# Introduction {#intro}

With extension of life expectancy over the last few decades, dementia has become a major burden that affects the elderly. In 2016, the global number of individuals who lived with dementia was over 40 [@gbd2016] million and by 2050 this number is expected to triplicate [@worldreport2018]. In 2020, deaths due to Alzheimer’s disease and other dementias have increased, becoming the 7th leading cause of death globally and overtaking stroke to become the second leading cause in high-income countries. Although dementia risk does not represent a leading cause of disease and death in low-middle income countries, it is projected to increase as the burden of preventable and curable diseases reduce[@who2020]. Likewise, women are disproportionately affected, worldwide [@women2015]. Furthermore, the burden of dementia does not only affect those who have the disease: dementia has large effects on the lives of caregivers, families, and health-care systems. Unequal distribution of opportunities, responsibilities and societal roles push women into the caregiver role more than men, which hampers even more their access to paid work and health, especially in poor and marginalized areas, creating negative feedback loops that increases all kinds of disparities [@swinkels2019; @brodaty2009; @etters2008]. However, these harmful consequences can be prevented if the burden of dementia is reduced, and this can be done by targeting prevention and delay of onset[@carrillo2013].

The challenge to reduce the burden of dementia is that this disease is complex and heterogeneous, with multiple etiologic and neuropathological processes likely related. The field is constantly focused on identifying the modifiable risk factors and therapies. To this purpose, in 2020, the Lancet Commission released an updated guideline with evidence on twelve modifiable risks factors, which would account for 40% of worldwide dementias that could have been prevented or delayed. These modifiable risk factors include: lower education, hypertension, hearing impairment, smoking, obesity, depression, physical inactivity, diabetes, low social contact, alcohol consumption, traumatic brain injury (TBI), and air pollution [@dementia\_lancet]. Furthermore, there is a growth on brain biomarker research, with the simultaneous intention of detecting proxies for early diagnosis, as well as identifying new molecular targets for intervention. Since there are few specific drugs that target amyloid and tau production, with small evidence of effect, new studies for drugs that target other mechanisms are in study. Drug repositioning and repurposing research is becoming more popular, offering a valuable alternative route for the identification of effective disease-modifying treatments for dementia[@ballard2020; @langedijk2015]. In 2020, Cummings et al. identified 121 agents in clinical trials for the treatment of Alzheimer's disease, out of which 43% (57) represented repurposed agents across all phases of the pipeline [@cummings2020].

All these objectives in dementia research are heavily reliant on observational studies and current availability of multiple sources of large data give us the opportunity to expand the field. Nevertheless, new data sources and sophisticated computational software and technology, may sometimes overshadow the process of asking clear questions and the steps to tie the questions to the data. In a time where we have deeply embraced that "causation is not correlation", the concept of "associations" and molding questions to hypothesis testing, has deeply overshadowed the critical step in research to clearly define questions and wanted interpretations. Acknowledging that etiologic research in dementia research is aimed at identifying causal effects, is probably the first challenge and impediment to conceptualize clear causal questions. This may be due to the misconception that causality can only be inferred from randomized controlled studies (RCTs), which is reinforced by high-impact journals such as JAMA[@jama]. Another reason why novice students may struggle to embrace causal thinking is due to the polarized debates among causal inference experts, about how to conceptualize exposures that cannot be intervened upon or manipulated (such as sex, race or BMI)[@waterkills; @schwartz2016]. During my first years of training in causal inference, this debate confused me and filled me with insecurities about how to study exposures related to dementia when I did not have measurements on interventional data. Nevertheless, the reader may find in the following chapters how I embraced this debate and developed my own criteria on how to study causal questions.

Throughout this book I will develop how the creative process of thinking and imagining what is what we truly want to ask helps to formulate clear questions, identify potential sources of bias, and articulate the analytic methods to match the question. To this matter, I will embrace the causal inference theory and corresponding methods and implement the target trial emulation framework to study the effect of several exposures (or potential targets of intervention) that had controversial results in previous observational studies. [@hernan\_robins\_2016; @labrecque2018]. To answer each question I will use data collected for the Rotterdam Study, a population-based cohort study with rich longitudinal data assessments [@ikram2020], described in the next section.

To give further context, in the decade of the sixties, Cochran reinforced the notion that observational studies were aimed to answer questions about cause-effect when randomized studies were not feasible or ethical [@cochran1965]. In the eighties, Robins formalized how to answer causal questions with observational studies following the RCT principles, and decades after the “target trial emulation framework” was conceived [@hernan\_robins\_2016]. Labrecque and Swanson define the target trial emulation as “the application of design principles from randomized trials to the analysis of observational data, explicitly tying the analysis to the trial it is emulating” [@labrecque2018]. This process includes specifying and emulating the key components of the trial protocol such as the eligibility criteria, treatment strategies, treatment assignment, follow-up period, outcome, causal contrasts and statistical analysis. The refinement of the causal question, and the target trial specifications will often be a back and forward process between the question we truly aim to answer and the availability of data to answer that question[@labrecque2018; @whatif2020].

Taking in consideration this framework and principles, in \*\*Chapter \@ref(chapter2)\*\* I present a target trial to study the effect of statins treatment in the 10-year risk of dementia and death. This work brings clarity to the idea that even in observational studies we can formulate causal contrasts like the intention-to-treat (ITT) and the per-protocol (PP) effect, as in pragmatic trials. In this setting, the intention-to-treat effect refers to a combination of the effect of the treatment under study and of any other patient and physician's behavioural changes triggered by the assignment itself. This effect is agnostic of any treatment decisions made after baseline, which makes it difficult to interpret to patients, clinicians and other decision-makers [@murray2019; @murray2018]. Thought this is what makes it appealing from an analytic perspective, since it can be conceptualized as a point treatment strategy. Instead, the per-protocol effect represents the effect of being assigned a treatment strategy and adhering to that assigned treatment strategy through-out follow-up, as specified in the study protocol. This effect can be conceptualized as a static or dynamic treatment strategy, since adherence to a treatment strategy over follow-up will depend on the evolution of an individual's time-varying covariates [@whatif2020]. Being explicit about the treatment strategy emphasizes on the necessity to collect data on the treatment adherence over follow-up, as well as time-varying confounders and predictors of adherence. It also introduces the major challenge with time-varying treatments, the treatment-confounder feedback.

Treatment-confounder feedback refers to the setting where, time-varying covariates affect treatment over time, but additionally, time-varying covariates are affected themselves by prior treatment[@robins1986; @whatif2020]. If we could conceptualize the modifiable risks factors proposed by the Lancet commission as potential targets of intervention, and define interventions as static or dynamic, we must conceptualize the potential treatment-confounder feedback loops that might challenge both randomize trials and observational studies. For example, the Lancet commission defines hypertension as one of the modifiable risk factors, though most references from observational studies considered blood pressure as a single time-point measurement. In \*\*Chapter \@ref(chapter3)\*\* I illustrate how to conceive a question where the interest is focused in learning how much would the risk of stroke and dementia change if, hypothetically, we could reduce and keep systolic blood pressure under different thresholds defined in clinical practice over follow-up, resembling a per-protocol effect. Given that blood pressure and hypertension is affected by other comorbidities, and it affects other comorbidities as well, proper analysis that accounts for treatment-confounder feedback was needed. Since traditional statistical methods cannot account for this feature, one of the highlights of this dissertation is the application of “\_G-methods\_”. G-methods (or generalized methods) are a set of causal models and analytic methods proposed by Robins beginning in 1986, consisting in the g-formula, marginal structural models and structural nested models [@robins1986; @whatif2020]. These methods have revolutionized the field of epidemiology and public health, by providing analytic tools to answer causal questions with longitudinal treatment/exposure data [@richardson2014].

In addition to the target trial framework and the G-methods, another key feature of causal inference methods is the development and application of directed acyclic graphs (DAGs). These causal diagrams are another representation of causal structures which allows to draw and visualize assumptions. DAGs were introduced by Judea Pearl [@pearl1994] and extended by Robins to settings of time-varying exposures[@robins]. In \*\*Chapter \ref(chapter4)\*\* I progressively build a DAG to represent a causal question of the effect of a Pin1-targeting drug on the risk of dementia, when we only have cancer diagnosis measured as the proxy for Pin-1. This project was motivated to understand all potential sources of bias that could be related to inverse association between cancer diagnosis and dementia that had been systematically reported in previous observational studies. Since we cannot understand bias if we do not have a clear causal question to start with, causal graphs helped elucidate all the steps to connect the unmeasured mechanism of interest to the observed data outlining the data generation process.

Just as it is important to clearly define what we mean by the exposure or intervention of interest, we must focus on other elements of a clear research question, such as how our question incorporates competing events and other censoring events. Given that more than 30% of the participants died prior to dementia diagnosis in the Rotterdam Study, death played a major role as a competing event in all the projects of this dissertation. At the same time I was drafting one of the first projects of this dissertation, a pre-print about competing events in a causal inference framework, by Young et al. [@young2020] was published. This paper helped me understand that death was not something to “fix” within the analysis of data, but rather to include as part of the question. This goes in hand with the definition of “estimands” by the ICH9 addendum [@ich9] that considers post-randomization events, defined as intercurrent events, that can affect outcome assessment or interpretation, as part of this research question.

As opposed to previous recommendations that suggest to use a “cause-specific hazard model” for etiologic research, and a “Fine and gray sub-distribution hazard model” for prediction research [@lau2009], Young et al. [@young2020] disengage from this recommendation, and rather start by framing the different estimands that are approached by traditional methods in survival analysis. Young et al. formalized how, under explicit assumptions, they allow for identification of different estimands, and present both directed acyclic graphs and single intervention world graphs for settings with competing events. The two causal questions or estimands discussed are the “controlled direct effect” and the “total effect”. The controlled direct effect responds to a question where death could have been prevented, hypothetically. The total effects responds to a question where death can also happen through-out the follow-up, thus it captures the effect mediated through death. In each chapter of this dissertation, I apply either of these estimands (or both), depending on the research aim, proving that different estimands will be better suited for different context. In \*\*Chapter \@ref(chapter5)\*\* I present a systematic review to describe the current practices and interpretations of longitudinal studies of dementia, where death plays the role of a competing event. Unfortunately, this work highlights how much in real applications censoring is treated as a synonym for ignoring and the large gap between methods development and applied research. This reality motivated me to write \*\*Chapter \@ref(chapter6)\*\*, were I will highlight the key conceptual definitions in regards to competing events in causal inference, with an applied example on smoking cessation in the risk of dementia, to make this novel framework accessible to applied researchers in the field.

To finalize, in \*\*Chapter \@ref(chapter7)\*\* I distill all the learning experiences from overcoming different methodological challenges in each project, the broader implications of my research and discuss the unsolved challenges as future lines of research. And, as a way to express my growth as a researcher and epidemiologist within this PhD trajectory, in \*\*Chapter \@ref(chapter8)\*\* I will zoom out from the current work in dementia research and discuss more broadly how this pandemic has shaped my understanding of becoming an epidemiologist interested in methods development.

\newpage

## Study setting

All the projects in this dissertation were designed and implemented using data from The Rotterdam Study, a population-based cohort that recruited participants living in the district of Ommoord, in Rotterdam, the Netherlands. The cohort was recruited from all inhabitants aged 55 years and older. Participants were invited in random clusters, through sampling from the municipal register[@hofman1991]. The Rotterdam Study, also known in the Netherlands as Erasmus Rotterdam Gezondheid Onderzoek (ERGO), started as a pilot study in July, 1989 and full recruitment and complete data acquisition started in January, 1990 and finished in September, 1993. Out of 11850 eligible participants, 7983 accepted (78% response rate). Participants were followed every three years, follow-up visits were held between 1993-1995, 1997-1999, 2002-2005, 2008-2014; however, time between visits varied for each participant ranging from 1 to 7 years between them.

In 1999, members of Ommoord who had become 55 years of age or moved into the study district since the recruitment were invited to participate; out of 4472 invitees, 3011 new participants were included to the Rotterdam Study [@hofman2007] between February 1999 and December 2001, with follow-up visits between 2002-2005 and 2008-2012. In 2006, the cohort expanded to with a new wave of recruitment, including Ommord participants who were 40 years and older. This recruitment represented 3932 new participants in the Rotterdam Study. Last, between 2016 and December 2020, a new wave of recruitment included 3005 new participants. For this reason, the total sample size of the Rotterdam Study is around 18000 participants, though this number does not represent the total sample size observed at a specific time-point, due to the active recruitment and since several participants may have died or were loss-to follow-up before the latter waves of recruitment. Data collection dates and representation of the recruitment waves included in this dissertation with corresponding sample size are presented in Figure 1.

When the Rotterdam Study was conceived, it was focused on four primary areas of research: neurogeriatric diseases, cardiovascular locomotor and ophthalmologic diseases, though it later expanded to explore several other areas of disease. Hofman et al defined the aims of the study as follows: 1) to investigate the determinants of diseases in order to assess their etiologic significance; 2) to investigate potentially modifiable determinants in order to be able to develop preventive strategies by providing specific recommendations for intervention studies[@hofman1991]. Thus, participants went through extensive examinations, including interviews, physical examination and collecting bio-samples for molecular and genetic studies. Furthermore, the Rotterdam Study integrated information from secondary sources; for example, data on death was collected from municipal registries from the early start. Since 1991, data from all pharmacies serving in Ommoord region was integrated and, linkage to the Dutch Hospital Data, which captures main discharge diagnosis from all nationwide hospital admissions was stablished in 1998. Several other examples of data integration are discussed by Ikram et. al [@ikram2020]. The immense effort to capture all this information about participants makes the Rotterdam Study a unique source to answer questions about time-varying exposures or interventions that could reduce the risk of dementia and other age-related diseases.

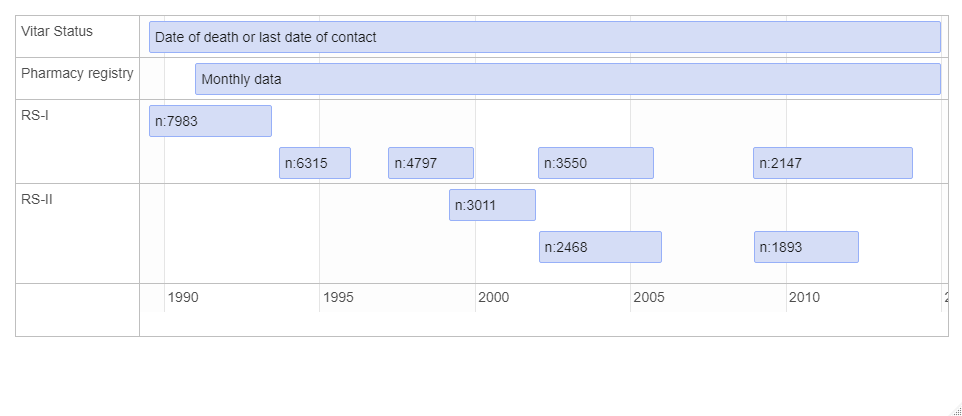


Figure 1. The Rotterdam Study data timeline and corresponding sample size (n). RS-I represents the first wave of recruitment and RS-II represents the second wave of recruitment. Though more recruitment waves and visits were held, this graph only represents the sources of information used in this dissertation.

\newpage

## References