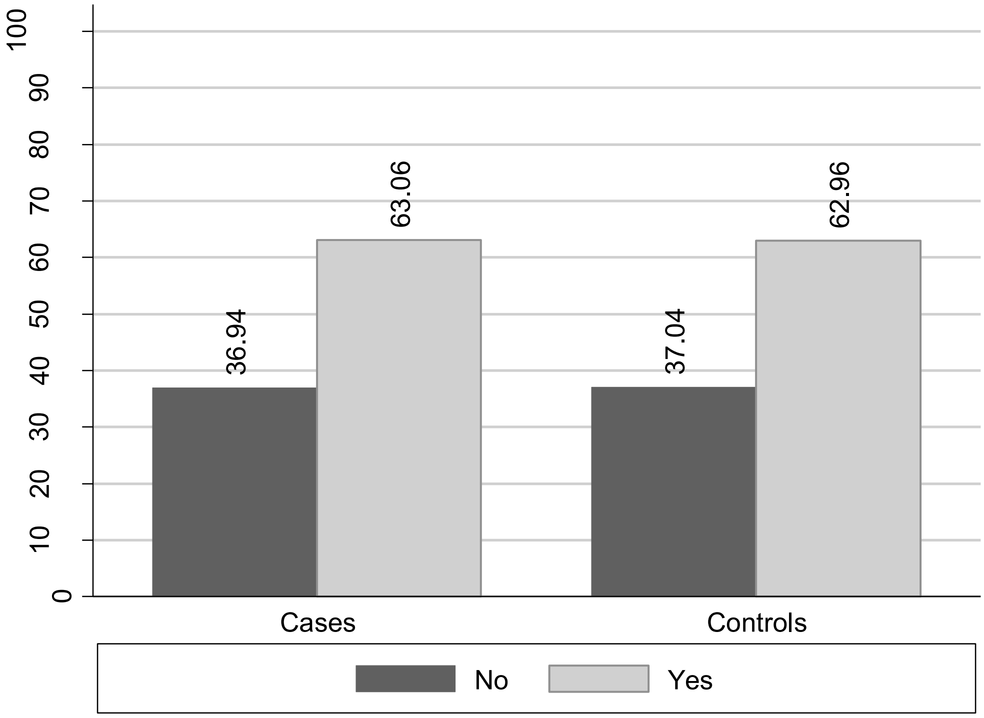

Figure 4.20 *– Distribution of exposure to road traffic (Q8 and Q9, combined) by case/control status.*

4.4.3. Lifestyle

4.4.3.1. Physical activity

The distribution of the answers regarding frequency ratings of different physical activities (Q10), by case/control status and on the overall sub-sample, are presented in ***Table 4.13***. None of the physical activities came out to be statistically associated with being a case or a control (jogging: Fisher's test, p=.79; walking: χ^2=6.25, p=.18; gym: Fisher's test, p=.99; swimming: Fisher's test, p=.79; cycling: χ^2=4.03, p=.40). 49 subjects marked the option "Other", rating it as at least sporadically practised; only 42 specified the other activities. Playing "Other" activities was not significantly associated with case/control status (Fisher's test, p=.64), but this result should be interpreted cautiously, because it comes from a miscellaneous of different activities.

Group		**Frequency rating**, *n (%)*					
		Never	*Every day*	*≥3 days/week*	*<3 days/week*	*Sporadic*	Total
Jogging	***Cases***	53 (79.1%)	0 (0.0%)	3 (4.5%)	2 (3.0%)	9 (13.4%)	67 (100%)
	Controls	180 (72.6%)	2 (0.8%)	9 (3.6%)	15 (6.1%)	42 (16.9%)	248 (100%)
	Overall	233 (74.0%)	2 (0.6%)	12 (3.8%)	17 (5.4%)	51 (16.2%)	315 (100%)
Walking	***Cases***	6 (6.2%)	23 (23.7%)	8 (9.3%)	14 (14.4%)	46 (47.4%)	97 (100%)
	Controls	38 (11.2%)	69 (20.4%)	48 (14.2%)	32 (9.5%)	151 (44.7%)	338 (100%)
	Overall	44 (10.1%)	92 (21.2%)	56 (12.9%)	46 (10.6%)	197 (45.3%)	435 (100%)
Gym	***Cases***	49 (79.0%)	0 (0.0%)	3 (4.8%)	6 (9.7%)	4 (6.5%)	62 (100%)
	Controls	190 (79.8%)	2 (0.8%)	12 (5.0%)	20 (8.3%)	17 (7.1%)	241 (100%)
	Overall	239 (78.9%)	2 (0.7%)	15 (5.0%)	26 (8.6%)	21 (6.9%)	303 (100%)
Swimming	***Cases***	51 (81.0%)	0 (0.0%)	0 (0.0%)	3 (4.8%)	9 (14.3%)	63 (100%)
	Controls	177 (78.3%)	0 (0.0%)	3 (1.3%)	7 (3.1%)	39 (17.3%)	226 (100%)
	Overall	228 (78.9%)	0 (0.0%)	3 (1.0%)	10 (3.5%)	48 (16.6%)	289 (100%)
Cycling	***Cases***	31 (38.3%)	13 (16.1%)	6 (7.4%)	6 (7.4%)	25 (30.9%)	81 (100%)
	Controls	84 (29.1%)	42 (14.5%)	37 (12.8%)	20 (6.9%)	106 (36.7%)	289 (100%)
	Overall	115 (31.1%)	55 (14.9%)	43 (11.6%)	26 (7.0%)	131 (35.4%)	370 (100%)
Other	***Cases***	12 (60.0%)	1 (5.0%)	3 (15.0%)	2 (10.0%)	2 (10.0%)	20 (100%)
	Controls	32 (43.8%)	11 (15.1%)	13 (17.8%)	12 (16.4%)	5 (6.9%)	73 (100%)
	Overall	44 (47.3%)	12 (12.9%)	16 (17.2%)	14 (15.1%)	7 (7.5%)	93 (100%)

Table 4.13 *– Distribution of the frequencies at which study participants declared to play different physical activities (Q10). Distributions are reported for cases and controls separately as well as for the overall sample.*

When the sub-questions regarding the different physical activities were recoded in a unique variable, as specified previously (see ***3.5.2 Survey variables used in the study***), 321 of the respondents (62.9%) were classified in the group of those who practise at least one activity regularly, with no difference in cases and controls (χ^2=1.20, p=.27), even if controls were showed to practise physical activity slightly more than cases. The distribution of the variable is reported in ***Figure 4.21***.

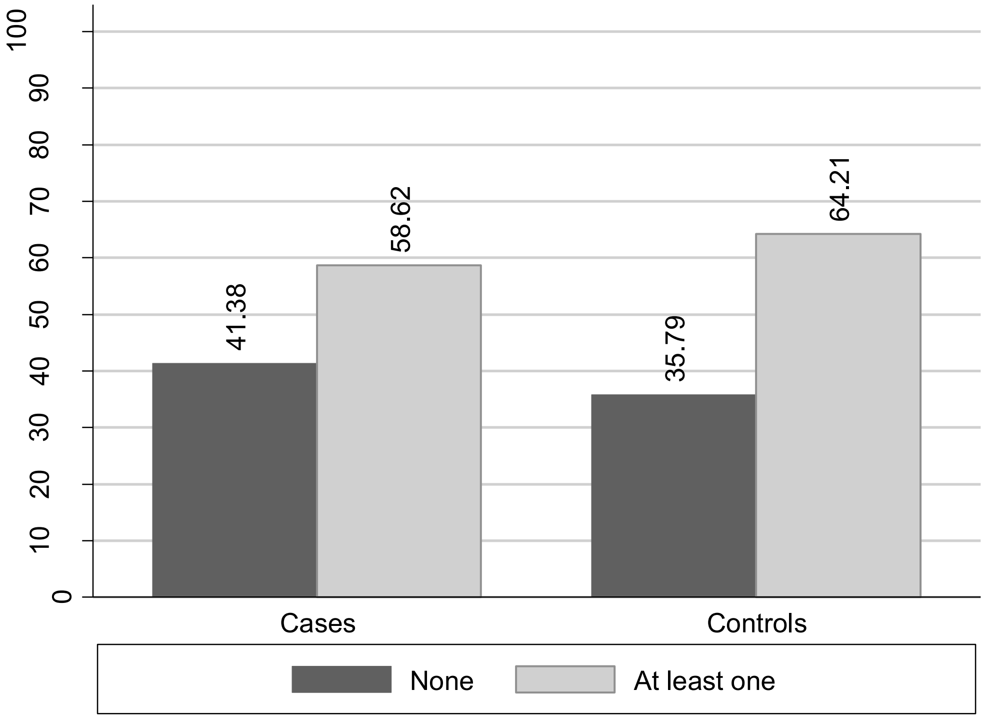

Figure 4.21 *– Distribution of regular practice of at least one physical activity (recoded from Q10.1-Q10.6) by case/control status.*

4.4.3.2. Alcohol consumption

With regards to consuming alcoholic beverages (Q11), 241 respondents (45.1%) answered that they consume these drinks, without any relevant difference between cases and controls (χ^2=0.53, p=.47). The distribution of alcohol consumption in cases and controls is shown in ***Figure 4.22***.

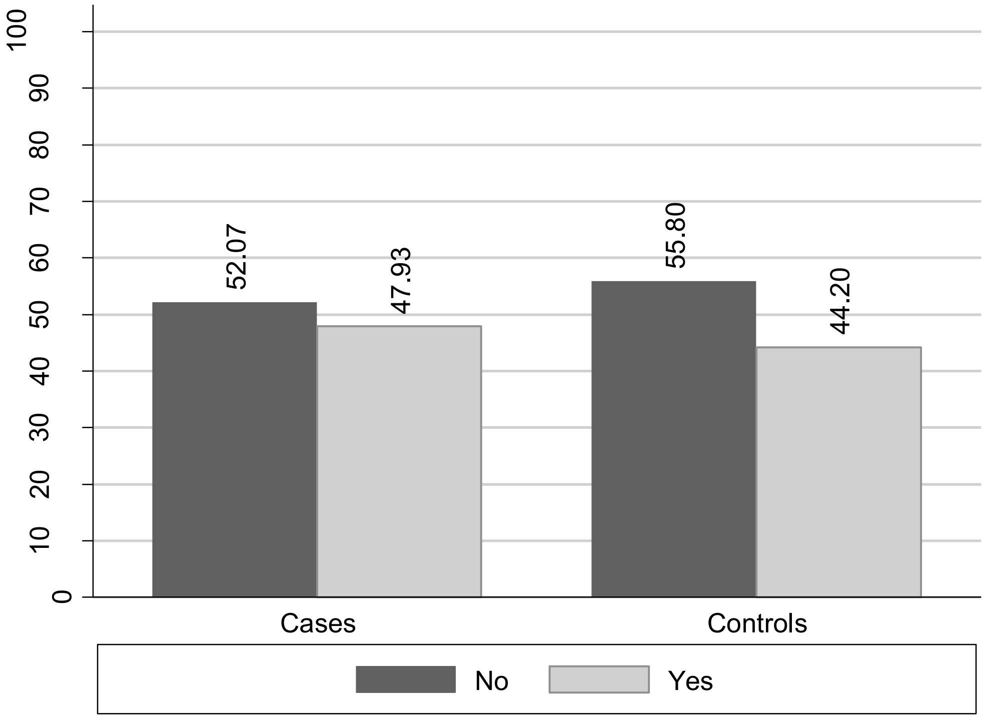

Figure 4.22 *– Distribution of alcohol consumption (Q11) by case/control status.*

As can be seen in ***Table 4.14***, the ratings given by respondents for the frequency of drinking different types of alcoholic beverages were similar in cases and controls (Fisher's tests: wine, p=.51; beer, p=.70; bitter/digestif, p=.78; hard liquors, p=.75; a negligible number of respondents marked the option "Other", thus this was not tested).

Group		**Frequency ratings**, *n (%)*					
		Every day	*3–4 per week*	*1–2 per week*	*Sporadic*	*Never*	Total
Wine	***Cases***	28 (50.0%)	6 (10.7%)	9 (16.1%)	13 (23.2%)	0 (0.0%)	56 (100%)
	Controls	81 (46.3%)	24 (13.7%)	16 (9.1%)	52 (29.7%)	2 (1.1%)	175 (100%)
	Overall	109 (47.2%)	30 (13.0%)	25 (10.8%)	65 (28.1%)	2 (0.9%)	231 (100%)
Beer	***Cases***	1 (2.3%)	2 (5.3%)	4 (10.5%)	24 (63.2%)	7 (18.4%)	38 (100%)
	Controls	5 (4.2%)	10 (8.4%)	23 (19.3%)	62 (52.1%)	19 (16.0%)	119 (100%)
	Overall	6 (3.8%)	12 (7.6%)	27 (17.2%)	86 (54.8%)	26 (16.6%)	157 (100%)
Bitter/ Digestive	***Cases***	1 (3.2%)	0 (0.0%)	2 (6.5%)	12 (38.7%)	16 (51.6%)	31 (100%)
	Controls	1 (1.0%)	2 (2.0%)	11 (10.9%)	39 (38.6%)	48 (47.5%)	101 (00%)
	Overall	2 (1.5%)	2 (1.5%)	13 (9.9%)	51 (38.6%)	64 (48.5%)	132 (100%)
Hard liquors	***Cases***	0 (0.0%)	0 (0.0%)	1 (3.2%)	11 (35.5%)	19 (61.3%)	31 (100%)
	Controls	1 (1.0%)	3 (2.9%)	6 (5.7%)	44 (41.9%)	51 (48.6%)	105 (100%)
	Overall	1 (0.7%)	3 (2.2%)	7 (5.2%)	55 (40.4%)	70 (51.5%)	136 (100%)
Other	***Cases***	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (50.0%)	1 (50.0%)	2 (100%)
	Controls	1 (11.1%)	0 (0.0%)	0 (0.0%)	2 (22.2%)	6 (66.7%)	9 (100%)
	Overall	1 (9.1%)	0 (0.0%)	0 (0.0%)	3 (27.3%)	7 (67.6%)	11 (100%)

Table 4.14 *– Distribution of the frequencies at which study participants declared to drink different alcoholic beverages (Q11). Distributions are reported for cases and controls separately as well as for the overall sample.*

4.4.3.3. Cigarette smoking

Among the 536 study participants that answered the question asking if they ever smoked cigarettes during their life (Q12), 302 declared they had (56.3%). Anyway, as represented in ***Figure 4.23***, lifetime cigarette smoking was only slightly less frequent in controls than cases, the difference being not statistically significant (χ^2=0.22, p=.64).

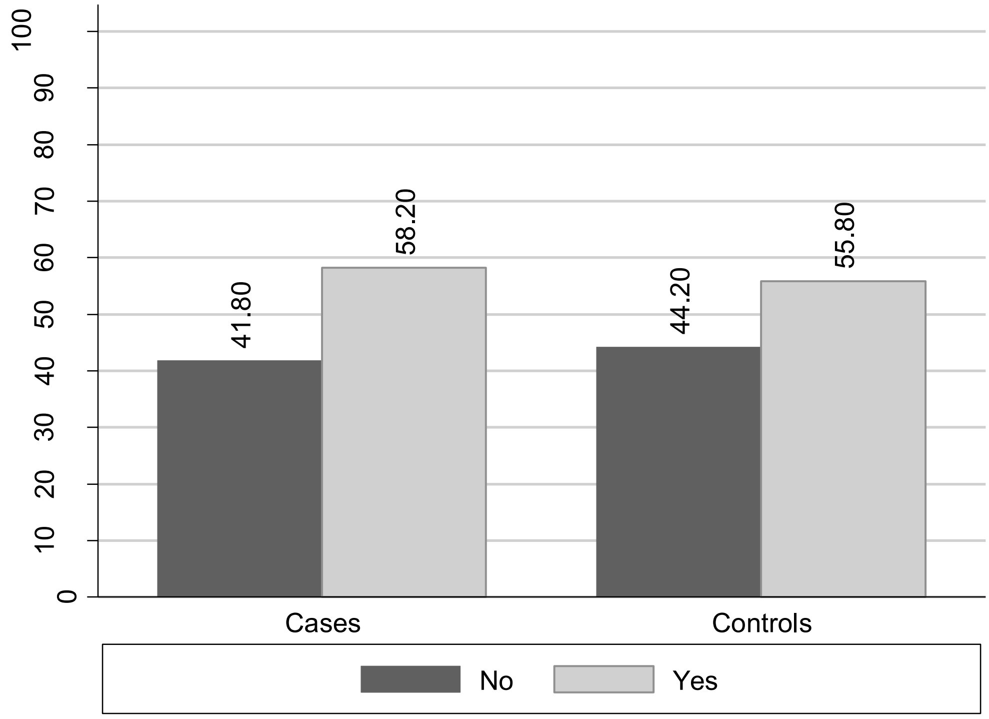

Figure 4.23 *– Distribution of lifetime cigarette smoking (Q12) by case/control status.*

Among those who declared they have smoked cigarettes, slightly more than a third (106, 36.3%) were identified as currently smokers (Q12.2, 10 missings). The share of current smokers (***Figure 4.24***) is significantly different depending on the case/control status (χ^2=6.72, p=.01). On average, age at which smokers started with the habit was 17.6±4.5 years (10 – 42 years), with no difference between cases and controls (t=0.43, p=0.67); age at cessation – for past smokers – was 40.2±13.1 years (16 –70 years), again with no difference in cases and controls (MW=-0.84, p=0.40). The number of cigarettes smoked daily (***Figure 4.25***) appeared slightly higher among cases than controls, but the difference did not reach statistical significance (MW=1.61, p=0.11).

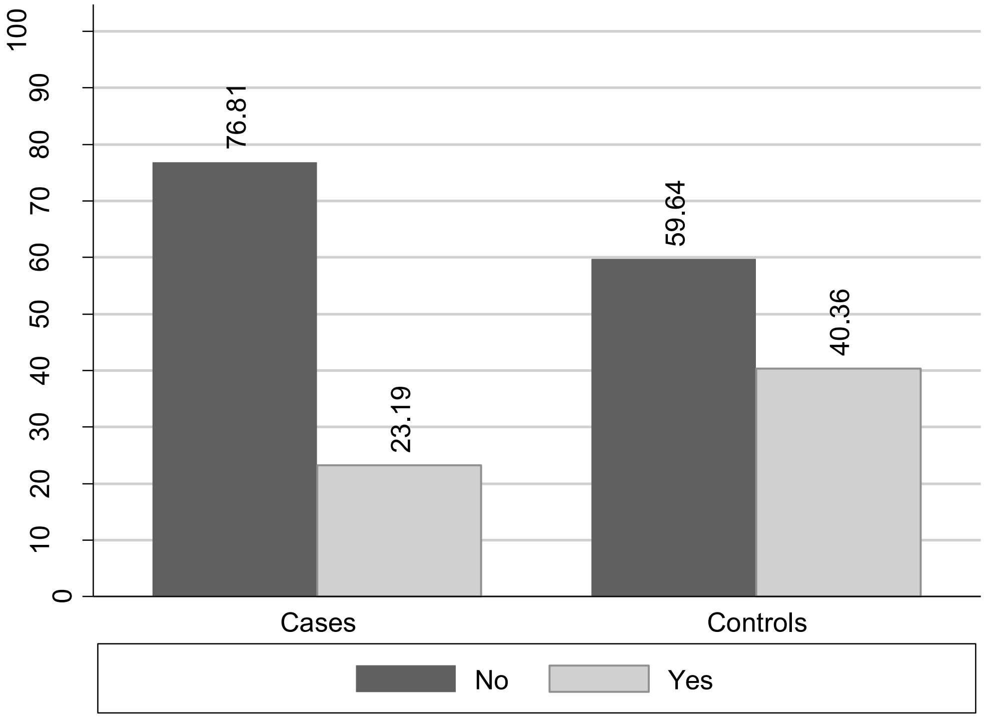

Figure 4.24 *– Distribution of current cigarette smoking (Q12.2) by case/control status among those who declared to have been smokers during their life in Q12.*

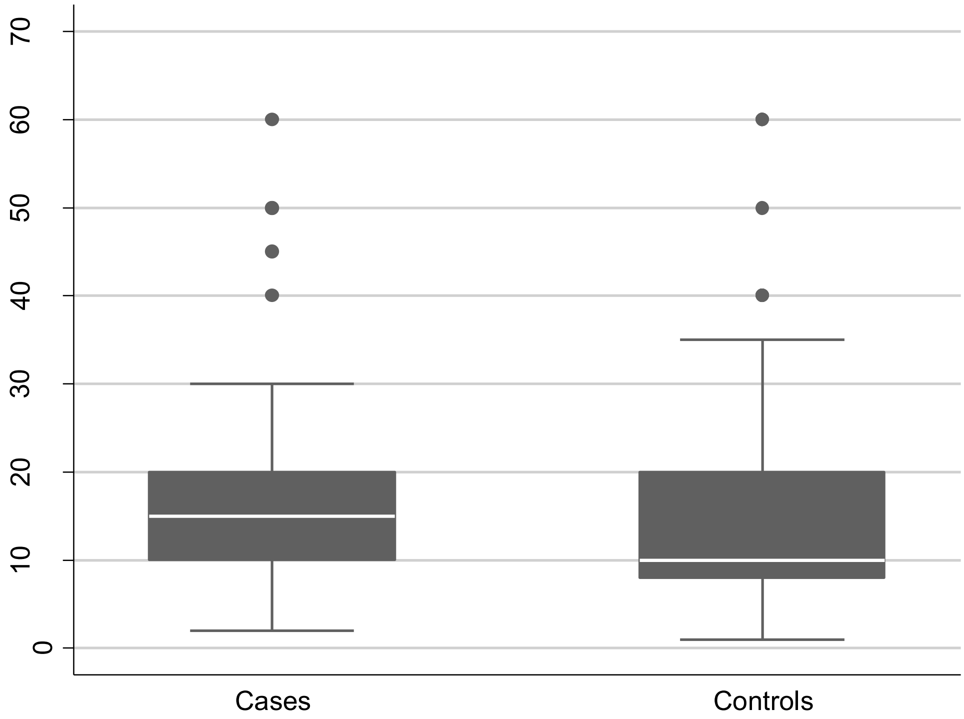

Figure 4.25 *– Description of the number of cigarettes smoked per day (Q12.4) in cases and controls. White line: median; box: 25th-75th centiles.*

Exposure to passive smoking (Q13) was comparable between cases and controls (***Figure 4.26***): the association of the two variables was not statistically significant (χ^2=1.83, p=.18).

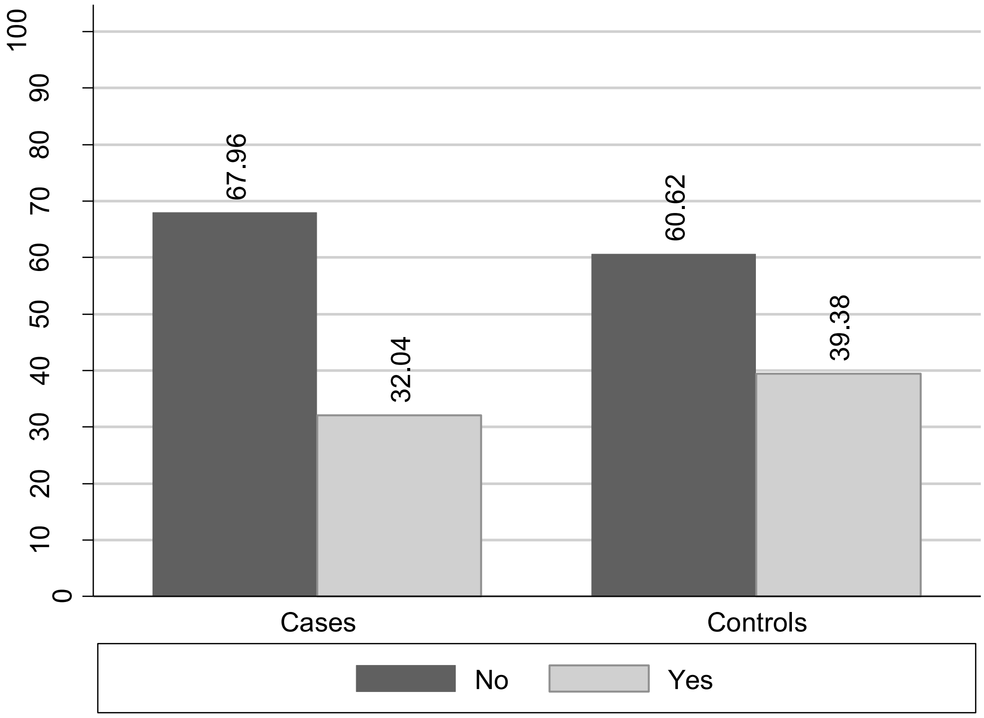

Figure 4.26 *– Distribution of exposure to passive smoking (Q13) by case/control status.*

4.4.3.4. Locally produced food

Consuming of various categories of locally produced food (fruit and vegetables, eggs, meat and fish, rice and cereals) appeared to be quite common among study participants, particularly with regards to rice, cereals, fruit and vegetables. No difference between cases and controls was observed in any of the food categories (fruit and vegetables: χ^2=0.08, p=.78; eggs: χ^2=0.21, p=.65; meat and fish: χ^2=1.39, p=.24; rice and cereals: χ^2=0.13, p=.72). Data are summarised in ***Table 4.15***.

Group		**Food consumption**, *n (%)*		
		No	*Yes*	Total
Fruit and vegetables	***Cases***	37 (31.1%)	82 (68.9%)	119 (100%)
	Controls	114 (29.8%)	269 (70.2%)	383 (100%)
	Overall	151 (30.1%)	351 (69.9%)	502 (100%)
Eggs	***Cases***	43 (37.7%)	71 (62.3%)	114 (100%)
	Controls	128 (35.4%)	234 (64.6%)	362 (100%)
	Overall	171 (35.9%)	305 (64.1%)	476 (100%)
Meat and fish	***Cases***	62 (55.9%)	49 (44.1%)	111 (100%)
	Controls	215 (62.1%)	131 (37.9%)	346 (100%)
	Overall	277 (60.6%)	180 (39.4%)	457 (100%)
Rice and cereals	***Cases***	24 (20.2%)	95 (79.8%)	119 (100%)
	Controls	73 (18.7%)	318 (81.3%)	391 (100%)
	Overall	97 (19.0%)	413 (81.0%)	510 (100%)

Table 4.15 *– Distribution of consuming of different categories of locally produced food (Q15). Distributions are reported for cases and controls separately as well as for the overall sample.*

4.4.4. Health

Concerning the general health status, as self-rated by study participants (Q16), the majority of respondents indicated a positive attitude, with more than 65% of them rating their health as “good”, “very good” or “excellent”. Controls referred more frequently than cases a very positive health perception and this difference was significant (Fisher’s test, p<.001) (***Table 4.16***).

Health status	**Cases** *n (%)*	**Controls** *n (%)*	**Overall** *n (%)*
Poor	17 (13.8%)	16 (4.0%)	33 (6.3%)
Passable	48 (39.0%)	103 (25.5%)	151 (28.7%)
Good	49 (39.8%)	219 (54.2%)	268 (50.9%)
Very good	7 (5.7%)	61 (15.1%)	68 (12.9%)
Excellent	2 (1.6%)	5 (1.2%)	7 (1.3%)
Total	123 (100%)	404 (100%)	527 (100%)

Table 4.16 *– Distribution of self-perceived general health status (Q16). Distributions are reported for cases and controls separately as well as for the overall sample.*

Data regarding diagnosis or administration of a therapeutic regimen for specific diseases and conditions, according to the answers given by study participants in the questionnaire (Q17 and Q18 respectively), are reported in ***Table 4.17*** (see next page). Generally, cases showed more frequently than controls to have received diagnoses or to be undergoing therapeutic regimens for those diseases and conditions.

Questions Q17 and Q18 were recoded into a single variable, summarising if whether the subject was suffering from a condition that could be ascribed to the same causes defined for cases (see ***3.5.2 Survey variables used in the study***). Among the 509 respondents of the survey who indicated at least one diagnosis or therapeutic regimen, 331 (101 cases, 230 controls) were affected by such diseases and conditions[89] (***Figure 4.27***).

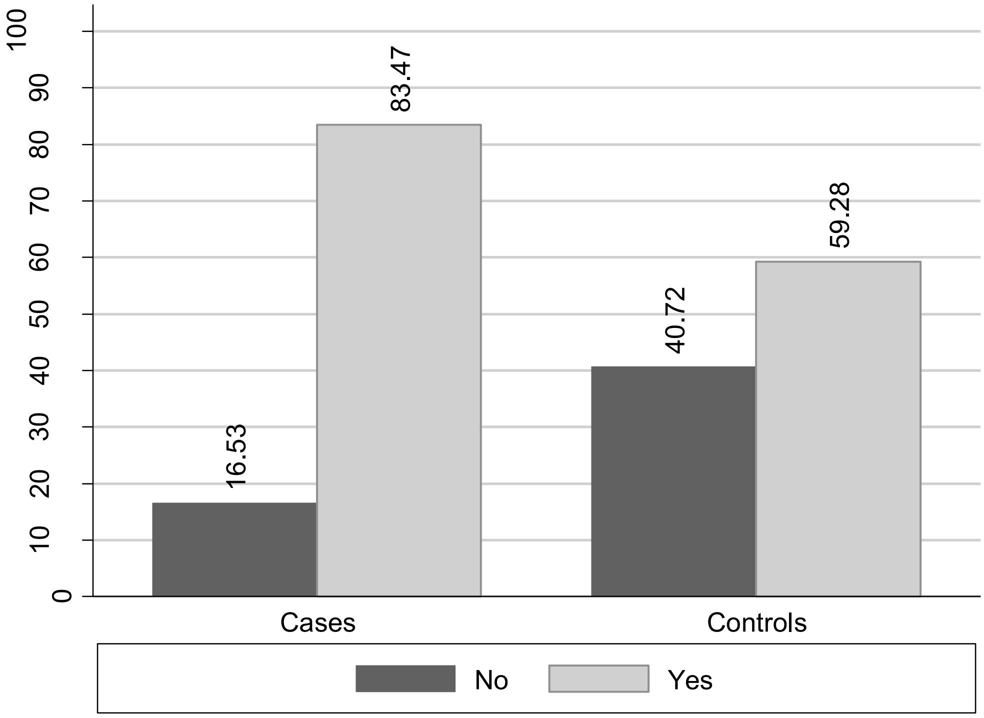

Figure 4.27 *– Distribution of subjects diagnosed with or treated for at least one condition that could be ascribed to the same ICD-IX-CM codes that were used in the definition of cases (arrhythmia, hypertension, asthma, COPD, diseases of the digestive system, recoded from Q17.1-Q17.7 and Q18.1-Q18.7), by case/control status.*

[89] This variable was not tested for association with case/control status: given that it summarises information about the same diseases and condition used for the sake of the identification of "case", a significant association would be meaningless.

Group		**Diagnoses**, *n (%)*				**Therapies**, *n (%)*			
		No	*Yes*	Total	**Test p-value**	*No*	*Yes*	Total	**Test p-value**
Arrhyth-mia	***Cases***	51 (57.9%)	37 (42.1%)	88 (100%)	$\chi^2=38.75$ $p<.001$	56 (65.1%)	30 (34.9%)	86 (100%)	$\chi^2=42.80$ $p<.001$
	Controls	282 (87.3%)	41 (12.7%)	323 (100%)		282 (92.5%)	23 (7.5%)	305(100%)	
	Overall	333 (81.0%)	78 (19.0%)	411 (100%)		338 (86.4%)	53 (13.6%)	391 (100%)	
Hyperten-sion	***Cases***	38 (35.5%)	69 (64.5%)	107 (100%)	$\chi^2=16.73$ $p<.001$	42 (40.8%)	61 (59.2%)	103(100%)	$\chi^2=11.14$ $p=.001$
	Controls	211 (58.0%)	153 (42.0%)	364 (100%)		209 (59.4%)	143 (40.6%)	352(100%)	
	Overall	249 (52.9%)	222 (47.1%)	471 (100%)		251 (55.2%)	204 (44.8%)	455 (100%)	
Dyslipi-daemia	***Cases***	58 (64.4%)	32 (35.6%)	90 (100%)	$\chi^2=4.97$ $p=.026$	58 (69.0%)	26 (31.0%)	84(100%)	$\chi^2=18.77$ $p<.001$
	Controls	246 (76.2%)	77 (23.8%)	323 (100%)		269 (88.5%)	35 (11.5%)	304(100%)	
	Overall	304 (73.6%)	109 (26.4%)	413 (100%)		327 (84.3%)	61 (15.7%)	388 (100%)	
Asthma	***Cases***	75 (90.4%)	8 (9.6%)	83 (100%)	$\chi^2=3.42$ $p=.064$	75 (91.5%)	7 (8.5%)	82(100%)	$\chi^2=8.37$ $p=.004$
	Controls	302 (95.6%)	14 (4.4%)	316 (100%)		294 (98.0%)	6 (2.0%)	300(100%)	
	Overall	377 (94.5%)	22 (5.5%)	399 (100%)		369 (96.6%)	13 (3.4%)	382 (100%)	
COPD	***Cases***	68 (79.1%)	18 (20.9%)	86 (100%)	$\chi^2=10.65$ $p=.001$	72 (90.0%)	8 (10.0%)	80(100%)	$\chi^2=3.2$ $p=.070$
	Controls	292 (91.5%)	27 (8.5%)	319 (100%)		286 (95.3%)	14 (4.7%)	300(100%)	
	Overall	360 (88.9%)	45 (11.1%)	405 (100%)		358 (94.2%)	22 (5.8%)	380 (100%)	
Disease of digestive system	***Cases***	66 (75.0%)	22 (25.0%)	88 (100%)	$\chi^2=4.39$ $p=.036$	64 (80.0%)	16 (20.0%)	80(100%)	$\chi^2=6.52$ $p=.011$
	Controls	269 (84.6%)	49 (15.4%)	318 (100%)		272 (90.4%)	29 (9.6%)	301(100%)	
	Overall	335 (82.5%)	71 (17.5%)	406 (100%)		336 (88.2%)	45 (11.8%)	381 (100%)	
Diabetes	***Cases***	69 (75.8%)	22 (24.2%)	91 (100%)	$\chi^2=15.78$ $p<.001$	66 (78.6%)	18 (21.4%)	84(100%)	$\chi^2=12.49$ $p<.001$
	Controls	292 (91.2%)	28 (8.8%)	320 (100%)		280 (92.1%)	24 (7.9%)	304(100%)	
	Overall	361 (87.8%)	50 (12.2%)	411 (100%)		346 (89.2%)	42 (10.8%)	388 (100%)	

Table 4.17 *– Distribution of diagnoses (Q17.1-Q17.7) and therapies (Q18.1-Q18.7) for various diseases and conditions, reported for cases and controls separately as well as for the overall sample.*

4.5. Association of the main exposure with other variables

As stated in ***3.6.4 Univariate analyses***, the variables that were selected from survey data (see ***3.5.2 Survey variables used in the study***) were tested for association with the main exposure, in order to investigate the existence of potential residual confounding bias. Because the subdivision in 5, 6 or 7 clusters have not been used in multivariate analyses (see ***4.3.3.3 Clustered PM10 concentration***), only municipality (roughly equivalent to 2-clustered PM10) and 3- and 4-clustered PM10 concentrations have been tested.

4.5.1. Demographics

Gender is significantly associated with municipality, with males being 58.9% of the subjects from Sannazzaro de' Burgondi and 46.8% of those from Ferrera Erbognone (χ^2=6.86, p=.009). On the contrary, a significant relationship is observed neither with the 3-clustered nor with the 4-clustered PM10 concentration (χ^2=4.17, p=.12 and χ^2=4.85, p=.18, respectively), in spite of differences up to 10% in the share of males in the lowest-exposure cluster (or clusters 1 and 2, in the 4-clustered variable) against the others. Data are reported in ***Table 4.18***.

Group		**Gender**, *n (%)*		**Total**
		Males	***Females***	*n (%)*
Municipality	***SdB***	228 (58.9%)	159 (41.1%)	387 (100%)
	FE	72 (46.8%)	82 (53.2%)	154 (100%)
3-clustered PM10	***Cluster 1***	83 (49.1%)	86 (50.9%)	169 (100%)
	Cluster 2	153 (58.0%)	111 (42.0%)	264 (100%)
	Cluster 3	64 (59.3%)	44 (40.7%)	108 (100%)
4-clustered PM10	***Cluster 1***	79 (49.1%)	82 (50.9%)	161 (100%)
	Cluster 2	12 (50.0%)	12 (50.0%)	24 (100%)
	Cluster 3	152 (59.1%)	105 (40.9%)	257 (100%)
	Cluster 4	57 (57.6%)	42 (42.4%)	99 (100%)

Table 4.18 – *Distribution of gender by main exposure (municipality, 3-clustered PM10 concentration, 4-clustered PM10 concentration). N=541.*

With regards to *age* (***Table 4.19***), the difference between the two municipalities is borderline significant, but it is of scarce practical relevance: on average, in Sannazzaro de' Burgondi survey respondents were 3 years older than in Ferrera Erbognone. No difference is observed with 3-clustered or with the 4-clustered exposure; anyway, the mean age of subjects in the lowest-exposed cluster is always roughly 3 years less than in the other clusters[90].

Group		**Age**, years *mean (SD)*	**Test** **p-value**

Municipality	***SdB***	57.9±0.5	MW=1.89
			p=.059

	FE	55.0±1.2	

3-clustered PM10	***Cluster 1***	55.0±14.1	KW=4.91
			p=.086

	Cluster 2	57.7±10.8	
	Cluster 3	58.8±10.4	

4-clustered PM10	***Cluster 1***	55.0±14.2	KW=4.46
			p=.22

	Cluster 2	58.2±10.0	
	Cluster 3	57.6±11.0	
	Cluster 4	58.9±10.1	

Table 4.19 *– Summary statistics of age (referred to 2014) by main exposure (municipality, 3-clustered PM10 concentration, 4-clustered PM10 concentration). N=541.*

4.5.2. Socio-demographic characteristics

Marital status (Q1) is borderline associated with both municipality and 4-clustered PM10 (χ^2=9.95, p=.041 and χ^2=22.56, p=.032, respectively), but not with the 3-clustered exposure (χ^2=11.58, p=.17)[91]. As can be seen in ***Table 4.20***, differences mainly reside in the share of singles and married persons: for instance, looking at the distribution of marital status by municipality, singles are more frequently encountered

[90] This can be easily explained by recalling that, as previously mentioned, the lowest-exposure cluster essentially coincides with Ferrera Erbognone and the higher-exposure clusters almost only include people from Sannazzaro de' Burgondi (see ***4.3.3.3 Clustered PM10 concentration***).

[91] The variable was tested with Pearson's Chi-squared test for independence, despite the optimal conditions for its applicability were not respected (specifically, expected frequencies were below 5 in various cells of the contingency table). Unfortunately, the high number of categories made the alternative road of applying a Fisher's exact test unfeasible because of its excessive computational requirements.

in Ferrera Erbognone than in Sannazzaro de' Burgondi (24.0% vs. 14.6%), while the opposite is for the married ones (61.7% vs. 73.5%). A similar situation is observed with the distributions of marital status by 3- and 4-clustered PM10.

Educational level (Q2) was not associated with municipality (χ^2=7.83, p=.098), nor with the 3-clustered (χ^2=13.77, p=.088) or the 4-clustered PM10 (χ^2=16.03, p=.19)[91] (***Table 4.21***). Anyway, the respondents from Ferrera Erbognone did not appear extremely different from those from Sannazzaro de' Burgondi respect to their educational level (e.g., respectively in Sannazzaro and Ferrera: primary school 22.5% vs. 25.3%; secondary school 29.2% vs. 34.2%; high school 37.0% vs. 29.2%; university degree 8.8% vs. 11.4%; other 2.6% vs. 0.0%); the same could be observed in the clustered variables, for cluster 1 versus the other clusters. Besides, it is relevant to mention that the associations tended to disappear if the 10 subjects who answered "Other" to Q2 – all registered in Sannazzaro – were excluded.

Group		**Marital status**, *n (%)*					Total *n (%)*
		Single	*Married*	*Divor-ced*	*Wido-wed*	*Other*	
Municipality	***SdB***	56 (14.6%)	283 (73.5%)	20 (5.2%)	20 (5.2%)	6 (1.6%)	385 (100%)
	FE	37 (24.0%)	95 (61.7%)	7 (4.6%)	13 (8.4%)	2 (1.3%)	154 (100%)
3-clustered PM10	***Cluster 1***	39 (23.1%)	104 (61.5%)	9 (5.3%)	14 (8.3%)	3 (1.8%)	169 (100%)
	Cluster 2	37 (14.1%)	195 (74.1%)	15 (5.7%)	12 (4.6%)	4 (1.5%)	263 (100%)
	Cluster 3	17 (15.9%)	79 (73.8%)	3 (2.8%)	7 (6.5%)	1 (0.9%)	107 (100%)
4-clustered PM10	***Cluster 1***	39 (24.2%)	100 (62.1%)	7 (4.4%)	13 (8.1%)	2 (1.2%)	161 (100%)
	Cluster 2	0 (0.0%)	18 (75.0%)	4 (16.7%)	1 (4.2%)	1 (4.2%)	24 (100%)
	Cluster 3	39 (15.2%)	188 (73.4%)	13 (5.1%)	12 (4.7%)	4 (1.6%)	256 (100%)
	Cluster 4	15 (15.3%)	72 (73.5%)	3 (3.1%)	7 (7.1%)	1 (1.0%)	98 (100%)

Table 4.20 *– Distribution of marital status (Q1) by main exposure (municipality, 3-clustered PM10 concentration, 4-clustered PM10 concentration). N=539.*

Group		Educational level, *n (%)*					Total *n (%)*
		Prima-ry	*Secon-dary*	*High school*	*Univer-sity*	*Other*	
Municipality	***SdB***	87 (22.5%)	113 (29.2%)	143 (37.0%)	34 (8.8%)	10 (2.6%)	387 (100%)
	FE	39 (25.3%)	53 (34.4%)	45 (29.2%)	17 (11.0%)	0 (0.0%)	154 (100%)
3-clustered PM10	***Cluster 1***	42 (24.9%)	61 (36.1%)	48 (28.4%)	17 (10.1%)	1 (0.6%)	169 (100%)
	Cluster 2	53 (20.1%)	74 (28.0%)	102 (38.6%)	29 (11.0%)	6 (2.3%)	264 (100%)
	Cluster 3	31 (28.7%)	31 (28.7%)	38 (35.2%)	5 (4.6%)	3 (2.8%)	108 (100%)
4-clustered PM10	***Cluster 1***	40 (24.8%)	58 (36.0%)	46 (28.6%)	17 (10.6%)	0 (0.0%)	161 (100%)
	Cluster 2	6 (25.0%)	8 (33.3%)	8 (33.3%)	1 (4.2%)	1 (4.2%)	24 (100%)
	Cluster 3	52 (20.2%)	71 (27.6%)	100 (38.9%)	28 (10.9%)	6 (2.3%)	257 (100%)
	Cluster 4	28 (28.3%)	29 (29.3%)	34 (34.3%)	5 (5.1%)	3 (3.0%)	99 (100%)

Table 4.21 *– Distribution of educational level (Q2) by main exposure (municipality, 3-clustered PM10 concentration, 4-clustered PM10 concentration). N=541.*

4.5.3. Housing (traffic near the house)

No significant association was disclosed between *traffic in the streets near the subjects' houses* (Q8 and Q9, combined) and municipality, 3-clustered or 4-clustered PM10 concentration (municipality: χ^2=0.25, p=.62; 3-clustered: χ^2=3.14, p=.21; 4-clustered: χ^2=3.33, p=.34). The distributions of traffic by each of the main exposures are reported in ***Table 4.22***. The share of subjects exposed to traffic due to the passing of cars or heavy vehicles in the streets near their houses is pretty similar in Sannazzaro and Ferrera, but differences of roughly 10% could be observed between the lowest-PM10 and the highest-PM10 exposure clusters in both the 3-clustered and 4-clustered PM10 concentrations.

Group		**Traffic**, *n (%)*		Total *n (%)*
		No	*Yes*	
Municipality	***SdB***	126 (36.3%)	221 (63.7%)	347 (100%)
	FE	55 (38.7%)	87 (61.3%)	142 (100%)
3-clustered PM10	***Cluster 1***	63 (40.9%)	91 (59.1%)	154 (100%)
	Cluster 2	87 (37.5%)	145 (62.5%)	232 (100%)
	Cluster 3	31 (30.1%)	72 (69.9%)	103 (100%)
4-clustered PM10	***Cluster 1***	60 (40.8%)	87 (59.2%)	147 (100%)
	Cluster 2	10 (45.5%)	12 (54.5%)	22 (100%)
	Cluster 3	82 (36.4%)	143 (63.6%)	225 (100%)
	Cluster 4	29 (30.5%)	66 (69.5%)	95 (100%)

Table 4.22 *– Distribution of traffic in the streets near the subjects' houses (Q8 and Q9, combined) by main exposure (municipality, 3-clustered PM10 concentration, 4-clustered PM10 concentration). N=489.*

4.5.4. Lifestyle

Regularly practising at least one *physical activity* (Q10, recoded) was essentially not associated with municipality or clustered exposure variables (municipality: χ^2=2.75, p=.098; 3-clustered: χ^2=2.92, p=.23; 4-clustered: χ^2=5.13, p=.16), even though differences up to approximately 15% could be observed (***Table 4.23***).

Group		**Physical activity**, *n (%)*		Total *n (%)*
		No	*At least one*	
Municipality	***SdB***	142 (39.3%)	219 (60.7%)	361 (100%)
	FE	47 (31.5%)	102 (68.5%)	149 (100%)
3-clustered PM10	***Cluster 1***	53 (32.7%)	109 (67.3%)	162 (100%)
	Cluster 2	101 (40.7%)	147 (59.3%)	248 (100%)
	Cluster 3	35 (35.0%)	65 (65.0%)	100 (100%)
4-clustered PM10	***Cluster 1***	49 (31.6%)	106 (68.4%)	155 (100%)
	Cluster 2	11 (50.0%)	11 (50.0%)	22 (100%)
	Cluster 3	98 (40.5%)	144 (59.5%)	242 (100%)
	Cluster 4	31 (34.1%)	60 (65.9%)	91 (100%)

Table 4.23 *– Distribution of regularly practising at least one physical activity (Q10, recoded) by main exposure (municipality, 3-clustered PM10 concentration, 4-clustered PM10 concentration). N=510.*

Whilst *alcohol consumption* (Q11) was disclosed to be less prevalent among respondents from Ferrera Erbognone (39.6%) than among those from Sannazzaro de' Burgondi (47.2%), the association with municipality was not statistically significant (χ^2=2.58, p=.11). On the contrary, a borderline-significant association was observed with 3- and 4-clustered PM10 (χ^2=6.23, p=.044 and χ^2=7.88, p=.049, respectively). In spite of the lack of statistically significant associations, differences up to 10% can be observed, in particular between the intermediate-exposure clusters and the extreme clusters (***Table 4.24***).

Group		**Alcohol consumption**, *n (%)*		**Total** *n (%)*
		No	***Yes***	
Municipality	***SdB***	201 (52.8%)	180 (47.2%)	381 (100%)
	FE	93 (60.4%)	61 (39.6%)	154 (100%)
3-clustered PM10	***Cluster 1***	101 (59.8%)	68 (40.2%)	169 (100%)
	Cluster 2	128 (49.4%)	131 (50.6%)	256 (100%)
	Cluster 3	65 (60.8%)	42 (39.2%)	107 (100%)
4-clustered PM10	***Cluster 1***	96 (59.6%)	65 (40.4%)	161 (100%)
	Cluster 2	11 (45.8%)	13 (54.2%)	24 (100%)
	Cluster 3	125 (49.6%)	127 (50.4%)	252 (100%)
	Cluster 4	62 (63.3%)	36 (36.7%)	98 (100%)

Table 4.24 *– Distribution of alcohol consumption (Q11) by main exposure (municipality, 3-clustered PM10 concentration, 4-clustered PM10 concentration). N=535.*

Analogously, *lifetime cigarette smoking* (Q12) was less prevalent in Ferrera than in Sannazzaro (47.7% vs. 59.8%), and in this case the difference was fully significant by municipality (χ^2=6.48, p=.011) as well as by the 3-clustered exposure (χ^2=7.78, p=.020), and borderline with the 4-clustered exposure (χ^2=7.21, p=.065). In this case as well, differences up to approximately 15% can be observed (***Table 4.25***), with the prevalence of smokers increasing with the clusters of PM10 exposure.

Group		Cigarette smoking, *n (%)*		Total *n (%)*
		No	*Yes*	
Municipality	***SdB***	154 (40.2%)	229 (59.8%)	383 (100%)
	FE	80 (52.3%)	73 (47.7%)	153 (100%)
3-clustered PM10	***Cluster 1***	87 (51.8%)	81 (48.2%)	168 (100%)
	Cluster 2	109 (41.8%)	152 (58.2%)	261 (100%)
	Cluster 3	38 (35.5%)	69 (64.5%)	107 (100%)
4-clustered PM10	***Cluster 1***	82 (51.3%)	78 (48.7%)	160 (100%)
	Cluster 2	12 (50.0%)	12 (50.0%)	24 (100%)
	Cluster 3	105 (41.3%)	149 (58.7%)	254 (100%)
	Cluster 4	35 (35.7%)	63 (64.3%)	98 (100%)

Table 4.25 *– Distribution of lifetime cigarette smoking (Q12) by main exposure (municipality, 3-clustered PM10 concentration, 4-clustered PM10 concentration). N=536.*

4.5.5. Health

Finally, diagnoses or therapies for *other diseases* (Q17 and Q18, combined) were not statistically associated with either municipality (χ^2=0.22, p=.64), 3-clustered (χ^2=1.37, p=.51) or 4-clustered PM10 (χ^2=1.35, p=.72). Data are shown in ***Table 4.26***.

Group		Other diseases, *n (%)*		Total *n (%)*
		No	*Yes*	
Municipality	***SdB***	125 (34.3%)	239 (65.7%)	364 (100%)
	FE	53 (36.5%)	92 (63.5%)	145 (100%)
3-clustered PM10	***Cluster 1***	57 (35.9%)	102 (64.1%)	159 (100%)
	Cluster 2	91 (36.4%)	159 (63.6%)	250 (100%)
	Cluster 3	30 (30.0%)	70 (70.0%)	100 (100%)
4-clustered PM10	***Cluster 1***	55 (36.2%)	97 (63.8%)	152 (100%)
	Cluster 2	7 (30.4%)	16 (69.6%)	23 (100%)
	Cluster 3	88 (36.4%)	154 (63.6%)	242 (100%)
	Cluster 4	28 (30.4%)	64 (69.6%)	92 (100%)

Table 4.26 *– Distribution of diagnosis or treatment for other diseases, compatible with those that determined the hospitalisation of cases (Q17 and Q18, combined), by main exposure (municipality, 3-clustered PM10 concentration, 4-clustered PM10 concentration). N=509.*

4.6. CRUDE ODDS RATIOS OF THE MAIN EXPOSURE

Crude (unadjusted) ORs were calculated for the environmental exposure alone. As the main exposure variable, municipality (equivalent to PM10 being subdivided in 2 clusters) as well as the 3-clustered and 4-clustered PM10 concentrations were used.

Crude ORs of being exposed given case or control status indicated that the higher the exposure level, the higher the odds of exposure, albeit none of the ORs was statistically significant (***Table 4.27***).

Exposure variable	OR	95%CI	Test* p-value
Municipality (used as equivalent to 2-clustered PM10)			
SdB vs. FE	1.595	0.990 – 2.569	χ^2=1.37 p=.51
3-clustered PM10			
Cluster 2 vs. 1	1.427	0.886 – 2.299	χ^2=2.16 p=.14
Cluster 3 vs. 1	1.427	0.796 – 2.558	χ^2=1.44 p=.23
4-clustered PM10			
Cluster 2 vs. 1	1.456	0.530 – 3.995	χ^2=0.54 p=.46
Cluster 3 vs. 1	1.448	0.888 – 2.361	χ^2=2.23 p=.14
Cluster 4 vs. 1	1.475	0.805 – 2.703	χ^2=1.60 p=.21

Table 4.27 *– Crude estimates of the ORs with 95%CIs for the main exposure (PM10 concentration), clustered in 3 or 4 groups or in 2 groups (approximated as municipality). (*) Test for homogeneity of odds.*

4.7. MAIN MULTIVARIATE ANALYSIS

The effect of environmental exposure (PM10 concentration) and of other individual variables was evaluated by means of a logistic regression model, as described in ***3.6.5 Main multivariate analyses***. Three different models were implemented, with three differently coded variables as main exposure: the first with municipality (as an approximation of 2-clustered PM10 concentration), the second with 3-clustered PM10, and the third with 4-clustered PM10. The specifications are presented only for the final models, although a brief summary of the intermediate steps is provided.

4.7.1. Main exposure: municipality

Marital status (Q1), *educational level* (Q2), *type of house* (Q6), *physical activity* (Q10, recoded) and *alcohol consumption* (Q11), when added one at a time to the baseline model (*municipality, age*, *gender*, *lifetime cigarette smoking*), proved neither to contribute to the model's fit (LR test) nor to be informative (AIC and BIC). For this reason, these variables were not included in the final model.

Traffic in the streets near the subject's house (Q8 and Q9, combined) was not informative or contributive (AIC and BIC were essentially the same of the baseline model, and the LR test was not statistically significant). It is worth noticing, however, that if the estimation of the baseline model was restricted to exclude those in which *traffic* was missing, but without including the variable, the OR of the main exposure changed slightly. The variable, in any case, was not included in the final model.

The variable concerning *other diseases* was showed to be both informative and contributive respect to the model without the variable, so it was considered for inclusion in the final model.

The estimates of the parameters for the final model, specified with only *other diseases* added to the baseline model, are reported in ***Table 4.28***. The model was statistically different from a null model (LR=29.19, p<.0001). The environmental exposure went in the direction of evidencing a risk (the higher the exposure level, i.e. moving from Ferrera to Sannazzaro, the higher the odds of being exposed), but the effect was not significantly different from a null effect. Alike, the covariates that were forced in the model (gender, age and lifetime cigarette smoking) showed no statistically significant effect on the outcome of being a case or a control. Only diagnoses or treatments for other diseases came out to be statistically significant.

Variable		OR	95%CI	Test p-value

Municipality	*SdB vs. FE*	1.476	0.899 – 2.423	WT=1.54
				p=.12

Gender	*F vs. M*	0.809	0.513 – 1.276	WT=-0.91
				p=.36

Age (in 2014)	*+1 yr*	0.989	0.970 – 1.009	WT=-1.11
				p=.27

Lifetime smoking	*Yes vs. No*	0.940	0.598 – 1.480	WT=-0.27
				p=.79

Other diseases	*Yes vs. No*	3.754	2.156 – 6.538	WT=4.67 p<.001

Table 4.28 *– Mutually adjusted ORs (and 95%CIs) of environmental exposure (municipality), gender, age, lifetime cigarette smoking and other diseases. N=505.*

4.7.2. Main exposure: 3-clustered PM10

Analogously to the model showed in the previous paragraph, the implementation of the model with 3-clustered PM10 as main exposure led to discard almost all the selected covariates – *marital status* (Q1), *educational level* (Q2), *type of house* (Q6), *traffic* (Q8 and Q9, combined), *physical activity* (Q10, recoded) and *alcohol consumption* (Q11) – because they were neither contributive nor informative comparing to the baseline model, containing only the main exposure with *gender*, *age*, and *current smoking* (Q12.2). Only *other diseases* (Q17 and Q18, combined) was informative and contributive, and thus was retained in the model.

The estimated final model is reported in ***Table 4.29***; the LR test, also in this case, suggested that it performed better in fitting data than a null model (LR=28.16, p=.0001). The interpretation of the model with the 3-clustered exposure was, in every aspect, similar to that of the model with *municipality* as main exposure, with an indication that the odds of being exposed among cases were higher than in controls and a small difference in the OR between cluster 2 and cluster 3 (both contrasted to cluster 1).

Variable		**OR**	**95%CI**	**Test p-value**

Cluster	*2 vs. 1*	1.309	0.794 – 2.158	WT=1.06
				p=.29

	3 vs. 1	1.370	0.744 – 2.523	WT=1.01
				p=.31

Gender	*F vs. M*	0.798	0.506 – 1.257	WT=-0.98
				p=.33

Age (in 2014)	*+1 yr*	0.989	0.970 – 1.009	WT=-1.10
				p=.27

Lifetime smoking	*Yes vs. No*	0.938	0.596 – 1.476	WT=-0.28
				p=.78

Other diseases	*Yes vs. No*	3.755	2.155 – 6.542	WT=4.67 p<.001

Table 4.29 *– Mutually adjusted ORs (and 95%CIs) of environmental exposure (3-clustered PM10), gender, age, lifetime cigarette smoking and other diseases. N=505.*

4.7.3. Main exposure: 4-clustered PM10

In the model where the 4-clustered PM10 was specified as the main exposure, the results were substantially consistent with those illustrated for the two previous models. Again, models including as covariates *marital status* (Q1), *educational level* (Q2), *type of house* (Q6), *traffic* (Q8 and Q9, combined), *physical activity* (Q10, recoded) or *alcohol consumption* (Q11) were not informative or contributive in comparison with the baseline model, which included only the main exposure with *age*, *gender* and *lifetime cigarette smoking*. The only variable to stand out as contributing to improve the model's fit and informativity was the diagnosis or therapy for *other diseases*.

The final model (***Table 4.30***) fit data significantly better than a null model (LR=28.38, p=.0002). The interpretation of the model was basically identical to the two already reported in the previous paragraphs.

Variable		OR	95%CI	Test p-value

Cluster	*2 vs. 1*	1.411	0.500 – 3.983	WT=0.65
				p=.52

	3 vs. 1	1.326	0.795 – 2.213	WT=1.08
				p=.28

	4 vs. 1	1.420	0.755 – 2.671	WT=1.09
				p=.28

Gender	*F vs. M*	0.796	0.505 – 1.255	WT=-0.98
				p=0.33

Age (in 2014)	*+1 yr*	0.989	0.970 – 1.009	WT=-1.10
				p=.27

Lifetime smoking	*Yes vs. No*	0.943	0.599 – 1.484	WT=-0.26
				p=.80

Other diseases	*Yes vs. No*	3.747	2.151 – 6.527	WT=4.66 p<.001

Table 4.30 *– Mutually adjusted ORs (and 95%CIs) of environmental exposure (4-clustered PM10), gender, age, lifetime cigarette smoking and other diseases. N=505. (*) Wald's test*

4.7.4. Comparison of the models

The three models described in the previous paragraphs (***4.7.1***, ***4.7.2***, ***4.7.3***) only differed by the number of clusters used to classify the main exposure, which implies a different number of parameters for the models. The three models were compared by means of information criteria in order to evaluate if the increase in the parametrisation was worth. The results of the comparison are reported in ***Table 4.31***.

Model (main exposure)	*df*	AIC	BIC
Municipality (≈ 2 clusters)	6	531.91	557.26
PM10 concentration, 3 clusters	7	534.94	564.51
PM10 concentration, 4 clusters	8	536.72	570.52

Table 4.31 *– Comparison of the informativity of the three logistic models used to study the effect of three different ways of describing the main environmental exposure (PM10 concentration) taking into account other individual characteristics as covariates. Df: Degrees of freedom.*

4.8. SECONDARY MULTIVARIATE ANALYSIS

In a secondary analysis, the self-perceived general health status (Q16, dichotomised) was used as response variable in a logistic model to assess if it was somehow influenced by distance of the subject's house from the refinery and other variables. Analyses were performed as described in ***3.6.6 Secondary multivariate analyses***.

The baseline model included the effect of *distance* (in kilometres), controlled for *age*, *gender* and *case/control status*. Three additional covariates were tested, namely *traffic* (Q8 and Q9, combined), *current cigarette smoking* (Q12.2) and *physical activity* (Q10, recoded); the first two were not contributive or informative in the model, and therefore were excluded. On the other hand, *physical activity* was both informative and contributed, so that this variable was finally added.

The results from the final model are reported in ***Table 4.32***. The model is overall (borderline) better fitting data than a null model (LR=63.18, p=.099). Distance from the refinery had no significant effect on perceived health, controlling for the other variables; nevertheless, it is suggested that increasing distance tends to make a subject more positive about his or her general health. The results were confirmed also when the model implementation was repeated with the PM10-related main exposures and in the context of an ordinal logit model instead of a binary logistic regression.

Variable		OR	95%CI	Test p-value

Distance	*+1 km*	0.859	0.629 – 1.173	WT=-0.96
				p=.34

Status	*Case vs. Control*	2.859	1.809 – 4.518	WT=4.50 p<.001
Age (in 2014)	*+1 yr*	1.048	1.029 – 1.068	WT=4.90 p<.001

Gender	*F vs. M*	1.744	1.166 – 2.607	WT=2.71
				p=.007

Physical activity	*Yes vs. No*	0.562	0.373 – 0.846	WT=-2.76
				p=.006

Table 4.32 *– Mutually adjusted ORs (and 95%CIs) of environmental exposure (distance) on self-perceived health status (Q16, dichotomised), taking into account case/control status, gender, age and physical activity as covariates. N=497.*

5. Discussion

5.1. MAIN RESULTS

5.1.1. Characteristics of survey respondents

The distribution of gender is not significantly different between cases (59.2% males) and controls (54.3% males); likewise, no relevant differences in mean age have been found (in 2014, 58.1 years among cases and 56.8 years among controls).

Lifetime cigarette smoking is highly prevalent in both cases (58.2%) and controls (55.8%), with no significant difference between the two groups. Among cases, 58.6% declares to practise regularly at least one physical activity; the share is 64.2% among controls, with the difference being not statistically significant. Finally, 83.5% of cases and 59.3% of controls declares to have been diagnosed or treated for a condition that was coded under the same ICD-IX-CM causes that were selected to define cases.

In addition, even though the variable was not included in subsequent analyses, it is interesting to mention that the educational levels are lower than the national ones: more than half of the respondents declares a low-grade schooling, against a nationwide share of 61% of people with at least a high school diploma[92]. However, low-grade schooling is more prevalent among retired people (i.e. the elder ones, perhaps consistently with the trend of increase in schooling started in the post-war period), and retired people are 45.6% of the respondents among cases and 41.6% among controls.

5.1.2. Environmental exposure among survey respondents

Regarding the environmental exposure, the distribution is bimodal both for the distance and for the PM10 concentration predicted by the AERMOD model. Cases show a greater exposure level to PM10 than controls. After clustering PM10 concentration, a trend of PM10 increasing with cluster ranking is found both in cases

[92] Referred to resident citizens aged 20-64. See *https://www.istat.it/it/archivio/219264* (opened on August 15th, 2018).

and controls. However, none of the clusters is significantly associated with the outcome.

It has to be remarked that predicted PM10 concentration attributable to the refinery (order of magnitude 10^{-1} µg/m^3) are substantially lower than total PM10 concentration (order of magnitude 10 µg/m^3).

5.1.3. *Health effects of exposure to the emissions from the refinery*

Crude ORs for the environmental exposure give an indication in terms of a possible excess of "risk" connected to the exposure to PM10: the odds of being exposed are roughly 60% higher in Sannazzaro de' Burgondi than in Ferrera Erbognone[93] (OR=1.595), and roughly 40-50% higher in any cluster compared to the lowest with either 3-clustered PM10 (cluster 2 vs. 1: OR=1.427; 3 vs. 1: OR=1.427) and 4-clustered PM10 (cluster 2 vs. 1: OR=1.456; 3 vs. 1: OR=1.448; 4 vs. 1: OR=1.475). Anyway, none of the ORs is statistically significant: the results do not allow to exclude the hypothesis of a null effect (i.e. OR=1) in any of the estimations. These results are similar to those that were obtained on the full sample of the 1046 enrolled subjects (unpublished data).

In the multivariate analyses that were performed, the effects of the environmental exposures remain essentially unchanged with respect to the germane crude estimates, whatever the variable used as main exposure. The main effect is controlled for gender, age and lifetime smoking, which inclusion was decided *a priori*, although none of those variables was proven to have a significant effect or was informative or contributive in the model. Only one covariate has been included in the model because it was informative, contributive and significant: having received a diagnosis or a treatment for other diseases that are potentially compatible with the causes of hospitalisation of cases. Subjects with such diseases are nearly 4 times more at "risk" than those without the same diseases; however, it should be noted that part of those who declared to have the diseases are the cases, so that the reason why cases are more likely to be "exposed" to this factor (i.e. to have an higher odds than controls of having those diseases) is somehow clear-cut, as it should be the reason for its inclusion.

[93] It is worth recalling that municipality was used as a *proxy* of 2-clustered PM10 concentration, as the two variables were almost perfectly concordant, and that exposure levels were higher in Sannazzaro de' Burgondi than in Ferrera Erbognone.

Other factors have been evaluated as potential covariates in the models[94], but they are not included in the final specifications because the criteria defined for informativity and contributivity were not met.

It should be highlighted that the models are extremely stable, whatever the variable used to describe the main exposure, and that adding covariates do not change relevantly the effect estimates of the main exposure. Moreover, a comparison of the three final models (with municipality, with 3-clustered PM10 and with 4-clustered PM10) shows a comparable informativity for all of them, with AIC and BIC slightly lower for the model with municipality, i.e. the less parametrised one: this seems to confirm that, with regards to the exposure to the emissions from the refinery, what matters more is the difference between Sannazzaro de' Burgondi (leeside of the stacks) and Ferrera Erbognone (upwind).

A secondary multivariate analysis investigated how the self-rated perception of health was influenced by various factors, among which the presence of the refinery, measured in terms of the distance from the house to the refinery[95]. The effect of distance is not significantly different from a null effect, albeit the odds of being closer to the refinery are reduced in those who have a more positive attitude towards their wellbeing with respect to those having a negative attitude (OR=0.859 per unit increment of distance in kilometres). Yet, it should be taken into account that the distribution of distance could have affected the estimates. The perception of one's health status is (obviously) negatively influenced by being a case rather than a control (OR=2.859), but the reason for that should be straightforward. In addition, females tend to have a worse attitude than males (OR=1.744) and also age negatively affect perceived health, even though the effect per each increment of one year is quite small (OR=1.048). Finally, also those who declared to practise regularly at least one physical activity are substantially more positive about their health (OR=0.562).

[94] *Marital status, educational level, type of house, traffic in the streets near the house, physical activity, alcohol consumption.*

[95] Distance was chosen because, to study a perception of the individual, it seemed closer to what a person can perceive rather than PM10 concentration.

5.1.4. Potential confounding bias

In the case of the present study, the main exposure does not actually have any significant effect on the outcome of being a case rather than a control, which could suggest that one of the sides of the "*confounding triangle*" [*KELSEY ET AL, 1986*][96] is actually suppressed. However, if a confounding bias exists, it could even result in masking the effect of the main exposure: in other words, confounding must be assessed anyway. In addition, as will be commented in the next paragraphs of this Chapter, the absence of a significant effect of the main exposure can have further explanations (i.e. reduced statistical power, selection bias, information bias) (see ***5.3 Strengths and Limitations***).

Findings from bivariate analysis underline that gender is significantly associated with municipality, but statistical significance is lost when using the 3- or 4-clustered exposure (probably because the same number of statistical units is split over more categories – and this consideration should be taken into account also for all the other factors commented in this section). Moreover, this factor is balanced[97] *by design* between cases and controls and, thus, it is not associated with the outcome. Likewise, age does not seem to act as a confounder: it is associated with the municipality and 3-clustered exposure (but not with the 4-clustered one), but its association with the outcome is prevented *by design*.

Marital status is borderline associated[98] with municipality and 4-clustered exposure, with the differences among the exposure groups seemingly due to the proportion of singles and married persons. Anyway, this factor do not show a convincing association with the outcome; even in the case, evidence supporting the role of marital status as a risk factor is mounting [e.g. *WEISS, 1973*; *MANFREDINI ET AL., 2017*], but its role as a

[96] *Confounding* occurs when, in the context of a (causal) relationship between an exposure and a certain outcome, another factor is associated to – but is not a consequence of – the main exposure and is itself also (causally) related to the outcome [*KELSEY ET AL, 1986*]: this is often referred to as the "*confounding triangle*". Otherwise, the factor would be a *mediator* rather than a confounder [*MACKINNON ET AL., 2000*]. The presence of confounding can either drown out or enhance the estimated main effect [*MACKINNON ET AL., 2000*]. Confounding bias can be prevented *by design*, i.e. by matching for those factors that are thought to potentially act as confounders, or *a posteriori* during the analysis, either by stratification or by applying multivariate methods (like logistic regression). See also ***Appendix B.5.1***.

[97] As stated in ***3.2.4 Sample selection***, the enrolment of cases and controls was balanced for some potential confounders.

[98] Again, the lack of a strong statistical significance could be due to a reduced study power.

determinant is less clear, albeit various hypotheses have been proposed (see, for instance, the Discussion in *MOLLOY ET AL., 2009*).

The association between educational level (maximum grade achieved) and the exposure was borderline when considering municipality and 3-clustered exposure, but it disappeared with 4-clustered exposure; moreover, excluding the 10 subjects who marked the option "other" in Q2 tended to smooth the difference between exposure groups (data not shown). A similar situation was observed when testing educational level against the outcome. In addition, in a causal pattern, educational level could be more a precursor than a direct cause of the outcome, for instance because it may determine differences in the profession or differences in the capacity to get access to healthcare.

Regarding the consumption of alcoholic beverages, this variable was borderline associated with the clustered PM10 exposures, but not with municipality; analogously, lifetime cigarette smoking was significantly associated with municipality and 3-clustered exposure, and the association remained borderline significant also with the 4-clustered exposure to PM10. Neither alcohol consumption nor lifetime cigarette smoking were significantly associated to the outcome of being a case or a control. However, the role of cigarette smoking as a causal factor of diseases (including those considered in the definition of case) has been widely recognised in countless studies, so anyhow this variable was included in multivariate analyses.

Finally, traffic levels near the subjects' houses, physical activity and diagnosis or treatment for other diseases were not associated with exposure and, thus, residual confounding due to these variables can be excluded.

5.2. COMPARISON WITH PREVIOUSLY PUBLISHED EVIDENCE

Prior to discussing any comparison with published literature, it should be made clear that the estimated risks for human health due to PM10 strongly depend on the way PM10 exposure is assessed; therefore, results from different studies can be compared for the "direction" of the disclosed effects – risk or protection – but putting together point estimates that arose from different ways of exposure classification is practically unfeasible. In addition, similar issues may come from the definition of outcome and, for what concerns PM10 specifically, on variability in the composition of the particles (e.g. because of the diversity in the type of sources from an area to one other) that can result in different health effects.

That said, the indication of an excess of risk with the increase in the level of exposure to PM10 from the refinery (albeit out of statistical significance) is consistent with most of the existing epidemiological literature, and according to toxicological evidence it is also plausible in terms of the biological pathways behind that risk. In fact, the evidence indicating that exposure to air pollutants (including PM10) is a risk factor for human health is large and robust for both short- and long-term exposure, also with hard outcomes like severe illnesses and mortality due to cardiovascular and respiratory diseases [*WHO, 2006*]. Notably, outdoor air pollution has been classified as carcinogenic to humans (*IARC Group 1*), thanks to the remarkably consistent evidence from epidemiological and toxicological research [*IARC, 2016*].

A review regarding the available evidences of the health impacts of PM10, updated as of 2004, was promoted by the US EPA in its document on air quality criteria [reported in *WHO, 2006*]; some published studies has been discussed also in other reports [e.g. *NAS, 2017*].

5.2.1. Toxicological evidence

Broadly speaking, the effects of PM10 have been investigated in lots of research, based both on animal models and on studies on humans (including induced exposure in volunteers). Several biological mechanisms have been identified to explain the negative effects of PM10 exposures; the mechanisms, together with the effects, may depend on the physical presence of the particles and/or on the action of the chemicals carried by the particles. Inflammation has a central role in the body's response; other pathways (partly coming as a consequence of inflammation) are impairment of pulmonary defences, exacerbation of pre-existing conditions, genotoxicity[99].

5.2.2. Epidemiological evidence

A vast literature of epidemiological studies regarding the exposure to particulate matter has mounted in the last two decades. In general, evidence of negative effects of PM10 exposure is strong for cardiovascular and respiratory diseases, which are two of the three macro-causes included in the definition of "case" for the study presented in this thesis; less evidence is available with regards to gastrointestinal diseases, albeit some studies identified excesses of risk and proposed mechanisms to explain them [e.g. *BEAMISH ET AL., 2010*; *ORAZZO ET AL., 2009*]. The indication of PM10 from the refinery being a risk factor, in our study, looked collectively at cardiovascular, respiratory and gastrointestinal causes; albeit a sensitivity analysis confirmed the results after excluding those cases admitted to hospital for a gastrointestinal condition, limitations in power of statistical tests (because sample size is reduced in each stratum) prevented us from repeating the analyses separately by cause.

With regard to short-term exposure, several metanalyses of daily time-series studies demonstrated a positive association between the concentration of particulates and mortality by any cause or by various specific causes (for instance, the excess of mortality was higher when looking at respiratory or cardiovascular diseases) [APHEA2 and NMMAPS, reported in *WHO, 2006*]. These results were substantially consistent (in spite of some negative findings) with cohort studies regarding long-term exposure, in which exposure level was assigned on the basis of the home address and the outcome was mortality [studies reported in *WHO, 2006*].

Lots of researches also looked at a wide range of different indicators related to morbidity, including hospital admissions, with different approaches such as time-series, panel studies, cohort studies, case-crossover studies and cross-sectional studies. Concerning hospitalisation, there is robust evidence supporting the increased risk of

[99] A detailed review of toxicological literature is out of scope for the present thesis; it can be found in various publications, e.g. in *IARC, 2016*; *WHO, 2006*; *BERNSTEIN ET AL., 2004*.

being admitted to a hospital for cardiovascular or respiratory conditions, including acute manifestations [studies reported in *WHO, 2006*].

The results from the present study are consistent with the findings of epidemiological research regarding industrial areas in Italy, with particular reference to refineries and petrochemical plants. An ecological study investigated deaths and hospitalisations – identified with the ReNCaM and the SDOs databases, respectively – in the resident population of several contaminated sites in Sicily (Gela, Priolo and Milazzo for petrochemical plants, refineries and other sources; Biancavilla for natural asbestos exposure). The researchers found excesses of mortality and morbidity rates by cancer and chronic pulmonary diseases in the area of Priolo; in Gela, there were excesses of mortality by cancer and morbidity by respiratory and cardiovascular diseases (among others); the risks were less relevant in Milazzo, with an increase of cardiovascular diseases in men and of respiratory diseases in women. However, it was recognised that differences in hospital admissions might be due to a different attitude of local practitioners and health personnel in hospital departments [*CERNIGLIARO ET AL., 2006*]. These results were consistent with another study carried out in the area of Gela, which found similar results for morbidity and mortality, with relevant differences in the excesses of mortality and morbidity for the less lethal conditions (particularly for respiratory diseases) [*FANO ET AL., 2006/A*]; consistent results were also observed in the area of Civitavecchia, which hosts a cement factory, several power plants and a harbour [*FANO ET AL., 2006/B*]. A study made among the workers of the refinery in Gela found a possible excess in mortality rates for the workers that used to live closer to the plant, after correcting for confounders like age and job (white or blue collar) [*PASETTO ET AL., 2008*].

5.3. Strengths and Limitations

5.3.1. Issues concerning case definition

The occurrence of the health outcome of interest was defined from the data stream of the healthcare system. This removes any chance of information bias due to erroneous recalling of past events by the subjects (i.e. *recall bias*), or any other sort of misconception, which could ultimately result also in a *misclassification bias* of cases and controls.

However, as already commented (see ***3.2.3.3 A focus on the Hospital Discharge Records***), the databases of the healthcare systems are primarily intended for administrative purposes – essentially surveillance on the functioning of the healthcare system and computation of the reimbursements owed to the facilities – and this can make SDOs unreliable for epidemiological research, introducing information bias. As stated by the Ministry of Health itself[100], the Hospital Discharge Records (SDO) might indeed be coded inhomogeneously [101] or, for some variables, inaccurately or incompletely.

Another limitation might come from the definition of case (as described in ***3.1.1 Case definition***), which is based only on the main cause of the first hospitalisation, as recorded in the SDO database. Looking at the first hospitalisation, obviously, has the limit that it disregards "changes" in the patient's health status. In addition, for any hospitalisation, the main cause is defined on an economical basis: it is the condition that implied the highest share of the spending, but this does not necessarily mean that it is the most relevant from a clinical perspective[102].

In addition, cases were defined as those who were hospitalised between 2002 and 2014 but were still alive at the end of 2015, and this might have determined an underestimation of the role of the environmental exposure, because of additional issues in terms of selection bias:

[100] See salute.gov.it/portale/temi/p2_6.jsp?id=1232&area=ricoveriOspedalieri&menu=vuot (opened on August 16th, 2018).
[101] It should be remembered that SDOs are filled in by the physician when the patient is discharged; there are more than 11 thousand diagnosis codes.
[102] See http://www.epicentro.iss.it/focus/sdo/pdf/2_deCampora.pdf (opened on August 16th, 2018).

- from one side, selecting only hospitalised cases might have excluded those with a less severe condition and, if they had no hospitalisations in the reference timespan, they could have been included as controls (although it should be reminded that, on this regard, two questions – Q17 and Q18 – were asked in the questionnaire);
- from another point of view, the most-severe cases might have been excluded due to death before the survey was conducted[103].

5.3.2. Issues concerning environmental exposure

5.3.2.1. Advantages and limitations of modelling

The choice of modelling the main exposure has an important advantage: it allows the attribution of an individual exposure level, which could otherwise be achieved only with individual monitoring and skyrocketing research costs[104].

For sure, this choice presents limitations respect to real-world measures, because modelling is by definition an approximation of reality[105] and itself has some limitations, affecting the extent to which its predictions can be trusted; anyway, limitations can be controlled if the model is chosen wisely. The AERMOD approach has been introduced in ***Appendix C*** and, as stated there, it is suitable for the sake of estimating ground-level concentrations of various pollutants. The reliability of its predictions can be affected by factors like low wind – which could be the case for the Po Valley. It should be said, anyway, that the AERMOD model was considered the "best shot" at the time when it was realised; indeed, it was – and, still, is – the model of choice for regulatory purposes in the United States and it was accepted by the Italian authorisation bodies in the context of the AIA-VIA procedure for the EST facility of the refinery.

[103] Anyway, oncological diseases were not included in the definition of case and were likely excluded from the definition of controls, because hospitalisation by any non-traumatic cause prevented a subject to be selected as control.

[104] "Global" measurements, for instance those by PM10 measurement stations, are generally too scarcely distributed in space to emphasise inter-individual differences, if the individuals are taken from the same geographical area.

[105] This said, it is also true that dispersion modelling takes into account, at least in part, the real world, and that its predictions are validated by years of usage.

5.3.2.2. Characteristics of exposure distribution

A relevant point is that exposure levels are extremely associated with the municipality, and within each municipality there is a low between-subjects variability in the estimated PM10 concentration[106]. This aspect, which was not expected when the study was designed, has been later explained by the fact that in the area there is only a low-speed but rather constant wind from north-west to south-east, so that Sannazzaro de' Burgondi is leeward to the refinery's stacks, while Ferrera Erbognone is essentially upwind. As cases and controls were balanced by municipality, this implies the effect of the main exposure could be underestimated and the OR will be closer to 1 than it should (biased towards a null effect). In addition, this has another implication (that will be discussed later) with the survey respondence.

5.3.3. Issues concerning the use of survey data

The use of self-administered questionnaires as a mean for data collection is for sure advantageous, being an efficient method to gather a lot of information from a high number of subjects at lower costs. Standardised self-administered questionnaires are often used in environmental and occupational epidemiology to assess main exposure or other factors [*NIEWENHUIJSEN, 2005*]. However, this method can be less accurate than others: subjects might misunderstand the questions, their answers might be affected by *recall bias*, or the questions (and, more relevantly, the answers) may not relate as expected with the exposures they are designed to investigate. This can lead to various forms of bias and, consequently, to unreliable conclusions from the study.

A first problem, in the present study, is that the survey asked the subjects about their status, habits and conditions at the time of the survey, while the exposure of interest was in the past. Whilst some traits are substantially stable over time – e.g. lifetime cigarette smoking – most of the others may theoretically undergo sudden changes. This, albeit limiting the chance of *recall bias*, can introduce *information bias* due to the relevance of the data in relation to time. Other issues are discussed in the next sections.

[106] It is also worth mentioning that the levels of PM10 attributed to the refinery's emissions is, both in Sannazzaro de' Burgondi and in Ferrera Erbognone, magnitudes below the total concentration of this pollutant (annual mean) in the area. In 2014, the measurement station of Sannazzaro de' Burgondi (area: urban, industrial) recorded a mean PM10 concentration of 28 μg/m^3 (limit: 40 μg/m^3), with 34 days in which it exceeded the limit of 50 μg/m^3 (limit: 35 days) [see *ARPA-L, 2014*].

5.3.3.1. Concerns regarding the questionnaire

A more than legitimate cause of concern, regarding potential information bias, is the use of non-validated questionnaires, as this means there is no actual guarantee that the questionnaire measures what it is meant to measure [*NIEWENHUIJSEN, 2005*][107]. Regarding the study presented in this thesis, the survey had a minor risk of a *validity bias*. The questionnaire is partly based on tools used in previous international studies [*http://www.ecrhs.org/*]. It mainly consists of isolated questions collecting semantically separated information, instead of a series of questions all related to the evaluation of a same trait. Moreover, the estimated time needed to fill it in (around 15 minutes) is reasonable and helps avoiding the "*yah-saying*" (or "*nay-saying*") *bias* [*NIEWENHUIJSEN, 2005*]. Anyway, biases like an extreme-response style, the "*faking-bad*" or social desirability cannot be excluded, and preliminary analyses of survey data might suggest these have indeed occurred in some questions (for instance, social desirability might have inflated the self-ratings given by subjects to their compiling in Q19, and extreme-response and faking-bad might have biased those questions related to alternative exposures than the main exposure because of subjects' inner beliefs about the effects of the refinery).

Most of the items included in the questionnaire are close-ended questions: on this regard, it must be taken into account that what can be seen from close-ended answers, e.g. those providing frequency ratings, descends strictly from the categories as they were defined in the designing phase. In addition, some questions (Q8, Q9, Q10) used by non-objective terms to define frequencies (e.g. "frequent" or "constant"), and different subjects may assign to these terms different thresholds.

A possible solution to reduce the risks described above is a pilot testing of the questionnaire. This was made, even if the tool was administered only to a convenience sample (colleagues and relatives of the investigators), and not on a representative random sample of the target population: this might have reduced the efficacy of this preliminary test [*NIEWENHUIJSEN, 2005*].

[107] In the paper by *NIEWENHUIJSEN* [*2005*], it is made clear that validating a questionnaire for exposure assessment purposes could be difficult, if possible at all, especially when the questionnaire is designed to measure a complex exposure with a conjunct of questions (resembling, to a certain extent, the issues with psychometric questionnaires).

The analyses regarding the consistency of data casted doubts about the average comprehension of some of the questions and instructions (e.g. Q10, Q11, Q15, Q17, Q18[108]), even when they seemed particularly straightforward in the design stage (for instance, Q12). This might be attributed to a lack of clarity, to the unexpectedly low educational level of enrolled subjects, or both. The complex of issues mentioned here implied the need to reject the use of some of the survey's data and to recode some of the questions, for instance by grouping together the ratings: this approach, albeit reducing the impact of non-objective labels, could result in a misclassification bias (in addition to the loss of information).

Finally, "*last but not least*", another potential issue when data are collected from paper questionnaires comes from mistakes during the data entry process. This possibility has been minimised by different strategies: checks in the entry process, which prevented to upload in the database inconsistent answers, and inspection *a posteriori* to look for weird values (when possible). In the most likely case, this type of error would have been stochastic.

5.3.3.2. Interpreting respondence

Given that a certain response rate is acceptable to the extent that it is adequate for the study question and for the target population, and no "general threshold" would really make sense[109], a general – and quite obvious – rule is *the higher, the better*.

A systematic review based on papers published in 1991 on US peer-reviewed medical journals, regarding studies based on mailed surveys, found that 192 articles reported information regarding response rate and it was on average 59%, with high variability; interestingly, in their findings, anonymous questionnaires are counter-correlated with respondence [*Asch et al., 1997*]. On the other hand, previous experience with a survey study on the environmental determinants of asthma (the *European Community Respiratory Health Survey*, ECRHS), conducted in the Province of Pavia, suggested

[108] In all these questions, asking to rate a frequency or a presence/absence for a series of items, during data entry a possible compiling pattern came to light. It was observed that various subjects might have answered only to those items that were applicable for them, e.g. in Q10 (physical activities) they might have rated only the frequency of those activities they were actually playing, ignoring the option "never" for the activities they did not use to play.

[109] A respondence around 85% would probably be more than acceptable to estimate a prevalence of 50%, but not of 5%. See *https://www.bmj.com/about-bmj/resources-readers/publications/epidemiology-uninitiated/5-planning-and-conducting-survey* (opened on August 15th, 2018).

that a response rate around 70-80% could be achieved, and on this basis the estimated sample size was adjusted for non-respondence. The global response rate reached in this study was below the *a priori* expectations: 54.6%.

Generally speaking, people might not participate in a survey for two different reasons. The first reason could be that no valid address is available to deliver the questionnaire but – thanks to the preliminary work with the Municipal Registries – this circumstance was minimised in the here-presented case. The second reason, if the subject has regularly received the questionnaire, is a refusal to participate, either implicit or explicit.

Meetings with the representatives of the residents of Sannazzaro de' Burgondi and Ferrera Erbognone, as well as comments received by the general practitioners and sometimes directly by the subjects[110], allow to advance some hypotheses to explain this substantial lack of respondence. A reason, apparently, is a lack of trust in the CONSAL Project, probably because the study is commissioned by the company running the refinery which, in people's opinions, might "influence" the results of the study towards certain findings.

In addition, criticism was raised by some local health professionals against the exclusion of oncological conditions from the definition of case. Finally, a lack of confidence regarding privacy was pointed out as well; on this regard, a statement with a detailed explanation of how privacy was dealt with was released during the first "wave" of contacting, but this did not seem to result in a substantial improvement. For both these reasons, perhaps, a better project communication would have helped solving – at least in part – the problem[111]; a deeper dissertation of this aspect is provided later in the Chapter (see ***5.4 Project communication: what could have we done?***).

Finally, even though this has not been formally confirmed and should be taken cautiously, another possible reason for non-respondence could be the worry of losing the job in the eventual case the refinery will be found to be a danger for public health, and thus compelled to reduce its production rate or even to end its activity. The refinery

[110] Several subjects, while returning their questionnaire, attached comments that we received.

[111] A nice read about communicating science to build trust was published by the Journal of Science Communication at *jcom.sissa.it/archive/15/05/JCOM_1505_2016_C00* (opened on August 17th, 2018).

is indeed a major economical player in the area and, among the enrolled subjects and/or their relatives, a lot are likely employed at the refinery.

5.3.3.3. Concerns regarding respondence in the present study

The lack of respondence recorded in the present study may alter the results in different ways. For sure, it may have determined a loss in terms of study power, so that non-significant results regarding the effects of the environmental exposure should not be interpreted as a conclusive evidence of a real lack of effect.

Anyway, the analyses of respondence, presented in ***4.1 Analysis of respondence***, suggested that the response rate has not been influenced by case/control status and gender, and only age was marginally different by respondence status. So, it may be concluded that respondents and non-respondents are comparable and *non-respondence bias*, a specific type of *selection bias* [*ASCH ET AL., 1997*], likely has not happened (at least for the variables that have been checked). In the context of a case-control study, this bias would be critical if occurring differentially in cases and controls or in relation to the exposure under investigation[112], and this could actually have happened in the present study. Indeed, respondence is strongly different between the enrolled subjects of Sannazzaro de' Burgondi and Ferrera Erbognone. Taking into account that municipality is associated with the environmental exposure as well (i.e. municipality itself was basically interpretable as exposure), this is likely to have imposed a bias, especially as a result of the combined effect of this and of the slight difference in case:control ratio between the two municipalities[113]. The reason for that can be easily understood by thinking at the OR formula, detailed in ***Appendix B.4***. Let's consider, for instance, the simplest case of a dichotomic exposure (i.e. the municipality of residence, which– as explained elsewhere in the thesis – is roughly the same as a 2-clustered variable for the individual-level PM10 concentrations): as there are more respondents in Ferrera than in Sannazzaro, the number of unexposed people – i.e. those from Ferrera – appears to increase; moreover, among the unexposed, the

[112] A selection bias, in case-control and in cohort studies, could also result from information bias (*misclassification*), respectively regarding outcome or exposure. In this case, however, cases and controls were defined regardless of the questionnaire.

[113] Indeed, the case:control ratio of the overall sample of respondents is slightly higher than the same ratio calculated on all the enrolled subjects, because the share of respondents is higher in Ferrera (where the ratio was higher as well).

case:control ratio was higher, and this could slightly inflate the numerator of the OR (because the number of unexposed controls is increased as a result of selection bias).

5.4. PROJECT COMMUNICATION: WHAT COULD HAVE WE DONE?

The results of the survey, if respondence and comments received by stakeholders are put together, seem to indicate a general distrust for institutions and experts, which is a widely discussed topic in sociology and science communication and affects both the implementation of the study and the possibilities in terms of transfer of knowledge and translation of scientific evidence into informed choices.

The problem of distrust goes far beyond the lack of specialised knowledge among non-experts[114], and also includes *confirmation bias*[115] and the consequent tendency towards *motivated reasoning*, i.e. "*the additional tendency to defensively reject information that contradicts deeply held worldviews and opinions*"[116]. Some sociological surveys conducted in other Countries (Germany, Sweden and Switzerland), for instance, found the distrust in scientists' motives higher than the distrust in science as a system, and in Germany 76% of respondents stated that funding is one reason for distrusting scientists[117].

In the words of the Ottawa Declaration of 1986, "*health promotion is the process of enabling people to increase control over, and to improve, their health*" [*DE CASTRO ET AL., 2016*]. There is indeed a mutual influence between science and society and, in years, citizens started to exert a more active role and to call for their right to be directly informed and engaged in the exchange with science of the society they are a part of [*DE CASTRO ET AL., 2016*]. In the context of investigations on contaminated sites, it is straightforward how this implies a reciprocity in the relationship between researchers specialised in health and environmental sciences and all the other *stakeholders* (public health officers, risk managers, policy makers, population, environmentalist

[114] See, for instance, *www.sciencealert.com/researchers-have-figured-out-what-makes-people-reject-science-and-it-s-not-ignorance*, or *www.psychologytoday.com/us/blog/intentional-insights/201807/distrust-in-science* (opened on August 17th, 2018).

[115] "*Confirmation bias, as the term is typically used in the psychological literature, connotes the seeking or interpreting of evidence in ways that are partial to existing beliefs, expectations, or a hypothesis in hand*" (R. S. Nickerson, 1998).

[116] See *https://www.scientificamerican.com/article/how-to-combat-distrust-of-science/* (opened on August 17th, 2018), where a study about a "treatment" for motivated reasoning can be found, together with a nice reflection about the social distrust of science and scientists, referred in particular to the US society.

[117] Data are reported and commented in *https://www.euroscientist.com/trusting-science-age-distrust/* (opened on August 18th, 2018).

organisations, patients' organisations). Such a relationship requires an appropriate framework, consisting of *stakeholder* engagement and dissemination activities based on adequate tools ensuring that any obstacle is removed from the communication process [*WHO, 2013*], and in the involvement of communities in research with a participatory approach [*Soskolne, 2016*; *Comba et al., 2016*]. These activities may help the researchers in making any assessment (and any proposal of intervention) more effective [*De Castro et al., 2016*; *Marsili, 2016*]. Moreover, it may also give support in building or reinforcing the trust in experts and institutions by making stakeholders less sceptical about data quality and more prone to accept the results, even if they are not in line with the population's perception of risk [*WHO, 2013*]. At the same time, it is well recognised that risk communication dynamics are challenging and might get out of hand even when well-planned and controlled [*De Castro et al., 2016*].

In other words, a direct involvement in the aforementioned dynamics might help in making the researchers trusted by the population, thus being pivotal for the success of a study. In fact, this dual relation facilitates the meeting of stakeholders' questions with the scientific rigour that is required to elaborate and answer those questions, and should comprehensively include scientific evidence and information needs of the different *stakeholders* [*De Castro et al., 2016*]. In this sense, the study presented in this thesis – as well as the Project it belongs to – is lacking in the implementation of a communication plan. In fact, even if different partners (such as town councils, ATS Pavia, ARPA Lombardia) are involved in the Project, the actions that has been taken failed to achieve adequate levels of compliance and to disseminate appropriately the progresses of the Project to the population.

In addition, a further aspect should be considered: studying environmental risk factors is indeed necessary, but not sufficient to address decisions. A public health study should be better completed with impact estimates aimed at assessing how relevant the risks are in the specific context of the target population: this is needed by stakeholders and policymakers to identify the priorities of intervention and represents the step from research as mere *information* to *information for action* [*WHO, 2013*].

6. Conclusions

The results presented and discussed in this doctoral thesis indicate a possible excess of hospitalisation risk among people living in Sannazzaro de' Burgondi in comparison with the near Ferrera Erbognone, due to the different exposure to pollutants emitted from the refinery (which turned out to fall with a substantially higher concentration on the former municipality). The crude estimates of risks are similar to those obtained after controlling for other covariates. The results are consistent with previous literature; however, they cannot be taken as conclusive evidence, because a null effect cannot be excluded – perhaps as a consequence of the reduction in study power due to the survey's response rate, which was lower than expected – and because the "risk" estimates may have been affected by bias. Still, they can be less relevant in the wider context of the CONSAL Project, in which they will be complemented with results from different epidemiological approaches.

The study shows that the predicted air concentrations of other pollutants are strongly correlated with those of PM10 in the fallout domain of the refinery's emissions, so that the observed effects might be due to the combined effect of different pollutants. The fallout of stack-emitted contaminants was modelled by means of the Gaussian AERMOD model, and individual exposure to PM10 from that source was assigned to each subject only on the basis of his or her registered home address. In the future, it might be interesting to use a different modelling approach – like a Lagrangian method, which may be more reliable with low-speed wind conditions – and eventually re-run the epidemiological analyses with contamination levels predicted by this model, in order to compare the results.

In addition, particularly in the context of the CONSAL Project, other two aspects might worth consideration for future work. First, integrating evidences from the other CONSAL studies might be helpful in order to identify the priorities for health management in this territory and to evaluate and define appropriate intervention plans together with the stakeholders. Second, committing to the study of the effects that living in a contaminated site may have on psychological wellbeing, which is a less investigated topic, could be intriguing.

REFERENCES
(ordered by first author and year)

Ackermann-Liebrich, U., Leuenberger, P., Schwartz, J., Schindler, C., Monn, C., Bolognini, G., ... & Grize, L. (**1997**). Lung function and long term exposure to air pollutants in Switzerland. Study on Air Pollution and Lung Diseases in Adults (SAPALDIA) Team. *American Journal of Respiratory and Critical Care Medicine*, *155*(1), 122-129. DOI:10.1164/ajrccm.155.1.9001300.

Agenzia Regionale per la Protezione dell'Ambiente della Lombardia (**ARPA-L**) (edited by: Mognaschi, G., & Carli, P.). (**2015**). Rapporto sulla qualità dell'aria della Provincia di Pavia. Anno 2015.

Agenzia Regionale per la Protezione dell'Ambiente della Lombardia (**ARPA-L**) (edited by: Mognaschi, G., Carli, P., & Guarnaschelli, G.). (**2014**). Rapporto sulla qualità dell'aria della Provincia di Pavia. Anno 2014.

Asch, D. A., Jedrziewski, M. K., & Christakis, N. A. (**1997**). Response rates to mail surveys published in medical journals. *Journal of Clinical Epidemiology*, *50*(10), 1129-1136.

Bacchieri, A., & **Della Cioppa, G.** (**2004**). Fondamenti di ricerca clinica. Springer-Verlag Italia, Milano 2004. ISBN: 8847002117.

Barregard, L., Holmberg, E., & Sallsten, G. (**2009**). Leukaemia incidence in people living close to an oil refinery. *Environmental Research*, *109*(8), 985-990. DOI:10.1016/j.envres.2009.09.001.

Beamish, L. A., Osornio-Vargas, A. R., & Wine, E. (**2011**). Air pollution: An environmental factor contributing to intestinal disease. *Journal of Crohn's and Colitis*, *5*(4), 279-286. DOI:10.1016/j.crohns.2011.02.017.

Benedetti, M., Lavarone, I., & Comba, P. (**2001**). Cancer risk associated with residential proximity to industrial sites: a review. *Archives of Environmental Health: An International Journal*, *56*(4), 342-349.

Bernstein, J. A., Alexis, N., Barnes, C., Bernstein, I. L., Nel, A., Peden, D., ... & Williams, P. B. (**2004**). Health effects of air pollution. *Journal of Allergy and Clinical Immunology*, *114*(5), 1116-1123. DOI:10.1016/j.jaci.2004.08.030.

Bertoldi, M., Borgini, A., Tittarelli, A., Fattore, E., Cau, A., Fanelli, R., & Crosignani, P. (**2012**). Health effects for the population living near a cement plant: An epidemiological assessment. *Environment International*, *41*, 1-7. DOI:10.1016/j.envint.2011.12.005.

Biggeri, A., Bellini, P., & Terracini, B. (**2001**). Metanalisi italiana degli studi sugli effetti a breve termine dell'inquinamento atmosferico. *Epidemiologia & Prevenzione*, *25*(2), 1-72.

Bluett, J., Gimson, N., Fisher, G., Heydenrych, C., Freeman, T., & Godfrey, J. (**2004**). Good Practice Guide for Atmospheric Dispersion Modelling. *Ministry of the Environment – New Zealand*, Wellington 2004. ISBN: 0478189419.

Briggs, D. J. (**2008**). A framework for integrated environmental health impact assessment of systemic risks. *Environmental Health*, *7*(1), 61. DOI:10.1186/1476-069X-7-61.

Budroni, M., Sechi, O., Cesaraccio, R., Pirino, D., Fadda, A., Grottin, S., ... & Tanda, F. (**2010**). Cancer incidence among petrochemical workers in the Porto Torres industrial area, 1990-2006. *La Medicina del Lavoro*, *101*(3), 189-198.

Bulat, P., Avramov Ivić, M. L., Jovanović, M. B., Petrović, S. D., Miljuš, D., Todorović, T., ... & Bogdanović, M. (**2011**). Cancer incidence in a population living near a petrochemical facility and oil refinery. *Collegium Antropologicum*, *35*(2), 377-383.

Burnham, K. P., & **Anderson, D. R.** (**2002**). Model selection and multimodel inference. A practical information-theoretic approach. 2nd edition. *Springer-Verlag*, New York 2002. ISBN: 0387953647.

Campbell, M. J., & **Swinscow, T. D. V.** (**2009**). Statistics at Square One. 11th edition. *BMJ Books, Wiley-Blackwell*, Singapore 2009. ISBN: 9781405191005.

Cernigliaro, A., Fano, V., Scondotto, S., Forastiere, F., Pollina Addario, S., Caruso, S., Perucci, C.A., & Mira, A. (**2006**). Esperienza della Sicilia sulle aree a rischio ambientale. In: Indagini epidemiologiche nei siti inquinati: basi scientifiche, procedure metodologiche e gestionali, prospettive di equità (edited by Bianchi, F., & Comba, P.). (2006). *Rapporti ISTISAN 06/19 Rev. ISS,* Roma 2006. ISSN 1123-3117.

Cimorelli, A. J., Perry, S. G., Venkatram, A., Weil, J. C., Paine, R. J., Wilson, R. B., ... & Brode, R. W. (**2005**). AERMOD: A dispersion model for industrial source applications. Part I: General model formulation and boundary layer characterization. *Journal of Applied Meteorology*, *44*(5), 682-693.

Comba, P., Iavarone, I., & Pirastu, R. (**2016**). Contaminated sites: a global health issue. Preface. *Annali dell'Istituto Superiore di Sanità*, *52*(4), 472-475. DOI:10.4415/ANN_16_04_02.

Dahlgren, J., Klein, J., & Takhar, H. (**2008**). Cluster of Hodgkin's lymphoma in residents near a non-operational petroleum refinery. *Toxicology and Industrial Health*, *24*(10), 683-692. DOI:10.1177/0748233708100553.

Daniel, W. W. (Italian edition edited by Attanasio, M., & Capursi, V.) (**1996**). Biostatistica. Concetti di base per l'analisi statistica delle scienze dell'area medico-sanitaria. *EdiSES*, Napoli 1996. ISBN: 8879590758.

De Castro, P., Roberto, P., Marsili, D., & Comba, P. (**2016**). Fostering public health awareness on risks in contaminated sites. Capacity building and dissemination of scientific evidence. *Annali dell'Istituto Superiore di Sanità*, *52*(4), 511-515. DOI:10.4415/ANN_16_04_09.

European Environment Agency (**EEA**) (edited by Guerreiro, C., Ortiz, A. G., de Leeuw, F., Viana, M., & Horálek, J.) (**2016**). Air Quality in Europe – 2016 Report. EEA Report No. 28/2016. *Publications Office of the European Union,* Luxemburg 2016. ISBN 9789292138240. DOI:10.2800/413142.

Eni S.p.A. Divisione Refining & Marketing (**ENI**) (edited by Snamprogetti S.p.A.). (**2008**). "Nuovo Impianto EST - Eni Slurry Technology - Progetto innovativo per la conversione di oli combustibili in gasoli, da realizzare presso la Raffineria di Sannazzaro de' Burgondi (PV)". Studio di Impatto Ambientale. Sintesi Non Tecnica.

Fano, V., Cernigliaro, A., Scondotto, S., Pollina Addario, S., Caruso, S., Mira, A., Forastiere, F., Perucci, C., A. (**2006/A**). Analisi della mortalità (1995-2000) e dei ricoveri ospedalieri (2001-2003) nell'area industriale di Gela / Mortality (1995-2000) and hospital admissions (2001-2003) in the industrial area of Gela. *Epidemiologia & Prevenzione, 30*(1), 27-32.

Fano, V., Forastiere, F., Papini, P., Tancioni, V., Di Napoli, A., & Perucci, C. A. (**2006/B**). Mortalità e ricoveri ospedalieri nell'area industriale di Civitavecchia, anni 1997-2004 / Mortality and hospital admissions in the industrial area of Civitavecchia, 1997-2004. *Epidemiologia & Prevenzione*, *30*(4-5), 221-26.

Fazzo, L., De Santis, M., Minelli, G., Bruno, C., Zona, A., Marinaccio, A., ... & Comba, P. (**2012**). Pleural mesothelioma mortality and asbestos exposure mapping in Italy. *American Journal of Industrial Medicine*, *55*(1), 11-24. DOI:10.1002/ajim.21015.

Fleiss, J. L. (**1973**). Statistical methods for rates and proportions. *Wiley series in probability and mathematical statistic*, *Wiley*, New York 1973. ISBN: 0471263702.

Götschi, T., Sunyer, J., Chinn, S., de Marco, R., Forsberg, B., Gauderman, J. W., ... & Ponzio, M. (**2008**). Air pollution and lung function in the European Community Respiratory Health Survey. *International journal of epidemiology*, *37*(6), 1349-1358.DOI: 10.1093/ije/dyn136.

Granieri, A. (**2015**). Community exposure to asbestos in Casale Monferrato: from research on psychological impact to a community needs-centered healthcare organization. *Annali dell'Istituto Superiore di Sanità*, *51*, 336-341. DOI:10.4415/ANN_15_04_14.

Holmes, N. S., & **Morawska, L.** (**2006**). A review of dispersion modelling and its application to the dispersion of particles: an overview of different dispersion models available. *Atmospheric Environment*, *40*(30), 5902-5928. DOI:10.1016/j.atmosenv.2006.06.003.

International Agency for Research on Cancer (**IARC**). (**2016**). Outdoor Air Pollution. IARC Working Group on the Evaluation of Carcinogenic Risks to Humans, 2013. *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans. Volume 109. WHO Press*, Lyon 2016. ISBN 9789283201755.

Istituto Superiore per la Protezione e la Ricerca Ambientale (**ISPRA**). (**2016**). Linee guida per la valutazione integrata di impatto ambientale e sanitario (VIIAS) nelle procedure di autorizzazione ambientale (VAS, VIA, AIA): Delibera del Consiglio Federale. Seduta del 22/04/2015 Doc. 49/15-Cf. *Manuali e Linee Guida 133/2016*. ISBN 9788844807580.

Kane, E. V., & Newton, R. (**2010**). Occupational exposure to gasoline and the risk of non-Hodgkin lymphoma: A review and meta-analysis of the literature. *Cancer Epidemiology*, *34*(5), 516-522. DOI:10.1016/j.canep.2010.05.012.

Kelsey, J. L., Thompson, W. D., & Evans, A. S. (**1986**). Methods in Observational Epidemiology. *Monographs in Epidemiology and Biostatistics. Volume 10. Oxford University Press*, New York 1986. ISBN: 0195036573.

Kirkeleit, J., Riise, T., Bråtveit, M., & Moen, B. E. (**2008**). Increased risk of acute myelogenous leukemia and multiple myeloma in a historical cohort of upstream petroleum workers exposed to crude oil. *Cancer Causes & Control*, *19*(1), 13-23. DOI:10.1007/sl0552-007-9065-x.

Liu, C. C., Chen, C. C., Wu, T. N., & Yang, C. Y. (**2008**). Association of brain cancer with residential exposure to petrochemical air pollution in Taiwan. *Journal of Toxicology and Environmental Health, Part A*, *71*(5), 310-314. DOI:10.1080/15287390701738491.

MacKinnon, D. P., Krull, J. L., & Lockwood, C. M. (**2000**). Equivalence of the mediation, confounding and suppression effect. *Prevention Science*, *1*(4), 173-181.

Mancini, F. R., Busani, L., Tait, S., & La Rocca, C. (**2016**). The relevance of the food production chain with regard to the population exposure to chemical substances and its role in contaminated sites. *Annali dell'Istituto Superiore di Sanità*, *52*(4), 505-510. DOI:10.4415/ANN_16_04_08.

Manfredini, R., De Giorgi, A., Tiseo, R., Boari, B., Cappadona, R., Salmi, R., ... & Fabbian, F. (**2017**). Marital status, cardiovascular diseases, and cardiovascular risk factors: A review of the evidence. *Journal of Women's Health*, *26*(6), 624-632. DOI:10.1089/jwh.2016.6103.

Marsili, D. (**2016**). A cross-disciplinary approach to global environmental health: the case of contaminated sites. *Annali dell'Istituto Superiore di Sanità*, *52*(4), 516-523. DOI:10.4415/ANN_16_04_10.

Martuzzi, M., Pasetto, R., & Martin-Olmedo, P. (**2014**). Industrially contaminated sites and health. *Journal of Environmental and Public Health*, *2014*. DOI:10.1155/2014/198574.

Molloy, G. J., Stamatakis, E., Randall, G., & Hamer, M. (**2009**). Marital status, gender and cardiovascular mortality: behavioural, psychological distress and metabolic explanations. *Social Science & Medicine*, *69*(2), 223-228. DOI:10.1016/j.socscimed.2009.05.010.

MONITER Working Group. (**2012**). Gli effetti degli inceneritori sulla salute. Studi epidemiologici sulla popolazione in Emilia-Romagna. *Quaderni di MONITER 06/12*. Regione Emilia-Romagna – ARPA Emilia Romagna. ISBN 9788890737022.

Musmeci, L., Bianchi, F., Carere, M., & Cori, L. (**2009**). Ambiente e salute a Gela: stato delle conoscenze e prospettive di studio / Environment and health in Gela (Sicily): present knowledge and prospects for future studies. *Epidemiologia & Prevenzione*, *33*(3 Suppl. 1).

National Academies of Sciences, Engineering, and Medicine (**NAS**). (**2017**). Using 21st century science to improve risk-related evaluations. *National Academies Press*, Washington DC 2017. ISBN 9780309453486. DOI:10.17226/24635.

Nieuwenhuijsen, M. J. (**2005**). Design of exposure questionnaires for epidemiological studies. *Occupational and Environmental Medicine*, *62*(4), 272-280. DOI:10.1136/oem.2004.015206.

Omoti, C. E. (**2006**). Socio-demographic factors of adult malignant lymphomas in Benin City, Nigeria. *The Nigerian Postgraduate Medical Journal*, *13*(3), 256-260.

Orazzo, F., Nespoli, L., Ito, K., Tassinari, D., Giardina, D., Funis, M., ... & Nosetti, L. (**2009**). Air pollution, aeroallergens, and emergency room visits for acute respiratory diseases and gastroenteric disorders among young children in six Italian cities. *Environmental Health Perspectives*, *117*(11), 1780-1785. DOI:10.1289/ehp.0900599.

Pasetto, R., Martin-Olmedo, P., Martuzzi, M., & Iavarone, I. (**2016**). Exploring available options in characterising the health impact of industrially contaminated sites. *Annali dell'Istituto Superiore di Sanità*, *52*(4), 476-482. DOI:10.4415/ANN_16_04_03.

Pasetto, R., Comba, P., & Pirastu, R. (**2008**). Lung cancer mortality in a cohort of workers in a petrochemical plant: occupational or residential risk?. *International Journal of Occupational and Environmental Health*, *14*(2), 124-128. DOI:10.1179/oeh.2008.14.2.124.

Pesatori, A. C., Garte, S., Popov, T., Georgieva, T., Panev, T., Bonzini, M., ... & Fontana, V. (**2009**). Early effects of low benzene exposure on blood cell counts in Bulgarian petrochemical workers. *Medicina del Lavoro*, *100*(2), 83-90.

Pirastu, R., Pasetto, R., Zona, A., Ancona, C., Iavarone, I., Martuzzi, M., & Comba, P. (**2013**). The health profile of populations living in contaminated sites: SENTIERI approach. *Journal of Environmental and Public Health*, *2013*. DOI:10.1155/2013/939267.

Pirastu, R., Zona, A., Ancona, C., Bruno, C., Fano, V., Fazzo, L., ... & Mitis, F. (**2011**). Risultati dell'analisi della mortalità nel Progetto SENTIERI / Mortality results in SENTIERI Project. *Epidemiologia & Prevenzione, 35*(5-6 Suppl. 4), 29-152.

Prüss-Ustün, A., Wolf, J., Corvalán, C., Neville, T., Bos, R., & Neira, M. (**2016**). Diseases due to unhealthy environments: an updated estimate of the global burden of disease attributable to environmental determinants of health. *Journal of Public Health*, *39*(3), 464-475. DOI:10.1093/pubmed/fdw085.

Ramis, R., Diggle, P., Boldo, E., Garcia-Perez, J., Fernandez-Navarro, P., & Lopez-Abente, G. (**2012**). Analysis of matched geographical areas to study potential links between environmental exposure to oil refineries and non-Hodgkin lymphoma mortality in Spain. *International Journal of Health Geographics*, *11*(1), 4. DOI:10.1186/1476-072X-11-4.

Rood, A. S. (**2014**). Performance evaluation of AERMOD, CALPUFF, and legacy air dispersion models using the Winter Validation Tracer Study dataset. *Atmospheric Environment*, *89*, 707-720. DOI:10.1016/j.atmosenv.2014.02.054.

Rossi, M., Bonafè, G., Scotto, F., Trentini, A., & Pasti, L. (**2012**). PM2.5 and PM1 aerosol monitoring near a waste incineration plant located next to Bologna, in the Po Valley: the MONITER Project Campaign. Conference Paper. European Aerosol Conference – EAC2012 Granada (ES).

Rothman, K. J., & **Greenland, S**. (**1998**). Modern epidemiology. 2nd edition. Lippincott-Raven, Philadelphia 1998. ISBN 0316757802.

Schikowski, T., Sugiri, D., Ranft, U., Gehring, U., Heinrich, J., Wichmann, H. E., & Krämer, U. (**2005**). Long-term air pollution exposure and living close to busy roads are associated with COPD in women. *Respiratory research*, *6*(1), 152. DOI:10.1186/1465-9921-6-152.

SENTIERI Working Group (edited by Pirastu, R., Ancona, C., Iavarone, I., Mitis, F., Zona, A., & Comba, P.) (**2010**). SENTIERI Studio Epidemiologico Nazionale dei Territori e degli Insediamenti Esposti a Rischio da Inquinamento: valutazione della evidenza epidemiologica. *Epidemiologia & Prevenzione, 34*(5-6 Suppl. 3).

Simonsen, N., Scribner, R., Su, L. J., Williams, D., Luckett, B., Yang, T., & Fontham, E. T. (**2010**). Environmental exposure to emissions from petrochemical sites and lung cancer: the lower Mississippi interagency cancer study. *Journal of Environmental and Public Health*, *2010*.

Soskolne, C. L. (**2016**). Ethical aspects of epidemiological research in contaminated sites. *Annali dell'Istituto Superiore di Sanità*, *52*(4), 483-487. DOI:10.4415/ANN_16_04_04.

StataCorp LP. (**2013**). Stata Statistical Software: Release 13. College Station, TX.

Sunyer, J., Jarvis, D., Gotschi, T., Garcia-Esteban, R., Jacquemin, B., Aguilera, I., ... & Heinrich, J. (**2006**). Chronic bronchitis and urban air pollution in an international study. *Occupational and Environmental Medicine*, *63*(12), 836-843. DOI:10.1136/oem.2006.027995.

United States' Environmental Protection Agency (**US EPA**). (**2018/A**). User's Guide for the AMS/EPA Regulatory Model (AERMOD). Publication No. EPA-454/B-18-001.

United States' Environmental Protection Agency (**US EPA**). (**2018/B**). AERMOD Implementation Guide. Publication No. EPA-454/B-18-003.

Vanni, F., Scaini, F., & Beccaloni, E. (**2016**). Agricultural areas in potentially contaminated sites: characterization, risk, management. *Annali dell'Istituto Superiore di Sanità*, *52*(4), 500-504. DOI:10.4415/ANN_16_04_07.

Weiss, N. S. (**1973**). Marital status and risk factors for coronary heart disease. The United States health examination survey of adults. *British Journal of Preventive & Social Medicine*, *27*(1), 41-43.

Weng, H. H., Tsai, S. S., Chiu, H. F., Wu, T. N., & Yang, C. Y. (**2008**). Association of childhood leukemia with residential exposure to petrochemical air pollution in Taiwan. *Inhalation Toxicology*, *20*(1), 31-36. DOI:10.1080/08958370701758734.

World Health Organization (**WHO**) Regional Office for Europe. (**2017**). Declaration of the Sixth Ministerial Conference on Environment and Health. *European Environment and Health Process Secretariat EURO/Ostrava2017/6*. Ostrava, Czech Republic, 15 June 2017.

World Health Organization (**WHO**). (**2016**). Ambient air pollution: A global assessment of exposure and burden of disease. ISBN 9789241511353.

World Health Organization (**WHO**), Regional Office for Europe. (**2013**). Contaminated sites and health: Report of two WHO workshops: Syracuse, Italy, 18 November 2011 Catania, Italy, 21-22 June 2012.

World Health Organization (**WHO**) Regional Office for Europe. (**2006**). Air quality guidelines: global update 2005. ISBN 9289021926.

Yu, C. L., Wang, S. F., Pan, P. C., Wu, M. T., Ho, C. K., Smith, T. J., ... & Christiani, D. C. (**2006**). Residential exposure to petrochemicals and the risk of leukemia: using geographic information system tools to estimate individual-level residential exposure. *American Journal of Epidemiology*, *164*(3), 200-207. DOI:10.1093/aje/kwj182.

APPENDIX

A. DISPERSION MODELLING

A.1. FOREWORD

This Appendix should be intended as an introduction to dispersion modelling techniques applied to assess individual exposure to air pollution, with a particular reference to the methods applied in this study. It starts with some basic vocabulary, then moves towards Gaussian modelling in general and concludes with AERMOD. The aim is not to give a thorough explanation, which would be out of scope for the present thesis.

A.2. KEY CONCEPTS

The concentration of a pollutant at ground levels depend on a series of factors: the characteristics of the pollutant itself, context, local characteristics of the Earth's surface, and meteorological conditions, this last being of utmost importance because the atmosphere has a "diluting effect" on the dispersion of pollutant [BLUETT ET AL., 2004]. Thus, a quick reminder about the atmosphere's structure (and how pollutants are released and removed from it) is needed before tackling dispersion modelling strategies.

The ***Planetary Boundary Layer*** (PBL) is the lowest part of the atmosphere (specifically, of the troposphere); it is in direct contact with the Earth's surface, and consequently with men and biosphere, thus implying the existence of interactions and exchanges between them. In other words, the PBL is influenced by Earth's surface. At night, air in the lowest part of the PBL is heated by the ground which is hotter than air in the PBL's upper part (the so-called "free atmosphere"), thus determining vertical convection; when air at ground level becomes colder (because the ground cease to heat it), vertical convective motion is inhibited. The PBL, especially in convective conditions, is upper-limited by a *thermal inversion layer*, preventing lower atmosphere to have exchanges with the upper atmosphere: this results in stagnation of pollutants[117].

The stacks of a refinery can be classified as ***point sources***. A point source is defined as a (spatially) single and identifiable emission of pollutants in the atmosphere; a point source is assumed to have no geometry (i.e. no dimensions) and is only characterised by its elevation. Other types of sources are *line sources*, *area sources* and *volumetric sources* [BLUETT ET AL., 2004]. All these sources emit pollutants in the PBL.

The release of pollutants by a point source in the atmosphere typically takes place in the form of a flow of smoke or vapours made up by the contaminants themselves. This flow forms an *emission plume*. ***Buoyant plumes*** are of particular interest for what concerns stack emissions: this definition refers to plumes "floating on ambient air" because, respect to surrounding atmospheric air, the smoke has either higher

[117] See the following link: *http://www.treccani.it/enciclopedia/atmosfera-lo-strato-limite_%28 Enciclopedia-della-Scienza-e-della-Tecnica%29/* (in Italian; opened on July 28th, 2018).

temperatures (like, in general, stack-emitted gases) and/or contains contaminants of lower density (e.g. methane). Other types of plumes are *dense plumes* (having lower temperatures and/or higher density than air) and *neutral plumes* (roughly as dense as surrounding air).

If a plume flows in a field with elevated buildings (or other structures), i.e. on a complex terrain rather than on a simple one, *downwash effect* will make the plume moving down sooner (***Figure A.1***). This is a result of the turbulences formed leeward of the building and it can increase ground deposition processes nearby.

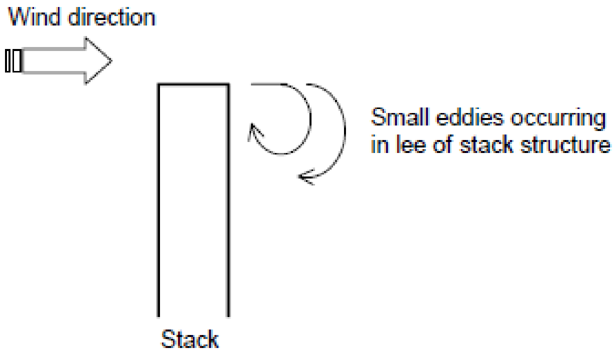

Figure A.1 *– Schematic representation of downwash effect. Picture from* BLUETT ET AL., *2004.*

The **removal of contaminants** from the plume can occur by two different processes. In *dry deposition*, gaseous or particulate pollutants are transferred to a surface (ground, water, or vegetation) by absorption and gravitational sedimentation, thus they are subtracted from the plume depending on deposition velocity (which itself results from the resistance the surface has in "accepting" the pollutant). *Wet deposition* is a result of rain.

A.3. DISPERSION MODELLING

A simple, essential definition of **dispersion modelling** is given by HOLMES & MORAWSKA [2006] as "[using] *mathematical equations, describing the atmosphere, dispersion and chemical and physical processes within the plume, to calculate concentrations at various locations*" (***Figure A.2***). In their review, they deeply compare and discuss various approaches; to sum up this (rather complicate) issue in a few words, the choice of one model over another depends on parameters like environmental context and complexity, scale, type of pollutants and concentrations.

Different modelling approaches, each one grounded on different assumptions, can be adopted to follow contaminants moving in the air matrix (and, from the air matrix, to other matrixes and to receptors).

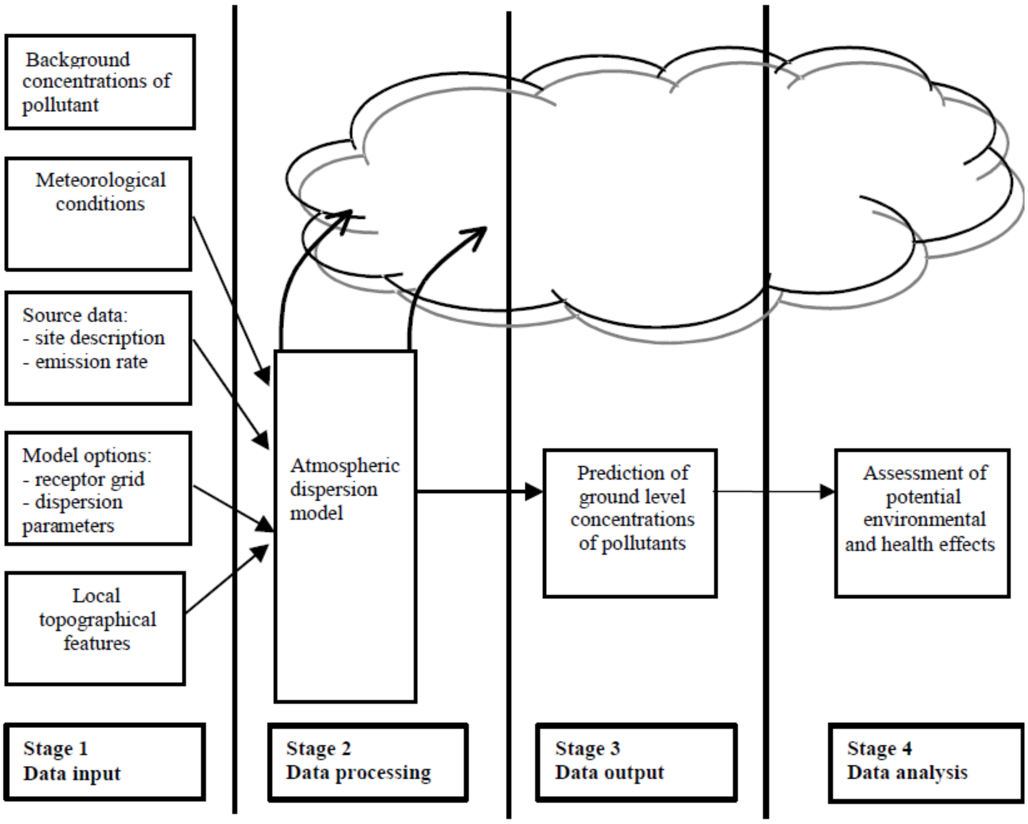
Background concentrations of pollutant
Meteorological conditions
Source data:
- site description
- emission rate
Model options:
- receptor grid
- dispersion parameters
Local topographical features
Atmospheric dispersion model
Prediction of ground level concentrations of pollutants
Assessment of potential environmental and health effects
Stage 1
Data input
Stage 2
Data processing
Stage 3
Data output
Stage 4
Data analysis

Figure A.2 *– Flowchart of the dispersion modelling approach. Picture from* BLUETT ET AL., *2004.*

Gaussian dispersion models[118] are widely used for the sake of studying atmospheric dispersion, particularly in regulatory settings; this class of models rely on the assumption that, under steady-state conditions, the distribution of a pollutant in the plume is fitted by the normal probability distribution in both the horizontal and the vertical direction [HOLMES & MORAWSKA, 2006]. Specifically, *Gaussian plume* models find wide applications in predicting how buoyant plumes, resulting from continuous emissions from (multiple) sources, are dispersed.

These models, albeit their wide application, have some major limitations. First of all, they do not consider the time needed for the pollutant to reach a receptor (as they assume steady-state conditions) and chemical transformations of contaminants are basically not accounted for. As a result, some tasks (e.g. modelling particle dispersion) need post-processing of model results. Furthermore, no interaction between plumes is assumed, which can be unrealistic in urban contexts [HOLMES & MORAWSKA, 2006]. Some of the aforementioned limitations can be resolved if emissions are approximated as a series of "puffs" over time: model restrictions are applied to each single puff, thus allowing conditions (e.g. wind speed) to change. In *Gaussian puffs* models, the integration of individual puffs in time gives the contribution of the source, whereas the exposure of a receptor in a certain position is computed as the sum of the contributions from different puffs.

The steady-state Gaussian approaches are unsuitable for predicting concentrations in positions less than 100 metres far from the source. At the same time, their application is either not recommended for predicting concentrations too far from the source, because meteorological conditions are approximated as homogeneous in the entire field, and this would be very unrealistic on large scale[119] [HOLMES & MORAWSKA, 2006]. In addition, they have been shown to over-predict the concentration of contaminants in places where the wind is low, even though this might be partially improved by hybrid plume-puff models [HOLMES & MORAWSKA, 2006].

[118] As it will be clarified later in the paragraph, AERMOD is a Gaussian model.

[119] In case the interest is in predicting at higher distances, other models could be used [HOLMES & MORAWSKA, 2006]. CALPUFF, in particular, is a non-steady-state Lagrangian puff model developed by US EPA; it allows for a spatially- and temporally-changing wind field. It is the current (as of July 2018) preferred model for regional-scale dispersion studies [ROOD, 2014].

A.4. BASICS OF AERMOD

The *American Meteorological Society (AMS) and United States Environmental Protection Agency (US EPA) Regulatory Model*, known by his acronym AERMOD, was developed by the AERMOD Improvement Committee (AERMIC), a collaborative group formed by the two agencies in 1991, and it has been continuously improved ever since to keep the model up to date with state-of-the-art evidence. AERMOD was officially recognised as the reference model for regulatory purposes in the United States on December 2006[120]. It replaced an older US EPA's system, the third version of the Industrial Source Complex Short-Term model (ISCST3), which dated back to 1995 and was intended to be used in the same scenarios. Currently (as of July 2018), AERMOD is still the first-choice model for the dispersion of pollutants in the near field (within 50 kms from the source) [ROOD, 2014]. A fully detailed and up-to-date explanation of the model and the implementation of its latest versions is available in the User's Guide document [US EPA, 2018/A] on US EPA's website[121].

AERMOD's roots are between the Seventies and the Eighties, when our understanding of PBLs started to improve, until they started to be integrated in applied dispersion modelling [CIMORELLI ET AL., 2005]. AERMOD's predecessor ISCST3 was also based on this theoretical framework, but it assumed an infinite boundary layer; in AERMOD, on the contrary, PBL is parametrised (thus allowing for plume reflection due to a "superior lid effect"). Besides, in AERMOD the model is complemented by plume interaction with simple and complex terrains, surface releases, and downwash effects [CIMORELLI ET AL., 2005]. The system is made up of different modules, of which the main relevant are: the processor AERMOD, producing simulations to estimate the concentration of pollutants; AERMET, pre-processing meteorological data to drive simulations; and AERMAP, pre-processing information regarding complex terrain as required by the simulation processor [US EPA, 2018/A].

AERMOD is a ***near-field steady-state Gaussian plume*** *dispersion model*, but in a certain sense it is an evolution of the overall idea of this kind of approaches as it was

[120] *Revision to the Guideline on Air Quality Models: Adoption of a Preferred General Purpose (Flat and Complex Terrain) Dispersion Model and Other Revisions; Final Rule*. US EPA, 2005.
[121] The document can be downloaded at the following link: *https://www.epa.gov/scram/air-quality-dispersion-modeling-preferred-and-recommended-models* (opened on July 27th, 2018).

described in ***A.3 Dispersion modelling*** [BLUETT ET AL., 2004]. AERMOD accounts for the effects of vertical variations in PBLs on contaminants dispersion: it considers a *Stable Boundary Layer* (SBL) in which the concentration of pollutants is normally distributed both horizontally and vertically, and a *Convective Boundary Layer* (CBL) in which those concentrations are normally distributed horizontally whereas they are fitted by a bi-Gaussian probability density function vertically. In general terms, the structure of a PBL and the dispersion of pollutants in it are influenced by heat and by surface effect[122]. PBL-related parameters required by AERMOD are computed at a 1-hour resolution by meteorological pre-processors like AERMET. Then, concentrations of pollutants are modelled by means of specific equations for SBL and CBL [CIMORELLI ET AL., 2005].

In order to account for the role of terrain in the model, AERMOD looks at the plume as the combination of a *terrain-impacting* (horizontal) and a *terrain-following* component (***Figure A.3***). The plume, in detail, will be the sum of the two components, with each component weighted for the residual amount of mass in it [CIMORELLI ET AL., 2005]. Terrain characteristics are pre-processed by AERMAP.

To sum up, AERMOD is suitable for application in estimating the concentration of pollutants emitted in form of continuous buoyant plumes by single or multiple sources of different types (point, line, area or volume), either elevated or at ground level, with constant or variable emission rates, and is able to model wet and dry deposition of particulate and gases [US EPA, 2018/A]. AERMOD can be used both in urban and rural contexts and for simple or complex terrains; a specific algorithm (Plume Rise Model Enhancement, PRIME) can also account for *downwashing*. Moreover, AERMOD can model the dispersion of particulates [HOLMES & MORAWSKA, 2006]. It is also very flexible regarding the specification of receptors [US EPA, 2018/A]. It is worth mentioning that the possible over-prediction of pollutants' concentrations in low wind conditions, already mentioned while commenting Gaussian dispersion models in general, is still a limitation that AERMOD improvements have not been able to overcome [US EPA, 2018/B].

[122] It is useful to say that enhanced vertical turbolences are implemented, thus accounting for the convective-like behaviour of PBLs during night-time, due to the "urban heat island" effect [CIMORELLI ET AL., 2005; US EPA, 2018/B].

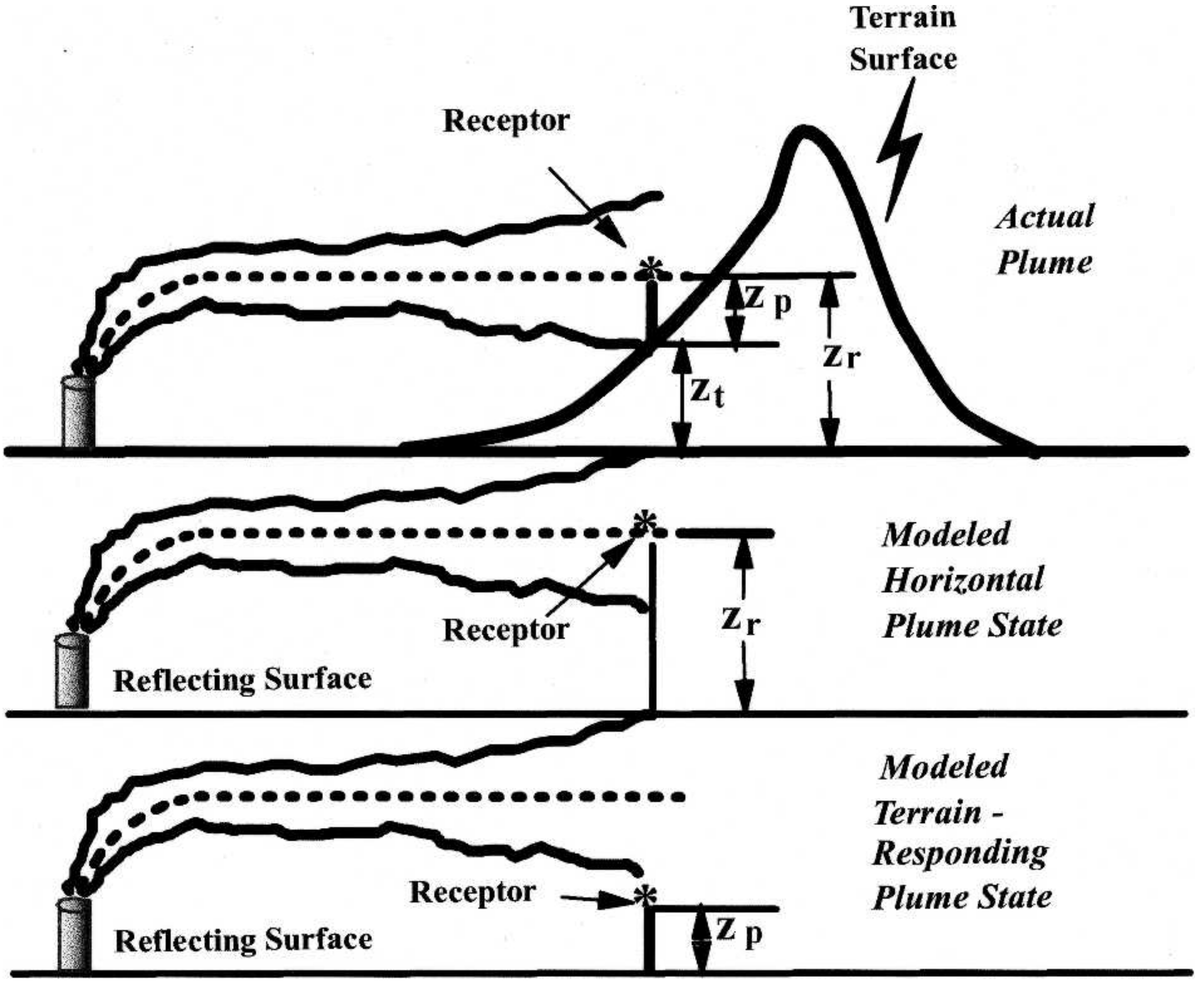

Figure A.3 *– An exemplification of the AERMOD approach to plume modelling. Up: real plume. Middle: terrain-impacting component (z_r is the vertical coordinate of the receptor). Down: terrain-following component (z_p is the height of the receptor referred to local ground level). Picture from* CIMORELLI ET AL.*, 2005.*

B. STATISTICAL METHODS

B.1. FLEISS FORMULA FOR SAMPLE SIZE ESTIMATION

The estimation of sample size is aimed at identifying the number of subjects that must be enrolled to the study in order to control the chance of failing by two different types of error. Type I error consists in concluding for an effect when the effect is actually null[123]: in other words, to refute the null hypothesis of a statistical test when it is actually true. Type I error is controlled by imposing a threshold (*significance level*) to the probability of making such mistake, usually indicated with the Greek letter "α" and set at .05. Type II error, on the contrary, occurs when a study fails in identifying an effect that actually exists, and is generally indicated with the Greek letter "β". The value of β can vary depending on the study and the aims of the investigators, but it is generally below .20 since its practical implications are less serious[124].

It is relevant to notice that a certain difference between two samples will always exist and, no matter how small, an exaggerated sample size could always make that difference reaching statistical significance [FLEISS, 1973]. For this reason, only imposing a threshold to statistical significance is not enough, and sample size should not be larger than the number of subjects needed to disclose effects of a relevant magnitude, and not of any magnitude. At this point, the core question that makes statistical significance of a practical meaning becomes: what is "relevant"? The answer to this question, or *effect size*, should be given *a priori* by the researcher, and should be based on previous data (regarding tendency and variability of the endpoint of the study) and on practical considerations.

Given the *effect size*, Type II error acquire itself a practical meaning, which is defined by its complement $1 - \beta$, or *study power*. For instance, if β=.20, then study power will

[123] In a parallelism with clinical tests, this would be a false positive finding.

[124] Again in a parallelism with clinical tests, this would be a false negative finding and it is considered a less serious event than a false positive. The reason for that becomes especially clear-cut if a Phase III clinical trial of a new drug versus another treatment is taken as an example: after the end of the study, a Type I error would imply giving an ineffective treatment to patients, while a Type II error would mean continuing to do what was done previously (e.g. other treatments, surgery…).

be 0.80, indicating that the study has an 80% probability of detecting an effect as significantly different from a null effect if the effect is at least as great as the pre-specified effect size. Thus, in order to compute the estimated sample size for a study, at least the three following parameters are needed: threshold for Type I error, threshold for Type II error, and effect size.

Various formulas for the estimation of sample size in case-control designs has been proposed in the past decades. For the sake of the study presented in this thesis, the one elaborated by Fleiss [FLEISS, 1973] for the comparison of proportions was used, and therefore that one is discussed here. Let a *contingency table* with an equal number *n* of cases and controls[125] be assumed; the *proportions* of exposed (and non-exposed) subjects among cases and among controls could be defined as follows:

	Exposed	***Non-exposed***	**Marginal**
Cases	p_1	$1-p_1$	1
Controls	p_2	$1-p_2$	1
Marginal	p	$1-p$	1

If the interest is in *comparing* the two proportions p_1 and p_2, namely exposed cases and exposed controls respectively, then the *test statistic* could be computed from the difference of the two proportions compared to their difference under the hypothesis of a null effect (i.e. a difference of zero):

$$z = \frac{(p_1 - p_2) - 0}{\sqrt{p(1-p)\left({}^{1}\!/_{n} + {}^{1}\!/_{n}\right)}} = \frac{p_1 - p_2}{\sqrt{\frac{2p(1-p)}{n}}} \qquad (*)$$

which is, in other words, the estimated difference between the two proportions divided by the standard error of this estimate. By recalling the properties of the standard distribution, it is straightforward that this test would give indication of a *statistically significant difference* if:

$$|z| > z_{\alpha/2}$$

with $z_{\alpha/2}$ being the *threshold* for a *two-tailed test* with *significance level* α[126].

[125] The assumption of an equal number of cases and controls is made only for the sake of simplicity of the mathematical proof. The formula for unequal number of cases and controls will be presented at the end of the paragraph.

[126] The threshold, for α=.05, would be *z*=1.96. The modulus was placed because the standard

If the true difference between the proportions is assumed to be $P_1 - P_2$, then *study power* could be defined as follows:

$$Pr\left\{\frac{|P_1 - P_2|}{\sqrt{\frac{2p(1-p)}{n}}} > z_{\alpha/2}\right\} = 1 - \beta$$

To make calculations easier, the difference between the proportions will be assumed to be positive (i.e. $P_1 > P_2$). A simple series of mathematical transformations could be applied to the formula reported above[127]:

$$1 - \beta = Pr\left\{p_1 - p_2 > z_{\alpha/2}\left(\sqrt{\frac{2p(1-p)}{n}}\right)\right\} =$$

$$= Pr\left\{(p_1 - p_2) - (P_1 - P_2) > z_{\alpha/2}\left(\sqrt{\frac{2p(1-p)}{n}}\right) - (P_1 - P_2)\right\} =$$

$$= Pr\left\{\frac{(p_1 - p_2) - (P_1 - P_2)}{\sqrt{\frac{p_1(1-p_1) + p_2(1-p_2)}{n}}} > \frac{z_{\alpha/2}\left(\sqrt{2p(1-p)/n}\right) - (P_1 - P_2)}{\sqrt{\frac{p_1(1-p_1) + p_2(1-p_2)}{n}}}\right\} \quad (\#)$$

The first member of the last inequation is essentially the test statistic defined in $(*)$, but in this case the reference effect is the true difference instead of a null difference. If n is large enough, this statistic has a standard distribution:

$$Z = \frac{(p_1 - p_2) - (P_1 - P_2)}{\sqrt{\frac{p_1(1-p_1) + p_2(1-p_2)}{n}}}$$

If $z_{1-\beta}$ is the threshold value of the standard distribution for a probability of $1 - \beta$, study power could be defined also as follows:

$$1 - \beta = Pr\{Z > z_{1-\beta}\}$$

If this expression is matched with the other definition of study power gave in $(\#)$, it is easy to understand that the following should be verified:

distribution is symmetric for z=0 and, in a two-tailed test, the difference between the proportions could be either positive or negative, i.e. p_1>p_2 and p_1<p_2 are both relevant.

[127] It is worth recalling that the standard error of a proportion p computed over n cases is $p(1-p)/n$.

$$z_{1-\beta} = \frac{z_{\alpha/2}\left(\sqrt{2p(1-p)/n}\right) - (P_1 - P_2)}{\sqrt{\frac{p_1(1-p_1) + p_2(1-p_2)}{n}}}$$

This equation can be solved to find the quantity n, given the other parameters:

$$n = \frac{\left[z_{\alpha/2}\sqrt{2p(1-p)} - z_{1-\beta}\sqrt{p_1(1-p_1) + p_2(1-p_2)}\right]^2}{(p_1 - p_2)^2}$$

Similarly, the formula for unequal numbers of cases and controls could be proven, being r the number of controls to be enrolled per each case:

$$n = \frac{\left[z_{\alpha/2}\sqrt{(r+1)p(1-p)} - z_{1-\beta}\sqrt{r\,p_1(1-p_1) + p_2(1-p_2)}\right]^2}{r(p_1 - p_2)^2}$$

where n will be the number of cases and $r \times n$ the number of controls.

The proportions required to complete the calculations can be computed from an OR and the marginal proportion of exposed among the general population (≈ the proportion of exposed among controls).

B.2. Comparison of quantitative endpoints between groups

The comparison of a quantitative endpoint (i.e. response variable) between two or more independent groups can be done via parametric or non-parametric statistical methods. The application of parametric methods is founded on the estimation of a parameter of the target population with a certain (known) sampling distribution, while generally non-parametric methods look at the ranks of the observations, so that the underlying assumptions regarding the parameters' distributions are released. However, non-parametric tests are less informative regarding the population than parametric tests [Campbell & Swinscow, 2009].

Consistently with the study design, the statistical methods presented here are meant to compare independent measures of the endpoint or, in other words, how the endpoint is distributed in *independent groups* made up of different statistical units.

B.2.1. Student's t *test*

The *Student's t test for unmatched data* (or *unmatched groups* or *independent groups*) is a parametric test used to compare the means of continuous variables in two independent groups[128]. It was proposed by William Sealy Gosset, who published it under the pseudonym "A. Student".

Two populations P_A and P_B, defined by a nominal dichotomous variable (two-levels group factor), and a quantitative endpoint y (response variable) are considered. The populations' means, μ_A and μ_B, and the respective variances σ_A e σ_B, are unknown. From these populations, independently one from the other, two representative samples C_A and C_B are randomly selected; observations taken in the samples will allow to produce an estimate of mean and variance (or standard deviation) in each of the two samples:

$$P_A: \mu_A, \sigma_A \rightarrow C_A: \bar{y}_A, s_A$$

$$P_B: \mu_B, \sigma_B \rightarrow C_B: \bar{y}_B, s_B$$

[128] Other versions of the test are suitable to compare a sample estimate of the variable's mean with the population mean (parameter) or to compare the mean values of matched groups of measurements.

Before applying the Student's test, the following assumptions must be checked with appropriate tests that will be discussed later in this paragraph:

- in each population, the endpoint y fits with a normal (Gaussian) distribution $N(\mu, \sigma^2)$;
- the two variances are homoscedastic (homogeneous), i.e. $\sigma_A{}^2 = \sigma_B{}^2 = \sigma^2$.

The methods adopted to test these assumptions are described in ***B.2.5 Testing the assumptions of parametric methods***. Meeting the assumption of homoscedasticity is even more important than the endpoint being normally distributed. The *t* tests are indeed quite robust against non-normality of the endpoint variable, especially if sample sizes are large, but the tests for unmatched data are severely biased by unequal variances in the two groups, with a strong reduction in test's power. It could be possible to adopt a correction for unequal variances in the formula of the test statistic (Welch's test) but, particularly when sample size is low, non-parametric methods might prove to have a higher power than the corrected *t* test[129].

As any other inferential test, the Student's test adopts a *reductio ad absurdum reasoning*, i.e. the existence of a difference is claimed if observations don't fit with the hypothesis of that difference being null. Thus, the research hypothesis is split into two complementary hypotheses:

- *null hypothesis* H_o: $\mu_A = \mu_B \rightarrow \mu_A - \mu_B = 0$
  (i.e. no difference is observed between the two population means – or, the two groups come from the same population);
- *alternative hypothesis* H_1: $\mu_A \neq \mu_A \rightarrow \mu_A - \mu_A \neq 0$
  (i.e. such difference exists).

Regarding the hypotheses, it is important to mention that the test could be either *one-tailed* or *two-tailed*. A one-tailed test is used when it is possible to exclude *a priori* one of the two possible directions of the difference, i.e. it is justifiable to state that one mean is for sure higher/equal (but not lower), or lower/equal (but not higher), than the other. If this cannot be justified, then the two-tailed test, which is more conservative with regard to the null hypothesis[130], should be used.

[129] For this reason, in case the assumption was not respected, non-parametric methods were used; therefore, the corrected formula for the Student's test is not presented.

[130] "*More conservative to the null hypothesis*" means that the test is less likely to reject the null

If the assumptions for applicability are met, the test statistic is calculated with the following formula[131]:

$$t = \frac{(\bar{y}_A - \bar{y}_B) - \overbrace{(\mu_A - \mu_B)}^{null\ under\ H_0}}{\sqrt{{s_p}^2\left({}^1\!/_{n_A} + {}^1\!/_{n_B}\right)}}$$

In the equation above, ${s_p}^2$ is the *pooled variance*, i.e. the weighted mean of the two sample variances on the respective degrees of freedom:

$${s_p}^2 = \frac{(n_A - 1){s_A}^2 + (n_B - 1){s_B}^2}{n_A + n_B - 2}$$

(It is worth noticing that the assumption of homoscedasticity is required for an unbiased estimate of the pooled variance, and only indirectly for the test statistic itself. Unequal variances would determine a biased estimation of the pooled variance, thus resulting in a biased estimation of the test statistic and, ultimately, in an inflated or deflated probability of refuting the null hypothesis).

Once the test statistic has been calculated, it is contrasted to the theoretical *t probability density function* (PDF) [132] to obtain the related *p-value* [133]. The *t* distribution is essentially a normal distribution in which the population variance is unknown, so that variability is its sample estimate, a fact that determines additional uncertainty in the estimation of standard errors (SE) (the denominator of the test statistic). Therefore, the *t* distribution is actually a family of distributions, each one defined on the basis of the degrees of freedom (*df*); the lower the *df*, the more spread is the distribution – or, in other words, the *t* distribution will be closer to a normal distribution as *df* increase. In the case of the Student's test for unmatched data, the degrees of freedom of the test are given by the degrees of freedom of the pooled variance, i.e. $df = n_A + n_B - 2$. The p-value is obtained by definite integration of the PDF.

hypothesis and uphold the existence of a difference.

[131] This formula is nothing more than dividing a difference of means by its standard error.

[132] A PDF is a probability function for continuous random variables. In that case, the probability of observing a specific value in the distribution tends to zero, and the function defines the density of probability over a certain value. Probability is calculated by integration over an interval of values.

[133] A *p-value* is defined as the probability of doing a Type-I error because the null hypothesis is refuted when it is actually true. It indicates the probability that the observed value of the test statistic (i.e. the observed difference) or a higher one could be obtained by chance.

B.2.2. One-way analysis of variance (ANOVA)

The Student's test is designed to compare the means between two groups. If the groups to be compared are more than two, applying that test to each possible couple of groups would result in an amplification of Type-I error, i.e. the probability of a false positive result would be increased. The reason for that is easy to prove: the results of the tests are independent events; therefore, the total probability of not refuting the null hypothesis when the null hypothesis is actually true in all the tests would be $(1-\alpha)^k$. Consequently, if the multiple comparisons are made with Student's tests, the final probability of making a Type-I error in at least one of the comparisons would be $\alpha_T = 1-(1-\alpha)^k$. It is clear-cut that such probability increases with k, so using a series of Student's tests would not be an appropriate analysis strategy.

The parametric statistical method with the highest power in the comparison of continuous variables in k different groups, defined on the basis of a polytomous nominal variable (k-levels group factor), is the *one-way ANOVA* (*ANalysis Of VAriance*). If more than one factor is considered, the ANOVA is said to be *multi-way.*

The categories of the factor can represent either a random sampling from a population of categories of the factor with a normal distribution (*random-effects model*, in which the objective of the analysis is not the factor itself but rather an estimation of a phenomenon's variability, represented by the factor) or they can be "assigned" – observed, in epidemiological studies – by the researcher (*fixed-effects model*, in which the factor is strictly relevant to the scope of the analysis). For the sake of the present thesis, the fixed-effects model will be presented, although the theoretical foundation is similar for the random-effects model.

A k number of populations P_i is considered, each one with the continuous endpoint y having μ_i and ${\sigma_i}^2$ as mean and variance, respectively. From each population, a representative sample C_i is randomly selected.

The analysis of variance can be seen as a generalisation of the Student's t test for unmatched data to k unmatched groups and, in fact, the assumptions are pretty similar:

- in each population from which the samples were drawn, y fits $N(\mu_i, {\sigma_i}^2)$;
- variances are homoscedastic between the populations, ${\sigma_1}^2 = {\sigma_2}^2 \ldots = {\sigma_i}^2 = \sigma$;
- the residuals are normally distributed.

The assumptions should be checked with appropriate statistical tests (the related methods are introduced in ***B.2.5 Testing the assumptions of parametric methods***). In case the assumptions are not met, non-parametric tests could better be applied.

ANOVA is based on the decomposition of the total variance and then in the comparison of observed between-groups variance ($MS_{between}$) and within-group variance (MS_{within})[134]. The test statistic is their ratio *RV* (for *Ratio of Variances*):

$$RV = \frac{MS_{between}}{MS_{within}}$$

The variance between groups can be attributed to the effect of stochastic processes and of the factor of interest, if the factor actually has an effect. The variance within groups, on the other hand, only depends on stochastic processes. In other words, it the factor exerts no effect, the statistical units would actually have been drawn from the same population, so that $MS_{between}$ and MS_{within} would be different estimates of the same stochastic variability. For this reason, for sure $MS_{between} \geq MS_{within}$ (i.e. $RV \geq 1$).

Analogously to any other inferential test, ANOVA is based on a *reductio ad absurdum* of the research hypothesis, which is split into the following statistical hypotheses, one complementing the other:

- *null hypothesis* H_o: $RV = 1 \rightarrow \mu_A = \mu_B = \cdots = \mu_k$
  (i.e. no difference is observed between the populations, the factor has no effect);
- *alternative hypothesis* H_1: $RV \neq 1$
  (i.e. such difference exists in at least one couple of populations, the factor has an effect).

The value of *RV* is contrasted with the theoretical Fisher's *F* PDF, which is non-symmetrical. The family of *F* distributions is defined on two parameters: the *df* of the numerator, which are those of the between-groups variance ($df_{between} = k - 1$), and the *df* of the denominator, i.e. of the within-groups variance ($df_{within} = N - k$). The p-value is computed by integration of the PDF in the interval from RV to infinite.

[134] The between-group variance is the average squared deviation of each group's mean from the general mean (i.e. an estimate of the effect of the group factor); the within-group variance is the average squared deviation of each observation from the mean value of the group it belongs to (i.e. and estimate of the residual).

If the ANOVA is statistically significant, then it is possible to conclude that a difference is observed in at least a couple of groups; the identification of the couple(s) in which such difference is observed requires a *post-hoc* testing to be carried out. A common *post-hoc* test is essentially a Student's test (see ***B.2.1 Student's t test***), in which within-group variances (with their respective degrees of freedom) are used instead of sample variances. In order to avoid the amplification of Type-I error, discussed while introducing ANOVA, the *post-hoc* tests must be adjusted for *multiple testing*. Among all the available options, in the present thesis the Bonferroni's method was chosen, because it is the most conservative to the null hypothesis. Basically, it consists in "spreading" the threshold for Type-I error (i.e. α) on all the m comparisons, thus defining a corrected significance threshold $\alpha' = \alpha/m$.

B.2.3. Mann-Whitney's test

Mann-Whitney's test, also indicated as U test, is a non-parametric test and can be considered as an alternative to Student's test for unmatched data, given that the two are suitable for similar circumstances. Mann-Whitney's test requires as endpoint a quantitative trait measured at least with an ordinal scale, considered in two populations defined by a two-levels group factor. This test, differently from the Student's one, works on observation's ranks instead of values or, said with other words, it uses medians as measures of central tendency instead of means. This implies that less information is used, but still the test can be conveniently applied when the conditions and assumptions for parametric testing are not met.

As previously commented concerning other inferential tests, also in this case the underlying reasoning is a *reductio ad absurdum* in which the null hypothesis is that the two groups are drawn from populations with the same median[135], i.e.:

- *null hypothesis* H_0: $median_A = median_B$
  (i.e. no difference is observed);
- *alternative hypothesis* H_1: $median_A \neq median_B$
  (i.e. a difference is observed).

[135] Actually, the hypotheses can be defined on medians only by assuming that the differences are given by a shift in the distribution from one group to the other. It is interesting to remind that such shift would move the median and the mean by the same extent [CAMPBELL & SWINSCOW, 2009].

Also in this case, the statistical hypotheses can be formulated as a two-tailed test or, if this can be justified *a priori*, as a one-tailed test.

As commented above, the Mann-Whitney's test works on the ranks of the observations to look for differences in the distribution of the endpoint variable in the two groups. In detail, it requires to order all the observations together, regardless of the group, and to assign ranks on the joint series of observations. Formally, the test statistic is based on enumerating how many observations of one group can be found before an observation of the other group and summing those counts per each group; the sums, U_A and U_B, are expected to be nearly equivalent under the null hypothesis, whereas if the groups are different $min(U_A, U_B)$ will be close to zero. The lower sum is always used to reduce computational efforts. It can be proven that the sums verify the following equation:

$$n_A n_B = U_A + U_B$$

Practically, the computing can be made easier, as they can be computed directly from the ranks with the following formula (its demonstration is beyond the scope of this Appendix), where $R_{k,i}$ is the rank of the i^{th} observation from group k and n_k is the sample size of that group:

$$U_k = \sum_{i,k} R_{k,i} - \frac{n_k(n_k + 1)}{2}$$

(The formula can be adjusted for the presence of ties, i.e. observations with the same value for which the rank is defined as the mean of the ranks those observations would have if they were consequential).

The exact p-values corresponding to the observed value of the test statistic are reported in a table. Anyway, if the sample sizes are large, it can be proven that the distribution of U is well approximated by a standard distribution, so that an asymptotic p-value could be easily obtained[136]. For the standardisation, the population mean is represented by the mean between U_A and U_B, i.e. $\mu_U = n_A n_B/2$, while the standard deviation is computed from the sample sizes of the two groups:

$$\sigma = \sqrt{\frac{n_1 n_2(n_1 + n_2 + 1)}{12}}$$

[136] The statistical software Stata only reports the asymptotic p-value.

So, the corresponding standardisation of the observed value of U is defined with the formula of the standard scores:

$$MW = \frac{U - \mu_U}{\sigma}$$

The asymptotic p-value for a two-tailed test is easily computed as the definite integral of the mathematical function of the standard PDF:

$$p - value = \int_{|Z|}^{+\infty} f(x) \qquad \text{where} \qquad f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-(x-\mu)^2 / 2\sigma^2}$$

B.2.4. Kruskal-Wallis' test

The *Kruskal-Wallis'* method is essentially the non-parametric counterpart of a one-way ANOVA for a quantitative endpoint[137] variable in k independent groups. In a certain sense, it is an extension of the Mann-Whitney's test. The test works on the null hypothesis that the k groups are drawn from populations with the same median:

- *null hypothesis* H_o: $median_A = median_B = \cdots = median_k$ (i.e. no difference is observed);
- *alternative hypothesis* H_1: $median_A \neq median_B$ for at least one couple of groups (i.e. a difference is observed).

Analogously to the Mann-Whitney's procedure, the observations from all groups are combined, and ranks are assigned to the joint series of observations. For each group, the sum of ranks is computed and, then, the mean of ranks ($\overline{R_k}$) by dividing for the n_k observations in that group. Likewise, the general mean of ranks ($\bar{R}$) is computed. The test statistic is computed as follows (N is the total sample size):

$$KW = \frac{12}{N(N+1)} \sum_{k=1}^{K} n_k(\overline{R_k} - \bar{R})^2$$

where the first part of the product represents the inverse of the ranks' variance, which only depends on the total sample size. The formula reported above can be adjusted for the presence of ties.

Under the null hypothesis, the mean of ranks should be similar in all groups and to the general mean of ranks, which implies that KW=0. How strongly the data support that

[137] Or, at least, an ordinal endpoint if the underlying trait is quantitative.

the observed test statistic is different from zero is indicated by the p-value. This can be found in specific tables but, if sample size is at least 5 in each group, the distribution of KW is well approximated by a χ^2 distribution with $df = k - 1$ degrees of freedom, so that an asymptotic p-value can be obtained by computing a definite integral.

If the test of Kruskal-Wallis indicates that the null hypothesis should be refuted, then it can be assumed that the medians are significantly different in at least a couple of groups. Analogously to parametric ANOVA, the identification of the couple(s) in which this difference is observed requires a *post-hoc* test to be performed. The groups can be compared couple by couple with a Mann-Whitney's test; the significance threshold for those tests can be adjusted for multiple comparisons with the Bonferroni's method, already discussed in ***B.2.2 One-way analysis of variance (ANOVA)***.

B.2.5. Testing the assumptions of parametric methods

B.2.5.1. Normality of distribution

Normality of a variable's distribution can be verified with different approaches. The most empirical one is to graphically assess the shape of the observed distribution and to compute a skewness index[138] (indicating symmetry) and a Curtosi index (indicating the spread of the curve). If the skewness indexes are given together with their SEs, then symmetry (i.e. skewness=0) could be tested: the test statistic would be the index divided by its SE, and it would have to be compared with a *t* distribution. The fit of the variable's distribution to a normal distribution can also be tested via other methods, like the Kolmogorov-Smirnov's method (suitable if $N \geq 30$) or the Shapiro-Wilk's test; the second was used in the statistical analyses conducted for the present thesis. The Shapiro-Wilk's test, in a nutshell, compares the *N* observed values with a sample of *N* values drawn from a normal distribution. The test statistic is computed as follows:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

The numerator is a non-parametric estimator, based on the linear combination of the values of a normally distributed random variable; $x_{(i)}$ is an order statistic, i.e. the *i*th

[138] The skewness indexes indicate a symmetrical distribution when their value is close to zero, a negative asymmetry (i.e. more observations than expected with low values) if the indexes are below zero, and a positive asymmetry (i.e. more observations than expected with high values) if the indexes are above zero.

smallest value of the sample, whereas a_i represents a weight coefficient calculated so that the numerator gives the best unbiased estimate of the standard deviation under the null hypothesis. The denominator is the sample variance, i.e. a parametric estimate of observed variability in the sample (it should be noted that the degrees of freedom would be $n-1$ both for the numerator and the denominator, so they can be reduced from the formula). The W statistic can vary in $[0,1]$; the null hypothesis of the test is that the distribution of the observed variable fits a normal distribution, i.e. $W=1$, whilst the alternative hypothesis is that data do not fit a normal distribution ($W<1$). The p-value for the computed test statistic can be obtained from specific tables of critical values of the W distribution under the null hypothesis, or they can be approximated; in the analyses presented in this thesis, the approximation proposed by Royston in 1992 was adopted.

It should be noted, however, that the assumption of normality for parametric tests could be overcome with large sample sizes thanks to the central limit theorem, which regards the sampling distribution of the means of a continuous variable. In fact, it can be proven that, if all the possible random samples of size n are drawn from a population with mean μ and standard deviation σ, if n is large enough (roughly more than 25-30 units) the means computed for all those samples would be distributed as $N(\mu,\sigma)$, regardless of the parent distribution of the variable in the population.

B.2.5.2. Homoscedasticity of the variances

Before introducing the methods to test homoscedasticity, it is useful to reflect on the causes of an eventual non-homoscedasticity. This could result from outlier observations or from a wrong study design, so it is necessary to assess if the differences are real or result from one of the abovementioned issues. It should also be considered that, if sample sizes are low, stochastic variability will have stronger impacts on the sample variability than it would have with higher sample sizes. Once this critical thinking has been made, various tests are available for the assumption of homoscedasticity of variances.

The *Bartlett's test* is based on computing the ratio of a deviation comparing group variances over the weighted mean of all the variances, and a correction factor accounting for the number of groups and the *df*. Under the null hypothesis of homoscedasticity[139], this test statistic is distributed as a χ^2. The test assumes that the groups are independent, normally distributed and randomly selected. The *Levene's test* looks at the deviations in each group (i.e. in each dispersion measure), calculating the mean of the deviations within every group. If the means are not significantly different, then the variability is comparable between the groups. These means are compared by means of a test statistic that, under the null hypothesis of homoscedasticity, is distributed as a Fisher's *F*. A robust version of this method, working on medians instead of means, has been proposed by Brown and Forsythe.

For the sake of the analyses presented in this thesis, Levene's robust test was used to test homogeneity of variances between two groups previous to a Student's test, and Bartlett's test was used previous to one-way ANOVA.

[139] It should be paid attention to not confusing this test with another Bartlett's test, aimed at testing the assumption of sphericity required by repeated-measures ANOVA models.

B.3. COMPARISON OF NOMINAL ENDPOINTS BETWEEN GROUPS

Comparing the distribution of a nominal endpoint between groups essentially corresponds to testing the association between two nominal variables, i.e. the endpoint and the group factor. This can be done by using a Pearson's Chi-Squared test for independence or a Fisher's exact test, depending on the circumstances[140]. Those tests have been the founding methods for non-parametrical analyses.

The two variables, which will be assumed to be a dichotomous exposure and a dichotomous outcome for the sake of mathematical simplicity (and coherently with an epidemiological study), are reported in a contingency table[141] defined as follows:

Exposure	***Outcome***		**Total**
	B1	***B2***	
A1	a	b	a+b
A2	c	d	c+d
Total	a+c	b+c	N

The number of subjects for the two categories of A (regardless of B) are expressed by row marginals, and the number of subjects for the two categories of B (regardless of A) are expressed by column marginals. The contingency table might be analysed by row or by column, depending on the interests of the researcher; in any case, both the Pearson's and the Fisher's tests are perfectly symmetrical.

B.3.1. Pearson's Chi-squared test for independence

The Pearson's Chi-squared test for independence[142], proposed by Karl Pearson at the beginning of the XX century, works on the following statistical hypotheses:

- *null hypothesis*: the variables are independent (no association)
- *alternative hypothesis*: the variables are not independent (association).

[140] If the nominal endpoint is dichotomous, then a test of proportions – which is a generalisation of Student's test – could be used.

[141] A contingency table represents the categories of the two variables in rows and columns respectively.

[142] There are other formulations of the Chi-squared test, to be applied in different circumstances (e.g. to test goodness of fit of an observed distribution vs. theoretical expectations, or for repeated measures).

Let's assume that the researcher is interested in investigating if whether the occurrence of the exposure changes between the two outcome groups (however, given that the test is symmetrical, the complementary reasoning could have been done as well). Under the null hypothesis, then, the frequency distribution of the exposure should be equal in the subjects with outcome B1 or B2, i.e. H_o: $\pi_{\mathrm{A}\,|\,B1} = \pi_{\mathrm{A}\,|\,B2}$.

Under the null hypothesis of independence, the frequency distribution that should be expected if the variables are actually independent can be easily calculated. Given that the probability of two independent events is the product of the two individual probabilities, the expected probability of subjects with A1 and can be defined as follows:

$$P(A_1 \cap B_1) = P(A_1)\,P(B_1) = \frac{a+b}{N} \times \frac{a+c}{N}$$

Therefore, under the frequentist definition of probability, the absolute frequency in the sample of size N would be obtained as:

$$f(A_1 \cap B_1) = P(A_1 \cap B_1) \times N = \frac{a+b}{N} \times \frac{a+c}{N} \times N = \frac{(a+b) \times (a+c)}{N}$$

The other cells of the expected contingency table could be completed with a similar procedure (or just by subtraction of the computed cell from the marginals).

The test statistic is based on the cell-by-cell comparison between observed and expected frequencies. It is defined as the sum of the squared deviation between observed (O) and expected (E) frequencies, each one normalised on the expected frequency:

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

The test statistic could be over-estimated if the expected frequencies are low because the denominator would be small and therefore the fraction would be inflated. No agreement has been achieved among theoretical statisticians regarding how small the frequencies should be for the estimator to observe such bias. Cochrane proposed that the test could be applied if no more than 20% of the cells have an expected frequency below 5, on the condition that such frequencies are above 1; a more restrictive criterion is to avoid using this test if any cell in the expected contingency table has a frequency below 5. This restrictive criterion was applied in the analyses conducted in this work.

Analogously, statisticians disagree regarding the use of the continuity correction, proposed by Yates in the Thirties to account for the fact that a contingency table contains discrete values, forming a discrete test statistic that is approximated to a Chi-squared (which is a continuous PDF). It is generally accepted that such adjustment is not needed with large contingency tables, but in the case of 2×2 tables its use is recommended by several authors and discouraged by others [DANIEL, 1996].

Under the null hypothesis, the test statistic is contrasted to a Chi-squared distribution. The family of the Chi-squared PDF is mathematically parented to the standard PDF; each curve is defined on the basis of the *df*[143]. For the test statistic defined above, the *df* are the number of cells in the contingency table where, given the marginals, a value of frequency could be randomly assigned before all the others are obtained by subtraction from the marginals: $df = (I - 1)(J - 1)$. The p-value can be calculated from the distribution ${\chi^2}_{df}$.

B.3.2. Fisher's exact test

Fisher's exact test was proposed almost simultaneously by three authors (Fisher, Irwin and Yates) in the mid-Thirties. Originally, it was formulated as a homogeneity test and was only applicable to 2×2 contingency tables; later, it has been generalised to wider contingency tables and reformulated as a test for independence. Despite its high computational requirements, it is useful as an alternative to the Pearson's test when the circumstances advise against the use of the latter.

The statistical hypotheses are similar to those of the Pearson's test (see ***B.3.1 Pearson's Chi-squared test for independence***), except for the possibility of a one-tailed formulation (however, only the two-tailed test will be presented). Let's assume that, in the contingency table, the number of subjects in group B1 is higher than in B2, i.e. $a + c > b + d$, and that $a/(a + c) > b/(b + d)$. The test statistic is the number of subjects in b. The p-value can be obtained from the tables of critical values or, for large sample sizes, by approximation to a standard distribution.

[143] For *df=1*, the Chi-squared curve is the squared transformation of the standard curve.

B.4. Crude Odds Ratio

An Odds Ratio (OR) is an effect measure which evaluates the strength of the relationship between two factors (the existence of such association can be tested by means of a Pearson's Chi-Squared test for independence). For the sake of simplicity, the circumstance in which the two factors are dichotomous will be presented, and the context of an epidemiological study will be assumed (so, the factors are an *exposure* and an *outcome*).

Let a contingency table be defined as follows:

	Diseased	***Disease-free***	**Total**
Exposed	a	b	a+b
Non-exposed	c	d	c+d
Total	a+c	b+c	N

First, the *odds*[144] of one of the first factor's categories (over the other category) are computed within the groups defined by the other factor. Consistently with the fact that this thesis presents a case-control study, the odds of being exposed (against non-exposed) among cases and among controls will be computed. The mathematical definition of the OR is the following[145]:

$$OR = \frac{odds_{\text{exposed } | \, diseased}}{odds_{exposed \, | \, disease-free}} = \frac{a}{c} \Big/ \frac{b}{d} = \frac{a \times d}{b \times c}$$

The confidence interval at a level $1-\alpha$ is calculated in logarithmic scale:

$$(1-\alpha)CI = e^{\ln OR \pm z_\alpha \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}}$$

The OR can only have positive values, i.e. $OR \in (0, +\infty)$ and indicates how many times the odds of being exposed are higher among diseased people rather than among disease-free people. If $OR < 1$, the odds of being exposed are higher in the disease-free

[144] The *odds* of an event are defined as the probability of success (the event is observed) over the complementary probability of not-success (the event is not observed), i.e. $OR = p(E)/1 - p(E)$.

[145] The estimates obtained by the application of the formula reported here also have the property of being the maximum likelihood estimate of the OR given the data.

group, therefore it can be concluded that the exposure protects the subjects from the outcome; on the contrary, if $OR > 1$, the odds of being exposed are higher among diseased subjects and, consequently, the exposure is interpreted as a risk factor. An OR of 1 would indicate that the exposure has no effect on the outcome, so if two variables are associated at an α significance level, the null value should not be expected to be included in the confidence interval of the estimated OR at the 1-α confidence level.

The mathematical definition of the OR, commented above, makes clear why this measure is typically used in case-control studies: first, it does not depend on the numbers of cases and controls, which in this type of studies are defined by the researcher; second, it is logically coherent with the fact that the subjects are enrolled by outcome status and not by exposure status. However, a similar procedure could be adapted to cohort studies if the odds of being diseased are calculated among exposed and non-exposed people.

As a conclusive remark, it is useful to say that the OR can be estimated also with different formulas, which are suitable for polytomous variables or even quantitative variables; in addition, the *Mantel-Haenszel method* allows a stratified analysis when a confounder is present.

B.5. LOGISTIC REGRESSION

Multivariate analyses, or more correctly *multivariable analyses*, are a series of different statistical techniques aimed at assessing the relationship between a dependent variable (or response variable) and various independent variables (or explicative variables, covariates, prognostic factors, determinants, causes, etc.).

The mathematical models are different depending on the specific technique, which in turn primarily depends on the type of response variable. In this thesis, all multivariate analyses had a nominal dichotomous variable, so that *logistic regression* was used.

Let's consider an outcome Y (response variable) and, for the sake of simplicity, a single dichotomous exposure X (explicative variable), both coded as 0 or 1. With a dichotomous Y, linear regression could not be applied because the mean value of Y would be the probability of Y, so it would be downward and upward limited. To apply a regression, the outcome variable must be transformed prior to specifying the model. If the probability of observing the outcome is considered, i.e. $p = P(Y = 1)$, the odds $p/(1-p)$ could assume any value in $[0, +\infty)$. The logarithmic transformation of the odds, called the *logit* (or *log-odds*) *transformation*, could assume any real value. In formulas:

$$logit(p)\colon \ \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

where β_0 is the intercept[146] of the regression equation and β_1 is the regression coefficient[147] of the exposure X. These parameters of the *logit* model are estimated with a maximum likelihood method (i.e. their estimated values will be the most-likely to have generated the observed data).

If solved for p, the last equation could be rewritten in the following form and, if graphically represented, it would be a peculiar S-shaped curve with codomain $[0,1]$:

$$\hat{p} = \frac{e^{(\beta_0+\beta_1 x)}}{1 + e^{-(\beta_0+\beta_1 x)}}$$

[146] The intercept is the value that the outcome (in this case, the *logit*) should be expected to assume, being zero all the covariates.

[147] The regression coefficient of a certain exposure indicates, for a unit increase in that exposure, how much the *logit* is incremented, being fixed all the other covariates. In the case of a *logit*-transformed outcome, the effects are additive.

The *logit* model could be similarly derived for multiple exposures, each one with its regression coefficient:

$$logit(p) = \beta_0 + \sum \beta_i \, x_i$$

The *logit* model can be transformed so that the dependent variable is not the *log-odds*, but instead an OR, thus obtaining a *logistic model*. The reason for that is the advantage in interpreting the model coefficients as changes in the odds of the outcome, rather than having exponentiated coefficients defining the probability of the outcome. From the definition of OR, and from the equations presented above, it is possible to define the model as follows:

$$OR = \frac{odds_{\exp \mid cases}}{odds_{exp \mid controls}} = \frac{\frac{P(X_1)}{1 - P(X_1)}}{\frac{P(X_0)}{1 - P(X_0)}} = \frac{e^{\beta_0 + \sum \beta_i X_{i_1}}}{e^{\beta_0 + \sum \beta_i X_{i_0}}} = e^{\sum \beta_i (X_{i_1} - X_{i_0})} = \prod e^{\beta_i (X_{i_1} - X_{i_0})}$$

It can be proven that the OR for the exposure X_i (being all the other exposures fixed) is obtained by exponentiating the germane regression coefficient:

$$OR_i = e^{\beta_i}$$

As explained at the beginning of this section, a multivariate model allows to investigate more explicative variables at the same time. The effects, as ORs, are multiplicative rather than additive, so the interpretation is that for a change in the explicative[148], the other explicative variables kept fixed (or "controlling" for the other exposures), the "risk" of being a case rather than a control is increased (or decreased) by a factor *OR*. Adjusted ORs are interpreted exactly as crude ORs (see ***B.4 Crude Odds Ratio***).

The estimation of the model's parameters is done with a maximum likelihood procedure: in short words, the log-likelihood[149] *L* is computed for arbitrarily assigned values of the parameters, until *L* is maximised.

[148] The "change" could be a 1-unit increase (for continuous exposures), being exposed rather than unexposed (for dichotomous exposures), or being in a certain category of exposure rather than in the reference category (for polytomous exposures).

[149] *Likelihood* could be defined as a measure of how strongly a certain estimated value of a parameter is supported by current data or, said differently, a measure accounting for the fact that if an estimate makes the observed data more probable, then it is better supported by data, i.e. "more *likely*" to be a good estimate.

B.5.1. Adding covariates to a regression model

In a regression model, *model specification*, i.e. the set of variables included in the model, can be decided either *a priori*, for "theoretical" reasons, or during the implementation of the model. In the latter case, a variable is generally added in a *stepwise* implementation of the model, which can be *forward* (the new variables are added to a less parametrised model) or *backward* (the variables are dropped from a more parametrised model). Stepwise procedures can be fully automatic or guided by the researcher. In any case, to avoid under- or over-fitting, choices should always be guided by parsimony: this is the epistemological precept known as the *Occam's razor*, or *lex parsimoniae*, enunciated in the XIV century by the Franciscan friar William of Ockham ("*nunquam ponenda est pluralitas sine necessitate*"). In other words, the best model is the most-simple (i.e. less parametrised) possible, which best describes the data by using the less covariates. For this reason, the inclusion of a variable should come after evaluating if it adds something to the model: this is essentially a "cost-benefit" approach and could be pursued by assessing if whether the variable is contributive and/or informative, as detailed in the next two sections. In addition, it is important to carefully discuss the opportunity of adding a new covariate to a multivariate model because – leaving aside its contribution – it could be affected by collinearity or interaction with other covariates or it could be biased by confounding.

Collinearity means that two (or more) variables tend to vary jointly, a fact that could affect the estimate of the regression coefficients by making the model unstable: indeed, a clue that a variable is collinear is the fact that, when added to the model, it makes the regression coefficients of other variables changing considerably. In this circumstance, the easiest thing is to drop one of the collinear variables, usually the less important one from a rational-interpretative point of view. Collinearity can be preliminarily excluded graphically or by computing a correlation coefficient; however, the existence of correlation between two variables is not enough *per se*. Other specific indexes, such as the *Variance Inflation Factors* (VIF), can be computed. In the analyses presented in this thesis, the uncentred VIF was used, and multicollinearity was assessed by adopting a widely used "rule of thumb" of $VIF \geq 10$ or the mean of the VIFs $\geq$ 1[150].

[150] These cut-offs are also suggested in the STATA 13 Manual.

Confounding occurs when the estimated association between an exposure and an outcome is in part attributable to another variable – the *confounder* – which exerts an effect on the same outcome and, at the same time, it is statistically associated with the exposure (the so-called "confounding triangle"). A confounder introduces a background difference, therefore making the comparison of exposed and unexposed subjects inconsistent because of the bias posed by confounding. For this reason, it is pivotal to check the covariates for possible confounding effects by means of appropriate analyses[151]; in case confounding is detected, the analyses should be stratified by the confounder.

Interaction between two covariates occurs when the effect of one variable on the outcome depends on the other variable. If no interaction exists, the effect due to a certain exposure (included as a covariate in a multivariate model) will be the same across all the strata defined on the basis of the other covariate. On the contrary, if such interaction is observed, the effect would change from a stratum to another. Interaction is not always relevant for the interpretation of a model: interaction between two confounders, for instance, would not be worth consideration. On the other hand, if a confounder interacts with an exposure, this would mean that the confounder could be more truly interpreted as an *effect modifier*, which would be extremely relevant to the interpretation of the results. Finally, if the interaction occurs between two exposures, it could indicate a synergic effect (either positive or negative) of the two exposures. The existence of relevant interactions can be inspected during model implementation.

B.5.2. Assessing informativity of a model

The concept of *informativity* accounts for the fact that, when reality is represented through modelling, some information will always be lost. Specifically, the indexes of informativity, called *information criteria*, are used to compare a set of models and decide which is the best in terms of informativity. These indexes represent the relative amount of information that is lost when passing from the real world to a modelled vision of the world; this said, it becomes clear that the lower the information criterion, the better the model. It is important to make clear that information criteria are not

[151] In the present thesis, as defined in Chapter 3 and described in Chapter 4, specific analyses were carried out to exclude confounding.

statistical tests, and that they are not meant to be interpreted as absolute measures: in other words, they "only" indicate the best model among the others, this without prejudice to the fact that the most-informative model in the set could still be largely perfectible.

Various different indexes have been proposed; among them, the most commonly used are the *Akaike's Information Criterion* (AIC)[152] and the *Bayesian Information Criterion* (BIC)[153], both based on the log-likelihood as an estimate of the loss of information[154] and on a penalty that is imposed when new parameters are added[155].

The mathematical definition of the AIC is the following:

$$AIC = -2L + 2k$$

where L is the log-maximum likelihood of the model undergoing the assessment and k is the number of parameters. So, adding a parameter to the model would increase the AIC of two units if the loss of information is not reduced by the new parameter. Looking at the formula, two things are worth commenting: first, the lowest index (i.e. the best model) would be the same regardless of the multiplicator of L; Akaike proposed a value of -2 because the logarithm of the ratio of two maximum likelihoods, multiplied by this value, is asymptotically distributed as a Chi-squared if certain conditions are met. Second, the penalty is obtained by multiplying the number of parameters times the same value (with opposite sign, as useless parametrisation should be discouraged) because K is the asymptotic bias correction for the loss of informativity and it is theoretically justified that it is multiplied by the same quantity of L [BURNHAM & ANDERSON, 2002].

The interpretation of the AIC requires to look at how much the index varies from one model to the other: if the difference is more than ten, the model with the lowest AIC is considerably better; if the difference is between 4 and 7, there is an indication that the

[152] It should be taken into account that computing the AIC when the number of parameters is high respect to the sample size is not recommended (however, this is not the case of the analyses presented in this thesis).

[153] The BIC is called "*Bayesian*" because it was developed in the context of bayesian statistics, but it was actually proposed by Schwarz and is sometimes referred to as SIC, for Schwarz's Information Criterion.

[154] The *loss of information* could be seen as the "distance" between the real process that generated the observed data and the conceptual model that has been specified to describe those data.

[155] Imposing such penalties is necessary to counterbalance the decrease in the loss of informativity that is given by adding parameters, so to make sure that adding parameters does not result in penalising the index only if the new parameters add enough information.

model with the lowest AIC is better, but this statement is less strongly supported; if the difference is between 0 and 2, it is practically considered null (so the two models are nearly equivalent) [BURNHAM & ANDERSON, 2002]. If the two models differ just by one variable, comparing the model with and without that variable can give an indication of how it contributes in reducing the loss of information.

The BIC differs from the AIC in the multiplicator of *K*, which here becomes related to the sample size *N*, so that the penalty is higher with greater sample sizes:

$$BIC = -2L + k \ \ln N$$

It has to be said that the BIC actually differs from AIC also in its theoretical foundation, and for this reason it has been suggested that it is less consistent with the world of biomedical research [BURNHAM & ANDERSON, 2002], but a dissertation on this topic is out of the scopes of this Appendix.

B.5.3. Assessing the contribution of a covariate in a model

The contribution due to a variable when it is added to the regression model can be assessed with the *Likelihood-Ratio test*. This test actually compares the goodness of fit of two models, one with more parameters (*fitted*) and another with less parameters (*null*), if the fitted model is "nested" in the null model (i.e. the fitted model contains all the parameters of the null model, plus others). The test statistic is computed as follows:

$$LR = -2(L_{null} - L_{fitted})$$

where L_{null} is the log-likelihood of the null model and L_{fitted} the log-likelihood of the fitted model. This test statistic is contrasted with a Chi-squared distribution; the *df* are computed as the difference in parametrisation[156] between the two models.

B.5.4. Wald's test of the estimated effect of a covariate

The Wald's test tests a single predicted regression coefficient for a covariate within a model, in order to assess if the effect of that covariate on the outcome is significantly different from a null effect in the specified model.

[156] It must be remembered that, for polytomous variables, the categories are contrasted to the reference category; each contrast adds one parameter to the model (e.g. if the variable has 3 categories, adding the variable will imply that two parameters are added).

The statistical hypotheses are formulated as follows:

- *null hypothesis* H_o: $\beta_i = 0$ (equivalently, $OR = 1$)
  (i.e. the effect of the covariate on the outcome is null);
- *alternative hypothesis* H_1: $\beta_i \neq 0$ (equivalently, $OR \neq 1$)
  (i.e. such effect is not null).

The test statistic, which is contrasted to the standard distribution, is computed as follows[157]:

$$Z = \frac{\widehat{\beta_\iota} - \overbrace{\beta_{null}}^{=0}}{\widehat{ES}(\beta_i)}$$

where $\widehat{\beta_\iota}$ is the maximum-likelihood estimate of the parameter, and its standard error $\widehat{ES}(\beta_i)$ is computed as the standard deviation obtained from the log-likelihood curve at its peak.

[157] The test could be expressed also in quadratic form. In that case, the test statistic is contrasted to a ${\chi^2}_{df=1}$ which is the squared transformation of $N(0,1)$.

C. QUESTIONNAIRE AND ANNEXES

In this Appendix, the Questionnaires used for the survey and the documents that were sent together with the questionnaire are reproduced. All the documents are in Italian.

Table of contents:

UNIVERSITÀ DI PAVIA   Regione Lombardia ASL Pavia   ARPA LOMBARDIA Agenzia Regionale per la Protezione dell'Ambiente Dipartimento di Pavia   Comune di Sannazzaro de Burgondi   Comune di Ferrera Erbognone

Conoscenza e Salute

Progetto CONSAL

Data di compilazione (gg/mese/anno): □□/□□/□□□□

ⓘ **Barrare una sola casella per domanda** ☒□

Genere: □ Maschio □ Femmina   Anno di nascita: □□□□

Peso: □□□Kg   Altezza: □□□cm

Caratteristiche socio-demografiche

1. Stato civile
   - □ a. Non coniugato
   - □ b. Coniugato
   - □ c. Separato/Divorziato
   - □ d. Vedovo
   - □ e. Altro, specificare: ____________________

2. Scolarità
   - □ a. Licenza elementare
   - □ b. Licenza media
   - □ c. Diploma scuola superiore
   - □ d. Laurea
   - □ e. Altro, specificare: ____________________

3. Qual è la sua professione attuale?
   - □ a. Inabile al lavoro
   - □ b. Disoccupato
   - □ c. Occupato saltuariamente
   - □ d. Occupato stabilmente
   - □ e. Casalinga
   - □ f. Studente
   - □ g. Pensionato
   - □ h. Altro, specificare: ____________________

   Se occupato o studente,

   3.1. Quanto tempo passa in media al giorno nel luogo dove lavora o studia? .... Ore □□

   3.2. Specifichi, cortesemente, il tipo di lavoro o studio che svolge:____________________

   ____

   3.3. Specifichi l'indirizzo del luogo attuale di lavoro o studio
   Comune: ____________________ Via e n. civico: ____________________

3.4. Se ha svolto per almeno sei mesi delle professioni precedenti le indichi dalla più recente alla meno recente:

Professione dall'anno all'anno

a. ______________________________ □□□□ □□□□

b. ______________________________ □□□□ □□□□

c. ______________________________ □□□□ □□□□

d. ______________________________ □□□□ □□□□

e. ______________________________ □□□□ □□□□

4. Abita ad un indirizzo diverso da quello di residenza? .... No □ Sì □

4.1. Se si, abita nello stesso comune di residenza? .... No □ Sì □

5. Quanto tempo passa in media al giorno fuori casa per motivi non lavorativi? .... Ore □□

Caratteristiche dell'abitazione

6. Tipologia di abitazione in cui vive

□ a. Appartamento in condominio
□ b. Edificio monofamiliare distanziato da altre case
□ c. Edificio monofamiliare vicino ad altre case
□ d. Villetta a schiera
□ e. Villetta bifamiliare
□ f. Altro, specificare: ______________________________

6.1. Se appartamento in condominio:

Su quale piano si trova la sua abitazione? .... N° piano □□

6.2. Se la tipologia di abitazione indicata è diversa dal condominio:

N° totale di piani dell'edificio (escluso sottotetto) ....N° piani □□

7. Se di fronte alla sua abitazione c'è una strada, che tipo di strada è:
(se ci sono più finestre, scelga quella di fronte alla strada più ampia)

□ a. Autostrada o superstrada o circonvallazione
□ b. Strada principale
□ c. Strada laterale
□ d. Area pedonale/giardino
□ e. Altro, specificare: ______________________________

8. Se di fronte alla sua abitazione c'è una strada, le macchine passano:

□ a. Costantemente
□ b. Frequentemente
□ c. Raramente
□ d. Mai

9. Se di fronte alla sua abitazione c'è una strada, i veicoli pesanti (ad es. camion/autobus) passano:

□ a. Costantemente
□ b. Frequentemente
□ c. Raramente
□ d. Mai

Stili di vita

10. Con quale frequenza settimanale svolge le seguenti attività?

	a) mai	b) Tutti i giorni	c) 3gg o più	d) meno di 3gg	e) occasionalmente
10.1.Corsa	□	□	□	□	□
10.2.Passeggiate all'aria aperta	□	□	□	□	□
10.3.Palestra	□	□	□	□	□
10.4.Piscina	□	□	□	□	□
10.5.Bicicletta	□	□	□	□	□
10.6.Altro		□	□	□	□

se ***Altro***, specificare:____

11. Assume bevande alcoliche?

□ No

□ Sì

Se ***Sì***:

Per ogni bevanda indichi rispettivamente quanto la consuma:

	a) Quotidianamente	b) 3-4 volte a sett.	c) 1-2 volte a sett.	d) saltuariamente	e) mai
11.1.Vino (1 bic)	□	□	□	□	□
11.2.Birra (1 bic. Circa 125ml)	□	□	□	□	□
11.3.Amari e digestivi	□	□	□	□	□
11.4.Superalcolici	□	□	□	□	□
11.5.Altro		□	□	□	□

se ***Altro***, specificare:____

12. Nel corso della sua vita ha mai fumato?

□ No

□ Si

Se ***Sì***:

12.1.A che età ha iniziato a fumare? .... Anni □□

12.2.Fuma attualmente (nell'ultimo mese)? .... No □ Sì □

12.3.Se ha smesso di fumare, specifichi a quale età .... Anni □□

12.4.Quante sigarette fuma o ha fumato al giorno in media? ....n° sigarette □□

13. In questo ultimo anno è stato esposto al fumo passivo? .... No □ Sì □

14. Quanti fumano dentro casa tra chi vi abita, compreso l'intervistato? .... N. □□

Se la risposta è uno o più persone, per favore indichi quante sigarette al giorno vengono fumate in media all'interno della casa da:

14.1.Persona A (l'intervistato) .... n° sigarette □□

14.2.Persona B .... n° sigarette □□

14.3.Persona C .... n° sigarette □□

14.4.Persona D .... n° sigarette □□

15. Consuma le seguenti tipologie di alimenti di provenienza locale (prodotte entro 10 Km)?

15.1. Frutta/verdura .... No □ Sì □
15.2. Uova .... No □ Sì □
15.3. Carne/pesce .... No □ Sì □
15.4. Riso/cereali .... No □ Sì □

Stato di salute

16. In generale direbbe che la sua salute è (nell'ultimo anno):

Scadente	passabile	buona	molto buona	eccellente
□	□	□	□	□

17. Le è mai stata diagnosticata da un medico una o più delle seguenti malattie?

17.1. Aritmie cardiache .... No □ Sì □
17.2. Ipertensione .... No □ Sì □
17.3. Ipercolesterolemia - dislipidemie .... No □ Sì □
17.4. Asma .... No □ Sì □
17.5. Bronchiti cronico-ostruttive .... No □ Sì □
17.6. Malattie dell'apparato digerente .... No □ Sì □
17.7. Diabete .... No □ Sì □

17.8. Altro (specificare) ____

18. È in terapia farmacologica per una o più delle seguenti malattie?

18.1. Aritmie cardiache .... No □ Sì □
18.2. Ipertensione .... No □ Sì □
18.3. Ipercolesterolemia - dislipidemie .... No □ Sì □
18.4. Asma .... No □ Sì □
18.5. Bronchiti cronico-ostruttive .... No □ Sì □
18.6. Malattie dell'apparato digerente .... No □ Sì □
18.7. Diabete .... No □ Sì □

18.8. Altro (specificare) ____

19. Ritiene di aver compilato il questionario con cura e in modo affidabile?

Per nulla	poco	abbastanza	molto
□	□	□	□

La ringraziamo per la Sua cortese collaborazione

1113600001

Questo è un codice anonimizzato, riportato sia in forma di numero che in forma grafica denominata "Codice QR", grazie al quale potremo utilizzare le informazioni da Lei fornite senza identificarLa.

UNIVERSITÀ DI PAVIA

Regione Lombardia
ASL Pavia

ARPA LOMBARDIA
Agenzia Regionale per la Protezione dell'Ambiente
Dipartimento di Pavia

Comune di
Sannazzaro de Burgondi

Comune di
Ferrera Erbognone

Conoscenza e Salute

Progetto CONSAL - Studio I

Gentile Signora/Signore,

Il Dipartimento di Sanità Pubblica, Medicina Sperimentale e Forense dell'Università di Pavia, in collaborazione con le amministrazioni comunali di Sannazzaro de' Burgondi e Ferrera Erbognone, l'ASL di Pavia e l'ARPA Lombardia stanno conducendo uno studio volto a valutare la relazione tra salute e inquinamento ambientale su un campione di residenti nel suo comune (Progetto Conoscenza e Salute, ConSal).

Dagli archivi sanitari dell'ASL di Pavia, risulta che lei è stata/o ricoverata/o per malattie dell'apparato respiratorio e/o cardiovascolare e/o digerente negli ultimi 15 anni.

La invitiamo a partecipare all'indagine rispondendo ad alcune domande di approfondimento sulla sua salute e sulle sue abitudini di vita raccolte in un questionario a risposte chiuse.

Il **questionario compilato** e la **dichiarazione di consenso firmata** dovranno essere riconsegnati nei punti di raccolta presso **l'ufficio anagrafe del comune, la farmacia o il poliambulatorio** nella busta bianca che troverà allegata e che dovrà sigillare.

La invitiamo cortesemente a riconsegnare il questionario entro due settimane da quando lo ha ricevuto.

Le informazioni raccolte attraverso il questionario saranno trattate in modo anonimizzato, ossia senza nome e cognome. Il questionario riporta un codice identificativo non ricollegabile direttamente ai Suoi dati anagrafici, i quali sono conservati separatamente. Solo il responsabile del trattamento dei dati o un suo delegato, in casi di motivata necessità, potranno collegare il codice ai Suoi dati anagrafici, che comunque non saranno oggetto di diffusione.

I dati raccolti tramite il questionario, trattati anche per mezzo di opportuni strumenti informatici, saranno valutati esclusivamente in modo aggregato, insieme cioè a quelli di tutte le altre persone selezionate per l'indagine; secondo questa stessa modalità (anonima e aggregata) potrebbero essere diffusi in occasione di relazioni, pubblicazioni scientifiche o convegni.

I dati saranno conservati in appositi archivi protetti, in forma sia cartacea che informatizzata, per un tempo di 20 anni a decorrere dall'inizio dello studio; trascorso questo periodo saranno definitivamente distrutti. Sarà nostra massima cura garantire un'adeguata tutela della Sua identità e di tutte le informazioni che vorrà fornirci.

La partecipazione al presente studio è un'opportunità fondamentale per conoscere lo stato di salute della cittadinanza e rappresenta un beneficio per tutta la popolazione. I risultati che si otterranno si trasformeranno in strumenti per le scelte di politica sanitaria e ambientale nel nostro territorio. L'adesione allo studio non comporta rischi, bensì è una occasione importante che si ha per accrescere le conoscenze scientifiche sugli effetti che può avere l'inquinamento ambientale.

La ringraziamo per la sua attenzione e per il tempo che ci ha voluto dedicare.

In caso ci voglia contattare può farlo utilizzando l'indirizzo di posta elettronica *consal@unipv.it* o telefonicamente al numero 0382 98 7180, dove risponderà il Dr. Marco Gnesi del Dipartimento di Sanità Pubblica, Medicina Sperimentale e Forense dell'Università di Pavia.

Cordiali Saluti
Prof. Gabriele Pelissero
Dipartimento Sanità Pubblica, Medicina Sperimentale e Forense
Università di Pavia
Via Forlanini 2, 27100 Pavia

UNIVERSITÀ DI PAVIA

Regione Lombardia ASL Pavia

ARPA LOMBARDIA Agenzia Regionale per la Protezione dell'Ambiente Dipartimento di Pavia

Comune di Sannazzaro de Burgondi

Comune di Ferrera Erbognone

Conoscenza e Salute

Progetto CONSAL - Studio I

Gentile Signora/Signore,

Il Dipartimento di Sanità Pubblica, Medicina Sperimentale e Forense dell'Università di Pavia, in collaborazione con le amministrazioni comunali di Sannazzaro de' Burgondi e Ferrera Erbognone, l'ASL di Pavia e l'ARPA Lombardia stanno conducendo uno studio volto a valutare la relazione tra salute e inquinamento ambientale su un campione di residenti nel suo comune (Progetto Conoscenza e Salute, ConSal).

Dagli archivi sanitari dell'ASL di Pavia, risulta che lei **non** è stata/o ricoverata/o per malattie dell'apparato respiratorio e/o cardiovascolare e/o digerente negli ultimi 15 anni.

La invitiamo a partecipare all'indagine rispondendo ad alcune domande di approfondimento sulla sua salute e sulle sue abitudini di vita raccolte in un questionario a risposte chiuse.

Il **questionario compilato** e la **dichiarazione di consenso firmata** dovranno essere riconsegnati nei punti di raccolta presso **l'ufficio anagrafe del comune, la farmacia o il poliambulatorio** nella busta bianca che troverà allegata e che dovrà sigillare.

La invitiamo cortesemente a riconsegnare il questionario entro due settimane da quando lo ha ricevuto.

Le informazioni raccolte attraverso il questionario saranno trattate in modo anonimizzato, ossia senza nome e cognome. Il questionario riporta un codice identificativo non ricollegabile direttamente ai Suoi dati anagrafici, i quali sono conservati separatamente. Solo il responsabile del trattamento dei dati o un suo delegato, in casi di motivata necessità, potranno collegare il codice ai Suoi dati anagrafici, che comunque non saranno oggetto di diffusione.

I dati raccolti tramite il questionario, trattati anche per mezzo di opportuni strumenti informatici, saranno valutati esclusivamente in modo aggregato, insieme cioè a quelli di tutte le altre persone selezionate per l'indagine; secondo questa stessa modalità (anonima e aggregata) potrebbero essere diffusi in occasione di relazioni, pubblicazioni scientifiche o convegni.

I dati saranno conservati in appositi archivi protetti, in forma sia cartacea che informatizzata, per un tempo di 20 anni a decorrere dall'inizio dello studio; trascorso questo periodo saranno definitivamente distrutti. Sarà nostra massima cura garantire un'adeguata tutela della Sua identità e di tutte le informazioni che vorrà fornirci.

La partecipazione al presente studio è un'opportunità fondamentale per conoscere lo stato di salute della cittadinanza e rappresenta un beneficio per tutta la popolazione. I risultati che si otterranno si trasformeranno in strumenti per le scelte di politica sanitaria e ambientale nel nostro territorio. L'adesione allo studio non comporta rischi, bensì è una occasione importante che si ha per accrescere le conoscenze scientifiche sugli effetti che può avere l'inquinamento ambientale.

La ringraziamo per la sua attenzione e per il tempo che ci ha voluto dedicare.

In caso ci voglia contattare può farlo utilizzando l'indirizzo di posta elettronica *consal@unipv.it* o telefonicamente al numero 0382 98 7180, dove risponderà il Dr. Marco Gnesi del Dipartimento di Sanità Pubblica, Medicina Sperimentale e Forense dell'Università di Pavia.

Cordiali Saluti
Prof. Gabriele Pelissero
Dipartimento Sanità Pubblica, Medicina Sperimentale e Forense
Università di Pavia
Via Forlanini 2, 27100 Pavia

UNIVERSITÀ DI PAVIA

Regione Lombardia
ASL Pavia

ARPA LOMBARDIA
Agenzia Regionale per la Protezione dell'Ambiente
Dipartimento di Pavia

Comune di
Sannazzaro de Burgondi

Comune di
Ferrera Erbognone

Conoscenza e Salute

Progetto CONSAL - Studio I

DICHIARAZIONE DI CONSENSO

La/Il Sottoscritta/o ____ acconsente a partecipare al progetto "Conoscenza e Salute" (ConSal) e autorizza il Dipartimento di Sanità Pubblica, Medicina sperimentale e forense dell'Università di Pavia al trattamento dei dati personali e sensibili, in conformità al D.Lgs. 196/2003 e successive integrazioni.

La/Il Sottoscritta/o è informata/o che:

- i dati raccolti saranno analizzati dal Dipartimento dell'Università di Pavia nell'ambito del progetto, nel rispetto delle vigenti disposizioni in materia di tutela dei dati personali e con l'adozione delle misure di sicurezza prescritte dal Codice sulla Privacy;
- fornire i dati è facoltativo e chi non vorrà fornirli non farà parte dello studio;
- la raccolta dei dati è fondamentale ai fini del progetto ed i dati stessi saranno trattati in modo anonimizzato e non riconducibile ai dati anagrafici;
- i dati personali e sensibili possono essere comunicati per le finalità di cui sopra, ai responsabili ed incaricati del trattamento del Dipartimento ed in forma anonima alle altre organizzazioni, istituzioni o enti facenti parte del progetto e non saranno in alcun modo oggetto di diffusione;
- il Responsabile del trattamento dei dati è il Professor Gabriele Pelissero, Dipartimento di Sanità Pubblica, Medicina sperimentale e forense dell'Università di Pavia;
- in ogni momento potrà esercitare i suoi diritti ai sensi dell'articolo 7 del D.Lgs 196/2003.

DATA ________________

FIRMA

____

D. Causes of hospital admission

The distribution of the main cause of the first hospital admission (as recorded in the SDOs database) is listed in ***Table D.1***, together distribution of the specific ICD codes within their macro-area. The descriptions of the ICD codes are reported in the list below, by chapter:

- *Chapter 7 – Diseases of the circulatory system:*
  - 401 Essential hypertension;
  - 402 Hypertensive heart disease;
  - 410 Acute myocardial infarction;
  - 411 Other acute and subacute forms of ischemic heart disease;
  - 413 Angina pectoris;
  - 414 Other forms of chronic ischemic heart disease;
  - 421 Acute and subacute endocarditis;
  - 427 Cardiac dysrhythmias;
  - 428 Heart failure;
  - 430 Subarachnoid haemorrhage;
  - 434 Occlusion of cerebral arteries;
  - 435 Transient cerebral ischemia;
  - 436 Acute, ill-defined cerebrovascular disease;
  - 442 Other aneurysm;
  - 447 Other disorders of arteries and arterioles;
  - 451 Phlebitis and thrombophlebitis;
  - 453 Other venous embolism and thrombosis;
  - 455 Haemorrhoids;
  - 458 Hypotension.
- *Chapter 8 – Diseases of the respiratory system:*
  - 462 Acute pharyngitis;
  - 463 Acute tonsillitis;
  - 464 Acute laryngitis and tracheitis;
  - 466 Acute bronchitis and bronchiolitis;
  - 475 Peritonsillar abscess;

    - 478 Other disease of upper respiratory tract;
    - 481 Pneumococcal pneumonia;
    - 482 Other bacterial pneumonia;
    - 486 Pneumonia, organism unspecified;
    - 487 Influenza;
    - 493 Asthma;
    - 515 Post-inflammatory pulmonary fibrosis.
- *Chapter 9 – Diseases of the digestive system:*
    - 522 Diseases of pulp and periapical tissues;
    - 530 Diseases of oesophagus;
    - 531 Gastric ulcer;
    - 536 Disorders of function of stomach;
    - 540 Acute appendicitis;
    - 550 Inguinal hernia;
    - 551 Other hernia of abdominal cavity, with gangrene;
    - 553 Other hernia of abdominal cavity, without obstruction/gangrene;
    - 555 Regional enteritis;
    - 556 Ulcerative colitis;
    - 558 Other and unspecified non-infectious gastroenteritis and colitis;
    - 560 Intestinal obstruction without mention of hernia;
    - 562 Diverticula of intestine;
    - 567 Peritonitis and retroperitoneal infections;
    - 568 Other disorders of peritoneum;
    - 571 Chronic liver disease and cirrhosis;
    - 574 Cholelithiasis;
    - 577 Diseases of pancreas.
- *Chapter 16 – Symptoms, signs, and ill-defined conditions:*
    - 785 Symptoms involving cardiovascular system;
    - 786 Symptoms involving respiratory system / other chest symptoms.

General cause, ICD Chap.		n (%)	ICD Code	n (%)
Cardiovascular conditions	*Chap. 7*	59 (47.2%)	401	2 (1.6%)
			402	4 (3.2%)
			410	14 (11.2%)
			411	7 (5.6%)
			413	3 (2.4%)
			414	5 (4.0%)
			421	1 (0.8%)
			427	9 (7.2%)
			428	1 (0.8%)
			430	1 (0.8%)
			434	3 (2.4%)
			435	1 (0.8%)
			436	1 (0.8%)
			442	1 (0.8%)
			447	1 (0.8%)
			451	1 (0.8%)
			453	1 (0.8%)
			455	2 (1.6%)
			458	1 (0.8%)
	Chap. 16	1 (0.8%)	785	1 (0.8%)
Respiratory conditions	*Chap. 8*	19 (15.2%)	462	1 (0.8%)
			463	1 (0.8%)
			464	1 (0.8%)
			466	2 (1.6%)
			475	1 (0.8%)
			478	2 (1.6%)
			481	1 (0.8%)
			482	1 (0.8%)
			486	5 (4.0%)
			487	1 (0.8%)
			493	2 (1.6%)
			515	1 (0.8%)
	Chap. 16	4 (3.2%)	786	4 (3.2%)
Gastrointestinal conditions	*Chap. 9*	42 (33.6%)	522	1 (0.8%)
			530	1 (0.8%)
			531	1 (0.8%)
			536	1 (0.8%)
			540	5 (4.0%)
			550	9 (7.2%)
			551	1 (0.8%)
			553	1 (0.8%)
			555	2 (1.6%)
			556	2 (1.6%)
			558	1 (0.8%)
			560	1 (0.8%)
			562	3 (2.4%)
			567	1 (0.8%)
			568	1 (0.8%)
			571	1 (0.8%)
			574	8 (6.4%)
			577	2 (1.6%)
Total		125 (100.0%)	Total	125 (100.0%)

Table D.1 *– Causes of hospital admissions of the cases among the respondents to the survey.*