UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Head Office: Università degli Studi di Padova

Department of Cardiac, Thoracic, Vascular Sciences and Public Health

Ph.D. COURSE IN: Translational Specialistic Medicine 'G.B. Morgagni'
CURRICULUM: Biostatistics and Clinical Epidemiology
SERIES: XXXIV

ANALYSIS OF HEALTH OUTCOMES AND COMORBIDITY PATTERNS IN PATIENTS WITH CHRONIC DISEASES. FORECAST MODELS AND PHENOMAPPING ON INTEGRATED ADMINISTRATIVE DATABASES.

Thesis written with the financial contribution of the Unit of Epidemiology, Regional Health Service ASL TO3.

Coordinator: Prof. Annalisa Angelini
Supervisor: Prof. Paola Berchialla

Ph.D. Student: Veronica Sciannameo

Ai nonni
Anna, Emilia, Leonardo, Marco

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SUPPLEMENTARY TABLES

LIST OF PUBLICATIONS

WITHIN THE FIRST TWO AUTHORS' NAMES

Opportunistic screening for type 2 diabetes in community pharmacies. Results from a region-wide experience in Italy. Gnavi R, Sciannameo V, Baratta F, Scarinzi C, Parente M, Mana M, Giaccone M, Cavallo Perin P, Costa G, Spadea T, Brusa P. PLoS One. 2020 Mar 18;15(3): e0229842. doi: 10.1371/journal.pone.0229842. eCollection 2020. PMID: 32187210.

Enrolment criteria for diabetes cardiovascular outcome trials do not inform on generalizability to clinical practice: The case of glucagon-like peptide-1 receptor agonists. Sciannameo V, Berchialla P, Orsi E, Lamacchia O, Morano S, Querci F, Consoli A, Avogaro A, Fadini GP; DARWIN-T2D study. Diabetes Obes Metab. 2020 May;22(5):817-827. doi: 10.1111/dom.13962. Epub 2020 Feb 6. PMID: 31943710.

Similar effectiveness of dapagliflozin and GLP-1 receptor agonists concerning combined endpoints in routine clinical practice: A multicentre retrospective study. Fadini GP, Sciannameo V, Franzetti I, Bottigliengo D, D'Angelo P, Vinci C, Berchialla P, Arena S, Buzzetti R, Avogaro A; DARWIN-T2D network. Diabetes Obes Metab. 2019 Aug;21(8):1886-1894. doi: 10.1111/dom.13747. Epub 2019 May 8. PMID: 30985052.

Transposition of cardiovascular outcome trial effects to the real-world population of patients with type 2 diabetes. Sciannameo V, Berchialla P, Avogaro A, Fadini GP; DARWIN-T2D Network. Cardiovasc Diabetol. 2021 May 10;20(1):103. doi: 10.1186/s12933-021-01300-y. PMID: 33971880.

Adjustment for baseline covariates to increase efficiency in RCTs: a comparison of Bayesian and frequentist approaches. Berchialla P., Sciannameo V., Urru S., Lanera C., Azzolina D., Gregori D., Baldi I. Int. J. Environ. Res. Public Health 2021, 18, 7758. https://doi.org/10.3390/ijerph18157758

Incidence and prevalence analysis of Non-Small-Cells and Small-Cells Lung Cancer using Administrative Data. A. Ricotti, V. Sciannameo, E. Ferracin, I Massa, W Balzi, A Roncadori, P Canavese, A Avitabile, and Paola Berchialla. Int. J. Environ. Res. Public Health 2021, 18(17), 9076; https://doi.org/10.3390/ijerph18179076.

Under Review

Targeted maximum likelihood estimation of treatment effectiveness under outcome data missingness and model misspecification: a simulation study to assess results from the DARWIN T2D study. Sciannameo V, Fadini GP, Bottigliengo D, Avogaro A, Baldi I, Gregori D, Berchialla P. Journal of Medical Systems

Deep Learning for predicting urgent hospitalizations in elderly population using administrative Electronic Health Records. Sciannameo V, Jahier Pagliari D, Ferracin E, Ricotti A, Ricceri F, Costa G, Berchialla P. Artificial Intelligence in Medicine.

CONTRIBUTOR

Alcohol use and misuse: a profile of adolescents from 2018 Italian HBSC data. Charrier L, Canale N, Dalmasso P, Vieno A, Sciannameo V, Borraccino A, Lemma P, Ciardullo S, Berchialla P; 2018 HBSC-Italia Group; the 2018 HBSC-Italia Group. Ann Ist Super Sanita. 2020 Oct-Dec;56(4):531-537. doi: 10.4415/ANN_20_04_18. PMID: 33346182

COVID-19 infection and diffusion among the healthcare workforce in a large university-hospital in northwest Italy. Garzaro G, Clari M, Ciocan C, Grillo E, Mansour I, Godono A, Borgna LG, Sciannameo V, Costa G, Raciti IM, Bert F, Berchialla P, Coggiola M, Pira E. Med Lav. 2020 Jun 26;111(3):184-194. doi: 10.23749/mdl. v111i3.9767. PMID: 32624560.

Connectedness as a protective factor in immigrant youth: results from the Health Behaviours in School-aged Children (HBSC) Italian study. Borraccino A, Berchialla P, Dalmasso P, Sciannameo V, Vieno A, Lazzeri G, Charrier L, Lemma P. Int J Public Health. 2020 Apr;65(3):303-312. doi: 10.1007/s00038-020-01355-w. Epub 2020 Apr 4. PMID: 32248262.

Characteristics of Breakthrough Pain and Its Impact on Quality of Life in Terminally Ill Cancer Patients. Gonella S, Sperlinga R, Sciannameo V, Dimonte V, Campagna S. Integr Cancer Ther. 2019 Jan-Dec; 18:1534735419859095. doi: 10.1177/1534735419859095. PMID: 31220961 Free PMC article.

Machine learning in clinical and epidemiological research: isn't it time for biostatisticians to work on it? The Machine Learning in Clinical Research Group - Danila Azzolina, Ileana Baldi, Giulia Barbati, Paola Berchialla, Daniele Bottigliengo, Andrea Bucci, Stefano Calza, Pasquale Dolce, Valeria Edefonti, Andrea Faragalli, Giovanni Fiorito, Ilaria Gandin, Fabiola Giudici, Dario Gregori, Caterina Gregorio,

Francesca Ieva, Corrado Lanera, Giulia Lorenzoni, Michele Marchioni, Alberto Milanese, Andrea Ricotti, Veronica Sciannameo, Giuliana Solinas, Marika Vezzoli. (2019). Epidemiol. Biostatist. Public Health 16:e13245. doi: 10.2427/13245

Factors Associated with Missed Nursing Care in Nursing Homes: A Multicentre Cross-Sectional Study. Sara Campagna, Alessio Conti, Marco Clari, Ines Basso, Veronica Sciannameo, Paola Di Giulio, Valerio Dimonte. Int J Health Policy Manag. 2021 Apr 21. doi: 10.34172/ijhpm.2021.23. Online ahead of print. PMID: 33949814

Alignment of Qx100/Qx200 Droplet Digital (Biorad) and QuantStudio 3D (Thermofisher) Digital PCR for quantification of BCR-ABL1 in Ph+ chronic myeloid leukemia. Carmen Fava, Simona Bernardi, Enrico Marco Gottardi, Roberta Lorenzatti, Laura Galeotti, Francesco Ceccherini, Francesco Cordoni, Filomena Daraio, Emilia Giugliano, Aleksandar Jovanovski, Marta Varotto, Davide Barberio, Giovanna Rege-Cambrin, Paola Berchialla, Veronica Sciannameo, Michele Malagola, Giuseppe Saglio, Domenico Russo. Diseases 2021, 9(2), 35; https://doi.org/10.3390/diseases9020035

To swab or not to swab? The lesson learned in Italy in the early stage of the Covid-19 pandemic. Paola Berchialla, Maria Teresa Giraudo, Carmen Fava, Andrea Ricotti, Giuseppe Saglio, Giulia Lorenzoni, Veronica Sciannameo, Sara Urru, Ilaria Prosepe, Corrado Lanera, Danila Azzolina, Dario Gregori. Appl. Sci. 2021, 11, 4042. https://doi.org/10.3390/app11094042

Clinical stability and propensity score matching in Cardiac Surgery: Is the clinical evaluation of treatment efficacy algorithm-dependent in small sample size settings? Bottigliengo, D.; Acar, A.S.; Sciannameo, V.; Lorenzoni, G.; Bejko, J.; Bottio, T.; Cozzi, E.; Vadori, M.; Soulillou, J.-P.; Roussel, J.C.; et al. Epidemiology Biostatistics and Public Health 2019, 16, doi:10.2427/13001.

The impact of the COVID-19 pandemic on nursing care: a cross-sectional survey-based study. Marco Clari, Michela Luciani, Alessio Conti, Veronica Sciannameo, Paola Berchialla, Paola Di Giulio, Sara Campagna, Valerio Dimonte. J. Pers. Med. 2021, 11(10), 945; https://doi.org/10.3390/jpm11100945.

Exploring the use and usefulness of educational resources among nurses during the first wave of the COVID-19 pandemic: a cross-sectional study. Alessio Conti, Marco Clari, PhD, RN; Michela Luciani, PhD RN; Veronica Sciannameo, MSC; Paola Berchialla, Phd, MSC; Paola Di Giulio, RN, MSC; Valerio Dimonte, Sara Campagna, PhD RN; Accepted by The Journal of Continuing Education in Nursing.

Under review

Long term results of rotating-hinge knee prosthetic Endo-Model: A Meta-analysis. Alessandro Bistolfi, Claudio Guidotti; Stefano Cremonese; Michele Boffano; Paola Berchialla, Veronica Sciannameo; Luigi Sabatini; Riccardo Ferracicni, Knee Surgery, Sports Traumatology, Arthroscopy,

Do sex-specific disparities in colorectal cancer chemotherapy can impact on drugs effect? De Francia, Silvia; Berchialla, Paola; Armando, Tiziana; Storto, Silvana; Allegra, Sarah; Sciannameo, Veronica; Soave, Giulia; Sprio, Andrea Elio; Racca, Silvia; Caiaffa, Maria Rosaria; Ciuffreda, Libero; Mussa, Maria Valentina. British Journal of Clinical Pharmacology.

Prediction of treatment outcome in clinical trials: a personalized medicine based approach. Paola Berchialla, Corrado Lanera, Veronica Sciannameo, Dario Gregori, Ileana Baldi. Scientific Reports.

Crosslinked versus conventional Ultra High Molecular Weight Polyethylene (UHMWPE) for total knee arthroplasty. Systematic review and meta-analysis of randomized clinical trials. A. Bistolfi, F. Bosco, F. Giustra, C.Faccenda, M. Viotto, L. Sabatini, P. Berchialla, V. Sciannameo, A. Massè. Knee Surgery, Sports Traumatology, Arthroscopy.

Primary Hyperoxaluria in Italy: the past 30 years and the next future of a (not so) rare disease. G. Mandrile, A. Pelle, V. Sciannameo, E. Benetti, M. M. D'Alessandro, F. Emma, M. Marangella, G. Montini, L. Peruzzi, R. Romagnoli, C. Vitale, B. Cellini, D. Giachino. Journal of Nephrology.

ABSTRACT

The “Centers for Disease Control” defines a chronic disease as a health condition which lasts at least one year, it limits daily activities and it requires continuous medical attention. Furthermore, when at the same time at least two chronic health diseases are co-occurring in the same individual, we refer to multimorbidity, which is a growing global public health issue, worsened by the aging of the population. Chronic diseases lead to poorer health outcomes which could heavily affect health care systems and their related costs in the future.

Diabetes is one of the most diffused chronic disease, and we refer to Type 2 Diabetes (T2D) when the body is not able to correctly use insulin. Glucose Lowering Medications (GLMs) are used in T2D patients to control blood glucose, Body Mass Index (BMI), blood pressure and lipids, to improve cardiovascular outcomes. In the last decades, lots of Randomized Controlled Trials (RCTs) have been conducted to evaluate the treatment effect of such medications, in comparison with placebo or between them.

However, these results which were obtained in RCT settings have to be confirmed by real world data (RWD), which are the ensemble of data related to the patient health status, routinely collected from different sources (i.e. disease registries, administrative databases where claims, billing activities, diagnoses of hospitalizations are collected). In fact, RCTs are a powerful tool to have scientific evidence about safety and efficacy of drugs, and they could help to understand the biological mechanisms undergoing the therapeutic actions stating if a medical product can ideally work, with a very high internal validity. Moreover, RCTs often have a low external validity and they are not sufficient to guide the decision-making process. It is therefore necessary to integrate knowledge from RCTs with Real World Evidence (RWE) which comes from RWD.

However, when dealing with RWD, lots of problems arise, such those relating to the absence of randomization in the treatment assignment, confounding, model specification, missing data, big data availability. In the last decades, lots of Machine Learning (ML) approaches have been developed to address these issues.

This thesis is focused on the application of advanced statistical approaches to analyze health outcomes and comorbidity patterns in patients with chronic diseases from RWD.

More in details, in the first contribution Propensity Score (PS) methods have been applied to evaluate the treatment effect of different GLMs, in terms of simultaneous reduction in HbA1c, body weight, and systolic blood pressure in T2D patients. Data were extracted from Dapagliflozin Real World evIdeNce in Type 2 Diabetes (DARWIN-T2D), a retrospective multicenter study conducted at diabetes specialist

outpatient clinics in Italy. In this study, we observed that in routine ambulatory care, initiation of Dapagliflozin (a SGLT2i drug) can be as effective as initiation of a GLP-1 receptor agonists (GLP-1RA) for the attainment of combined risk factor goals.

However, in this first work, we had to deal with lots of issues related to RWD: the absence of randomization in the treatment assignment, the high amount of missing data (about 50%, both on covariates and outcome measures), the misspecification of treatment and outcome models.

It follows that in the second contribution I tried to limit the size of biases occurring in observational context related to such issues applying different advanced statistical approaches, mainly focusing on the particular case in which a high percentage of missing not at random (MNAR) data are present in a dichotomous outcome. Covariate adjustment, PS adjustment, PS matching, inverse probability of treatment weighting, targeted maximum likelihood estimator (TMLE), were compared using DARWIN-T2D data and also in a simulation setting, done through Bayesian Networks (BNs) to resemble DARWIN-T2D characteristics. TMLE showed less biases and higher precision in estimating the Marginal Treatment Effect in an observational setting, in which the outcome and/or treatment models could be misspecified, regardless the amount of MNAR outcome data.

Then, in the third contribution, the aim was to evaluate generalizability of cardiovascular outcome trials (CVOTs) on GLP-1RA to the real-world population of T2D patients. The proportion of real-world patients which constitute CVOT-like populations were assessed, using as target population DARWIN-T2D. We developed a novel approach, based on BNs which allow to assess conditional dependencies among variables in DARWIN-T2D. Such method was then used to sample the greatest subsets of real-world patients yielding true CVOT-like populations. A very small proportion of real-world patients constitute true CVOT-like populations. These findings question whether any meaningful information can be drawn from applying trial Inclusion/Exclusion criteria to real-world T2D patients. Clinical practice transferability of CVOT should rather rely on observational effectiveness studies.

In the fourth contribution, the aim was transferring results obtained in CVOTs to the real-world setting, using again DARWIN-T2D as target population. A post-stratification approach based on aggregated data of CVOTs and individual data of a target population of diabetic outpatients was used. Stratum-specific estimates available from CVOTs were extracted from publications to calculate expected effect size for DARWIN-T2D by weighting the average of the stratum-specific treatment effects according to proportions of a given characteristic in the target population. The main finding was that, based on CVOT stratum-specific effects, cardiovascular protective actions of investigational GLMs are transferrable to a much different real-world population of patients with T2D.

Finally, in the fifth contribution, I worked on administrative databases of Piedmont, a Northern Italy region, to forecast urgent hospitalization in people aged more than 65 years. I applied the Bidirectional Encoder Representations from Transformers (BERT), which is a deep learning approach developed by Google in 2018. The aim was to deal with healthcare trajectories, defined as a sequence of medication purchases and hospitalization diagnoses, to forecast urgent hospitalizations within 3 months. Results suggested that BERT is able to embed administrative health records, into patients' medical histories to predict future urgent hospitalizations. This could be a tool which could help to improve the quality of life of elderly people, preventing adverse outcomes in a personalized way, and to optimize the allocation of healthcare resources in the future.

CHAPTER 1

INTRODUCTION

Chronic diseases and multimorbidity

The Centers for Diseases Control defines “chronic” all those diseases which last more than one year, which require continuous medical attention and limit daily activities. However, in the biomedical literature is still lacking a uniformed definition of chronic diseases (1). For example, the World Health Organization (WHO) defines chronic diseases all those illnesses that “are not passed from person to person. They are of long duration and generally slow progression. The four main types […] are cardiovascular diseases (like heart attacks and stroke), cancers, chronic respiratory diseases (such as chronic obstructed pulmonary disease and asthma) and diabetes” (2). The Australian Institute for Health and Welfare defines chronic diseases if there is a complex causality, a long development period without symptoms, a prolonged course of illness and associated functional disabilities (1) (3).

When at least two chronic diseases are simultaneously present in the same individual we refer to multimorbidity or comorbidity (4). Multimorbidity is a growing global public health issue, worsened by the increasing aging of the global population, which leads to adverse outcomes that contribute to heavily affect the health-care systems (4)(5). Furthermore, multimorbidity is strictly connected to polypharmacy (6), i.e. the use of multiple medications by the same individual at the same time, that increases the complexity of managing such patients (7).

Multimorbidity is then highly correlated with “frailty”, a decline of the health condition of a subject, which is strictly related to aging. Frailty leads to reduced physical and mental health faculties, resulting in higher vulnerability and increased risk of bad health outcomes, with negative implications for both older people themselves and for the entire society (8).

A better knowledge of the epidemiology of chronic diseases and multimorbidity is therefore necessary to develop intervention tailored to prevent them, to manage them in a personalized way and to better allocate healthcare resources (5).

An example of chronic disease: type 2 diabetes

T2D is one of the most diffused chronic disease, which affected 108 million of people in 1980 and increased to 422 million in 2014, with a higher prevalence in particular in the low- and middle-income countries (9). More in detail, in 2014 the 8.5% of adults (> 18 years) were diabetic people. In 2019, 1.5 million deaths were directly due to diabetes (9). These data highlights that diabetes is a huge global healthcare concern (10).

T2D occurs when pancreas produces not enough insulin or when the body is not able to correctly use the insulin produced (9). The main consequence of T2D is hyperglycemia, which leads to many problems related to blood vessels, nerves, heart, eyes, and kidneys (9). In fact, in many scientific studies it has been shown that adults affected by T2D have a two- or three-fold increase in the risk of heart attacks or strokes (11). Furthermore, if nerves damage in the feet is concomitant with reduced blood flow, there is an increase in the risk of having infections, foot ulcers and eventual need for amputations (9). Moreover, 2.6% of global blindness is caused by diabetes (12), and T2D is one of the main causes of kidney failure (13).

The main risk factors which lead to T2D, are excess of body weight, physical inactivity, obesity, hypertension and dyslipidemia (9) (10).

Glucose Lowering Medications (GLMs) are the main drugs used to control the level of glucose in blood, Body Mass Index (BMI), blood pressure and lipids to improve cardiovascular outcomes (10).

Sodium glucose co-transporter 2 inhibitors (SGLT2i) are GLMs which prevent renal glucose resorption. SGLT2i were available for the treatment of T2D in Italy (as Dapagliflozin) from March 2015 (10). In the last years, lots of RCTs have been conducted to evaluate the efficacy of Dapagliflozin. It was evaluated in comparison with placebo, or as add-on to metformin, sulphonylurea, DPP-4i, insulin or versus active comparators (14). Many meta-analyses showed non inferiority of Dapagliflozin 10 mg if compared with Glipizide and Saxagliptin. It has been shown that Dapagliflozin reduces Hba1c by 0.5-0.7% if compared with placebo at 24 weeks, and had sustained glucose lowering effects over periods of 48-102 weeks (15,16).

In conclusion, lots of studies showed that Dapagliflozin improves glycemic control in T2D patients, and that it is protective against cardiovascular risks by acting on well known risk factors.

Research on chronic diseases: from RCT to RWD

Results about the protective effect of Dapagliflozin against cardiovascular risk in T2D have been obtained mainly within RCTs, which are ideal settings, and patients enrolled in the trial are very selected ones.

It is therefore necessary to verify if the same conclusions are reached also analyzing RWD, extracted from routinely accumulated clinical data (10). Many studies have jet been conducted in this setting, substantially confirming the glycemic and extra-glycemic effects observed in RCTs and they support the protection against cardiovascular diseases (17–19).

Randomized Controlled Trials (RCTs)

Clinical trials are prospective studies in which subjects receive an experimental intervention. The design of RCTs is more complex at earlier phase of the study (20). In particular, phase I trials have the main aim of studying the pharmacokinetic, pharmacodynamics and safety. Typically, they are small, in fact often less than 100 healthy volunteers are enrolled, and there is a single arm.

Phase II trials may instead be composed by two or more arms, and subjects are randomized to the active treatment/intervention or to the control arm, which can be placebo or a comparative medication/ intervention. Randomization allows researchers to apply very basic statistical approaches, which lead to robust causal effect estimates, with simple outcome comparisons between the two arms (21). In this phase, safety and preliminary efficacy are assessed, and from 100 up to 300 patients are generally enrolled.

Then, often one of the aims in phase II trials is to determine the optimal medication dose, which will be used in phase III trials, that are generally used to establish the efficacy of a drug. They are bigger studies, involving generally more than 1000 patients, they can be single, double or triple blinded, and follow-up is typically longer than trials in the previous phases.

Finally, in phase IV trials the long term risks and rare adverse events are monitored and the optimal use of the drug in clinical practice is investigated (22).

RCTs are considered the most reliable method of generating evidence on the efficacy of interventions or medications (23). In fact, for many years RCTs and meta-analysis of RCTs were at the top of the hierarchy of study designs for the evidence of a treatment effect (24).

However, in 2016 Murad et colleagues published the "New evidence pyramid", in which they suggested another way of looking at the evidence-based medicine pyramid (25). They pointed out that the study design alone is not sufficient to state the risk of bias. In fact, methodological limitations of a study, imprecisions, and inconsistencies could be factors which influence the quality of a study (25). So, the authors replaced the straight lines that in the classic pyramid separate the different study designs with wavy lines (25). Then, they removed meta-analysis from the top of the pyramid and instead they use them as a lens through which other types of studies should be seen (25).

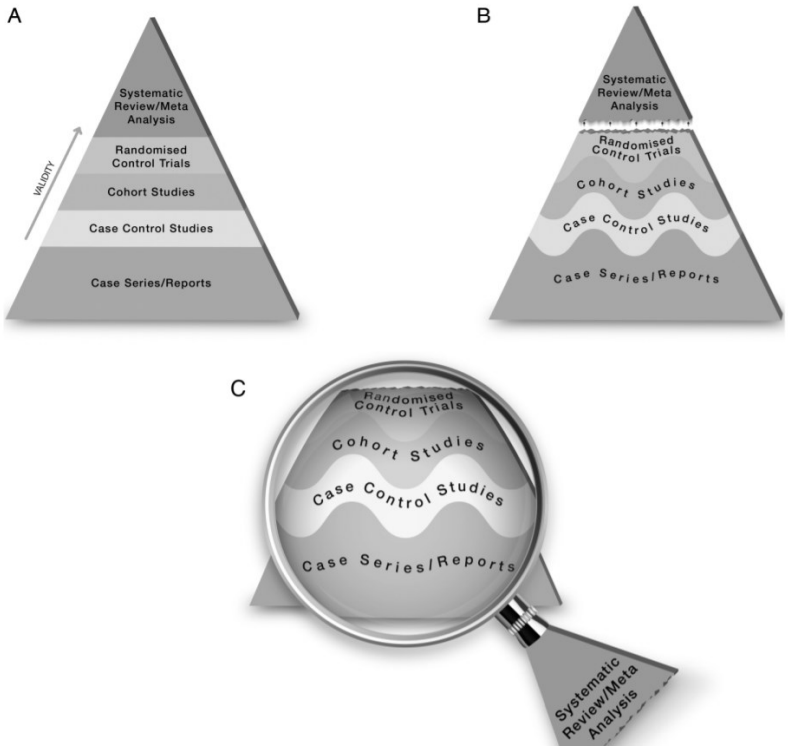


Figure 1. From «New evidence pyramid», M. H. Murad et al., Evid. Based Med 2016 (25).

The strongest point of RCTs is their excellent internal validity, due to the power of randomization that potentially remove confounding, ensuring that if a difference between groups is observed, it is highly probable that it is due to intervention itself and not to other confounding factors (23).

However, RCTs tend to be conducted in very selected populations and in particular environments which are different from the real-world clinical or home settings (26) (23). In fact, strict inclusion/exclusion (I/E) criteria are applied, that lead larger effect with respect to the expected in the general population. Strata of population which are vulnerable to side effects, i.e. children or older people affected for example by multimorbidity, are often excluded from trials (21). For example, in the oncological field, only <10%

of patients with cancer are enrolled into RCTs, and elderly subjects, with more than two diseases and with a lower socio-economic status are strongly under-represented (23).

RCTs can provide important information about efficacy, which is the drug effect under ideal settings, but not about effectiveness, i.e. the true benefit in routine clinical practice (23). Furthermore, RCTs have a limited capability in detecting toxicity or side effects, due to the relatively small sample size and the usually short follow up period (23).

However, in such specific settings it is easier to have a higher control of the quality of data, using detailed case-report forms that exist separately from ordinary medical records and which are designed specifically for research purposes. Moreover, when RCTs are conducted, an intensive monitoring controls the strict adherence of the research to a well-characterized protocol that ensure precision in data recruitment (26).

Real World Data and Real World Evidence

Evidence which comes from RCTs has to be integrated with evidences assessed in the real world context. RWD can be used to understand how different treatments or exposures affect outcomes of interest in real settings (21). Analyses of RWD leads to RWE, which is defined as the clinical evidence regarding the benefits and risks of the medication/intervention investigated (26). RWE provides information about safety surveillance and effectiveness of drugs and interventions, and can investigate which factors can influence treatment effects (26).

The availability of RWD, which can be used for healthcare research purposes, is constantly growing, due by the rise of electronic health records, disease registries, administrative databases which collect billing activities, diagnoses of hospitalizations, and data obtained from electronic devices (26). From these data, it is possible to reconstruct healthcare trajectories, which are defined as the ensemble of cares received and outcomes experienced by a subject. This leads to evidence which could be used to learn how to effectively treat patients in the future (27). However, RWD coming from claims databases, electronic devices, and disease registries were not designed for research aims, but for administrative purposes (26). For this reason, some critical issues are found when they are used to conduct medical research. For example, population databases often do not report the plan of specific treatments, but it could be reconstructed indirectly linking different data sources, with a questionable quality of data (23). Furthermore, confounding is a big issue in the observational context, due to the absence of randomization in the assignment of the treatment.

To deal with such intrinsic limitations of RWD, it is necessary using more advanced statistical approaches (10). In fact, the considerable sizes of data sets, the uncertain quality of data with high amount of missing data, the absence of randomization, and the presence of measured and/or unmeasured confounders could lead to incorrect conclusions (26). Statistical approaches typically used to deal with confounding are multivariable analysis and propensity score (PS) approaches. However, such approaches are regression-based models, which require lots of assumptions to be satisfied (28).

However, RWD leads to research with high external validity and provide insights into delivery of care in routine clinical practice to all patients, even those elderly and with comorbidity, that better represent population that effectively use medications in the real world (23).

Furthermore, RWD require less time and costs, allow to conduct safe research on high-risk groups and on effectiveness, and they provide higher sample size which allows to detect rare side effects or toxicity (24,31).

The complementarity of RCTs and RWD

RCTs are a powerful tool to have scientific evidence about safety and efficacy of medications, and they help to understand the biological mechanisms undergoing the therapeutic actions stating if a medical product can work, with a very high internal validity (26).

However, in the last decades, the idea that RCTs have a low external validity and that they are not sufficient to guide the decision processes is gaining traction, because the contexts in which they are conducted could be very different from the real-life clinical settings. In fact, many factors such as the heterogeneity the characteristics of patients, the simultaneous use of different drugs over time (i.e. polypharmacy), and the adherence to the treatments, are factors that could lead to discrepancies between the evidence generated in RCTs and their generalizability in the real world clinical practice.

Furthermore, there are several reasons for which in some particular situation, a RCT cannot be conducted, and it is necessary to replace it with an observational study. More in detail, Black and colleagues identified four main reasons that do not allow to perform an RCT. In fact, experimentation could be (24) (29):

1) unnecessary, if the treatment effect is so dramatic that unknown confounding elements could be ignored
2) inappropriate, if the analyzed outcome is rare, because long follow up are necessary

3) impossible, due to clinician's refusal to participate or to legal or ethical obstacles
4) inadequate if the external validity is low.

So, for many different reasons, RWE has a big potential in complementing the knowledge deriving from traditional RCTs, whose limitations make it difficult to generalize finding to the population of patients that uses medical products in practice (23).

Summarizing, RCTs are essential tools with a strong internal validity but a weak generalizability to real life context. For these reasons, there is a growing interest in RW studies, due to their close association with the routine clinical practice. However, RW studies are weak in terms of internal validity, and strong statistical tools are necessary to overcome their numerous intrinsic limitations, such as missing data and confounding.

In conclusion, RCTs and RW studies are complementary and only if they are used together we can obtain a better evidence, jointly in terms of internal and external validity of evidence. So, first conduct a RCT to demonstrate efficacy of a medication. Then, real world studies should evaluate patterns of care, toxicity and effectiveness in routine clinical practice (23) (30).

Table 1: Advantages and disadvantages of RCTs and RWE

	RCTs	**RW studies**
Advantages	• Efficacy • Randomization • Blinding • Control arm • Rigorous analysis methods • Simple statistical analyses • High quality of data	• Effectiveness • Non-selected population • Ethical feasibility • Rare or late side effects • Clinical routine practice setting
Disadvantages	• Selected population • Setting and monitoring bias • Ethical restrictions • Not able to detect side effects • Short duration	• Lack of randomization • Confounding factors • Lack of blindness • Complex statistical analyses • Low quality of data (missing data, inaccuracies,…)

Two examples of Real World databases

DARWIN-T2D

Dapagliflozin Real World evIdeNce in Type 2 Diabetes (DARWIN-T2D) is a real world multicenter retrospective nationwide Italian study, promoted by Italian Diabetes Society and AstraZeneca (ESR-14-11441) (10).

DARWIN-T2D uses clinical data which are routinely accumulated (31), with the aim to describe which are the baseline clinical characteristics of T2D patients, and to control glycemic and extra-glycemic parameters in patients undergoing Dapagliflozin compared with patients initiated on comparator GLMs (DPP-4i, i.e. Sitagliptin, Saxagliptin, Vildagliptin, Alogliptin) in Italian diabetes outpatient clinics (10).

The analysis is retrospective and starts on 13th March 2015, when Dapagliflozin was approved in Italy, and ends on 31st December 2016.

Secondary aims of DARWIN-T2D are to describe heterogeneity, regional variations, and temporal trends of baseline characteristics of T2D patients (10).

The study includes four groups of patients, in accordance with their main therapy (i.e. Dapagliflozin, DPP-4i, Gliclazide, GLP-1RA). As typically occurs in observational studies, the comparison of such groups is made more difficult by the absence of randomization (10).

T2D patients can be included in DARWIN-T2D if the following inclusion criteria are met (10):

- Age: 18-80 years;
- diagnosis of T2D since at least 1 year;
- had initiated Dapagliflozin 10 mg as add-on to metformin and/or insulin from 13th March 2015 to 31st December 2016 OR
- patients taking full-dose DPP4i OR
- patients taking Liraglutide 1.2 mg or 1.8 mg OR
- patients taking Exenatide QW 2 mg OR
- patients taking Gliclazide modified release 30 mg or higher;

On the other side, the exclusion criteria were the following (10):

- Diagnosis of type 1 diabetes
- Age < 18 years or age >80 years
- Previous therapy with SGLT2i
- Previous Chronic Kidney Disease (CKD) defined as an estimated glomerular filtration rate of less than 60 ml/min/1.73 mq.

To uniform the data extraction in all the centers involved in DARWIN-T2D, an automated software draws out data from the same electronic chart system (MyStar Connect [MSC], Me.te.da) (10). The specialist diabetes outpatient clinics enrolled in DARWIN-T2D were 46, uniformly distributed in the Italian territory, to better represent the diabetic population in Italy (10).
The patients enrolled in DARWIN-T2D are 281 217, which represent about one fifth of all T2D patients attending specialist clinics in Italy, but only 17 285 (6.1%) were included in longitudinal assessments (10).

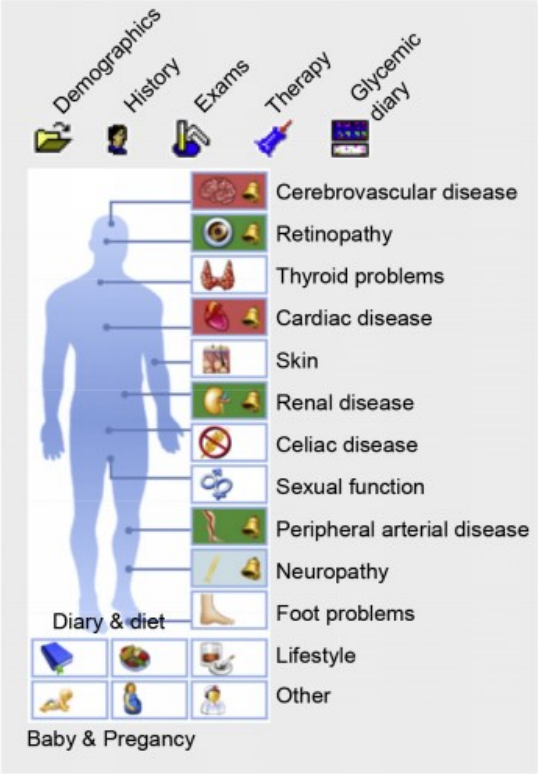


Figure 2: A selection of data contained in the MyStar software, which are present in the DARWIN-T2D study. Figure extracted from (10).

Healthcare Administrative databases (HADs)

The Healthcare Administrative Databases (HADs) are a source of RWD, extracted from databases of the National Health System (NHS).

Italian HADs are composed by (27):

- Pharmaceutical data, containing the Anatomical Therapeutic Chemical (ATC) classification system codes of medications purchased in territorial pharmacy or prescribed by general physicians, and data about pharmaceutical assistance services in direct distribution.
- Hospital discharge forms, reporting the International Classification of Diseases and Related Health Problems (ICD-9-CM) codes of the causes of hospitalizations carried out in public hospitals, equivalent and private affiliated hospitals.
- Access to the emergency room, with the causes' ICD-CM-9 codes.
- Outpatient services, that shows the codes of specialist services, i.e. visits and laboratory tests, and instrumental diagnostics which are provided by the public clinics and private hospital and extra-hospital specialists;
- Data about childbirth assistance, vaccinations and any other form of assistance guaranteed by the Essential Assistance Levels.

These data are used retrospectively to conduct medical research, i.e. when the study is planned, both expositions and outcomes have been verified in the past. However, HADs are a secondary source of data, which means that they were recruited for administrative purpose and not with the aim of conducting epidemiological studies. In fact, often information about severity of disease, concomitant diseases, dose of prescribed medications, BMI, and blood pressure, are not reported (27). Furthermore, many diagnostic procedures are under-reported due to the lack of financial incentive to document them (32). For example, Campbell et al. (33) observed that infections, venous thromboembolism, neurologic deficits, and the need for re-operation were significantly under-reported in patients undergoing spine surgery by administrative databases if compared with a prospective data collection. Moreover, in Italy, only drugs that are reimbursable by the NHS are registered, leaving a lack of information for entire classes of drugs.

HADs are interconnectable sources, that is data from different sources are linked to the same subject by means of a unique code that allows to reconstruct the health trajectory experienced by an individual in the NHS (27).

CHAPTER 2

SIMILAR EFFECTIVENESS OF DAPAGLIFLOZIN AND GLP-1 RECEPTOR AGONISTS CONCERNING COMBINED ENDPOINTS IN ROUTINE CLINICAL PRACTICE: A MULTICENTRE RETROSPECTIVE STUDY

Introduction

Glucose Lowering Medications (GLMs) are routinely used in the clinical practice for the management of T2D patients. However, the choice between the different GLMs have to keep into account the presence of atherosclerotic cardiovascular disease (CVD) or chronic kidney disease (CKD) (34). In fact, lots of cardiovascular outcome trials (CVOTs) showed improved cardio-renal outcomes if sodium-glucose cotransporter-2 inhibitors (SGLT2i) (35,36) or glucagon-like peptide-1 receptor agonists (GLP-1RA) (37–39) were used, which are both GLMs. Such CVOTs showed a cardio protective effect for both of them, but it seems that GLP-1RAs are more effective in lowering glucose if compared with SGLT2is. However, a recent network meta-analysis which compares GLMs, suggested that there is not a significant difference in the glycemic control between these two medications (40).

The cardio-vascular protection of GLMs is mainly due to their potential in lowering Hba1c, BW and systolic blood pressure (SBP)(33). In fact, in recent years it has been shown that managing simultaneously multiple risk factors (like hba1c, BW and SBP) could improve micro and macro vascular outcomes (41).

The focus in this work is to compare in a RW context the patients that are undergoing to Dapagliflozin (SGLT2i) and patients that were initiated to GLP-1RA, to assess differences in the changes in glycemic efficacy parameters.

However, as highlighted in the previous chapter, RWD have lots of issues to address, and appropriate statistical approaches are needed. In particular, the absence of randomization makes the comparison between the two treatment (GLP-1RA vs SGLT2i) challenging, due to confounding factors.

To date, the most diffused way to deal with the absence of randomization in RWD, is the potential outcome framework (42), with propensity score (PS)-based techniques (43–45). Such methods are able to simulate the randomization process that typically occurs in RCTs, being able to balance, on average,

the individual baseline characteristics (46). The PS-based approaches more applied in biomedical research are PS matching (PSM) and Inverse Probability of Treatment Weighting (IPTW) (47), which reduce the effect of confounding factors. More details about PS-based methods could be found in the "Material and Methods" section below.

Material and Methods

Real Word Data: DARWIN-T2D

In this study we used data from DARWIN-T2D, which was described in the previous chapter.
We compared T2D patients initiated on SGLT2i (Dapagliflozin, at the full dose of 10 mg,) or a GLP-1RA, which have not been treated with a member of the same drug class in the past and who continued to use the drug at the time of follow-up. Dapagliflozin was chosen among the SGLT2is because it was the most widely used in Italy when the study was designed, meanwhile among GLP-1RAs were included exenatide and liraglutide.
The primary endpoint was the proportion of patients with a simultaneous reduction in HbA1c, BW and SBP, without thresholds.
Secondary endpoints were:

(a) the proportion of patients that had a simultaneously reduction of Hba1c > 0.5%, BW > 2 kg, and SBP > 2 mm Hg

(b) the proportion of patients that achieved specific values at follow up: Hba1c ≤ 0.7%, BW loss > 3kg and SBP<140 mm Hg

(c) change in the individual components of the composite endpoints.

Data about age, sex, BMI, diabetes duration, systolic and diastolic blood pressure (SBP and DBP, respectively), smoking status, fasting glucose, HbA1c, complete lipid profile, serum creatinine, estimated glomerular filtration rate (eGFR, using the CKD-EPI equation), urinary albumin excretion rate (in mg/g of creatinine or equivalent), prior and concomitant GLM and other concomitant medications were collected. Then, micro-angiopathy was defined as the presence of at least one between retinopathy, neuropathy (somatic or autonomic), nephropathy (CKD stage III or higher or micro-/macro-albuminuria). Finally, Macro-angiopathy was defined as the presence of at least one

between ischemic heart disease, stroke/transient ischemic attack, peripheral arterial disease, revascularization of coronary, carotid or peripheral arteries.
At the end of follow, i.e. at the date of the first date between 3 and 12 months after baseline up data about Hba1c, BW and BP were collected.
More details about DARWIN-T2D were reported in Chapter 1.

Statistical analysis

Continuous variables were described as means and standard deviations (SDs) or as medians and inter-quantile ranges (IQRs), when the distributions were not Gaussian (evaluated through Kolomogorov-Smirnov test). Categorical variables were described as frequencies and percentages.
Comparisons between the two treatment groups (Dapagliflozin vs GLP-1RAs) were performed via Student's t tests or via Chi-squared tests. Comparisons between baseline and end of follow-up measurements in continuous variables were performed using the paired two-tailed Student's t test.
An high amount of missing data are present in DARWIN-T2D (about 50%), so multiple imputation (MI) was performed, using the Multiple Imputation by Chained Equation (MICE) algorithm (48–50). In this way, five imputed datasets were obtained. Only covariates with less than 40% of missing values were included as predictors in the imputation process, including observed outcome values. Outcome variables were not imputed, which means that only patients with observed outcome data were retained for the analyses.
I applied multivariable adjustment (MVA) and PS-based methods.
In the first one, logistic regression models were used when the outcome was dichotomous, and a linear regression model was used with continuous outcomes. The clinical characteristics that differed at baseline between the two groups were included as covariates. Kolmogorov–Smirnov tests were performed to evaluate if variables were Gaussians, and if $p<0.05$ they were log-transformed.
In each of the five imputed datasets, PS was computed including the following baseline covariates: age, gender, duration of diabetes, BW, BMI, FPG, HbA1c, SBP and DBP, total and HDL cholesterol, triglycerides, eGFR, insulin and metformin therapy, micro-angiopathy and macro-angiopathy.
PS Matching (PSM) and outcome analyses were then performed in each imputed subset. Finally, estimates of the treatment effect were pooled together to obtain the final treatment effect estimate. A sensitivity analysis was also performed with IPTW to estimate the Average Treatment Effect (ATE).

Statistical analyses were performed using R version 3.4.0 and a two-tailed p value less than 0.05 was considered statistically significant.

Propensity Score

Potential outcome framework is a way to quantify causal effects, introduced in 1974 by Donald Bruce Rubin (51). Each subject enrolled in the study has two potential outcomes. $Y_i(1)$ is the outcome that the i^{th} subject would have experienced if he/she had been treated (Z = 1), and $Y_i(0)$ is the outcome that the i^{th} subject would have had if he/she had not been treated (Z = 0). It follows that for each subject it is possible to observe only one of these potential outcomes $Y_i(0)$ and $Y_i(1)$, because each subject can only be treated or untreated. In mathematical notation, the observed outcome for the i^{th} subject, which is indicated with Y_i, is (46):

$$Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0)$$

Furthermore, for each subject the treatment effect is computed as $Y_i(1) - Y_i(0)$, from which follows that the average treatment effect (ATE) is defined as $E[Y_i(1) - Y_i(0)]$ (46).

However, in observational studies the treated and the untreated subjects often differ in lots of baseline characteristics, due to the absence of randomization. It follows that $E[Y_i(1)|Z = 1] \neq E[Y_i(1)]$ (46). Thus, an unbiased estimate of the ATE cannot be computed by directly comparing outcomes between treated and untreated groups (46).

When the outcome is dichotomous, the ATE could be estimated through the marginal Odds Ratio (OR). A way of estimating the marginal OR in the potential outcome framework is the PS approach, introduced by Rosenbaum and Rubin (52).

More in detail, PS is defined for each subject as the probability to be assigned to the binary treatment Z, conditionally to the observed baseline covariates $\mathbf{W} = (\mathbf{W_1}, \ldots, \mathbf{W_n})$. Mathematically, PS is denoted as

$$e_i = P(Z_i = 1|\boldsymbol{W_i})$$

for each subject *i* in the sample.

PS is a balancing score, i.e. conditionally on PS, the measured baseline covariates share a similar distribution between treated and untreated subjects, resembling a RCT setting (52).

Rosenbaum and Rubin defined treatment assignment strongly ignorable if (52):

(a) $(Y(1), Y(0)) \perp\!\!\!\perp Z|\mathbf{W}$, i.e. the treatment Z assignment is independent from the potential outcomes conditionally on baseline covariates **W** or, alternatively, there are not unmeasured confounders;

(b) $0 < P(Z=1|\mathbf{W}) < 1$, i.e. each subject has a probability of receiving the treatment Z that lies between 0 and 1, but it is not exactly 0 or 1.

These two conditions demonstrate that if treatment assignment is strongly ignorable, conditioning on the PS allows to obtain unbiased estimates of ATE (46). However, in RWD often these assumptions are not testable, and a biased treatment effect estimate could be obtained via PS methods.

Typically, the procedure applied to estimate PS is logistic regression, in which treatment status Z is regressed on observed baseline characteristics **W**.

Generally, PS techniques are advantageous if compared with regression-based approaches for at least 5 reasons. First, all pre-treatment variables are summarized into a single score (the PS) which reduce the dimensionality. Second, PS-based approaches come from the potential outcome framework, which is a formal model for causal inference. So, causal questions can be well-defined and explicitly specified. Third, PS methods do not require to model the mean for the outcome, which can help avoid bias from mis-specification of that model. Fourth, PS-based techniques avoid extrapolating beyond the observed data unlike parametric regression modeling for outcomes which extrapolate whenever the treatment and control groups are disparate on pretreatment variable. Finally, PS adjustments can be implemented using only the pre-treatment covariates and treatment assignments of study participants without any use of the outcomes, which eliminates the potential for the choice of model specification for the pre-treatment variables to be influenced by its impact on the estimated treatment effect (45).

Then, one of the PS-based approach involves the matching on the basis of the computed PS values, to create paired sets of treated and untreated subjects that share a similar value of the PS (52). When the matched sample has been formed, the treatment effect is estimated comparing outcomes between treated and untreated patients, miming a RCT setting (46). This method is referred as PS matching (PSM).

Propensity Score Matching (PSM)

The aim of PSM is to create paired sets of treated and untreated subjects that have a similar value of the PS, which means that in average they share similar baseline characteristics, simulating a RCT setting (52). When the matched sample has been formed, outcomes between treated and untreated patients are compared (46).

In this study, PSM was performed with a 1:1 ratio without replacement. This means that once an untreated subject has been selected to match a treated subject, he/she cannot be selected anymore to be a match for other subjects in the treatment group (46).

Then, PSM was performed with the nearest neighbour approach with a calliper of 0.15 standard deviations of the distribution of the PSs on the logit scale (46). If there were more than one untreated subject that had PS values equally close to that of the treated subject, one of these untreated subjects was selected randomly. No threshold has been set as maximum acceptable difference between the PS values of two matched subjects (46). The *MatchIt* R package was used (53).

When a PS analysis is performed, a crucial point is to evaluate whether the PS model has been adequately specified. PS is a balancing score, i.e. in strata of subjects that share the same PS value, the distributions of measured baseline covariates **W** will be quite similar between treated and untreated groups. So, a way for assessing whether the PS model has been correctly specified, involves the examination of the overlapping distributions of measured baseline covariates **W** between treated and untreated subjects that share the same estimated PS.

Then, a comparison of the means of continuous covariates and the distribution of the categorical ones between treated and untreated subjects has to be performed in the matched sample. The standardized mean differences are usually used to compare the treatment groups (46). Usually, a standardized difference less than 0.1 is considered acceptable.

Inverse Probability of Treatment Weighting (IPTW)

The second most diffused PS-based approach is the inverse probability of treatment weighting (IPTW), a doubly robust (DR) estimator, which means that it remains consistent even if either a model for the PS or the outcome is correctly specified (44,54). in IPTW, weights based on the PS value are computed, to create a sample in which the distribution of observed baseline covariates **W** is independent of treatment assignment (46).

More in detail, when we are interested in estimating ATE, weights are computed as follows (46):

$$w_i = \frac{Z_i}{e_i} + \frac{1 - Z_i}{1 - e_i}$$

for each subject i in the sample, where Z_i is 1 if the ith subject is treated and Z_i is 0 if the ith subject is not treated, and e_i is the PS value for the ith subject. In other words, to each subject is assigned a weight

that is equal to the inverse of the probability of receiving the treatment that the subject actually received.

Assuming that Y_i is the outcome measured for the i[th] subject, the estimate of ATE is computed by

$$ATE = \frac{1}{n}\sum_{i=1}^{n}\frac{Z_i Y_i}{e_i} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-Z_i) Y_i}{1-e_i}$$

where n is the total number of subjects recruited in the sample.

Furthermore, the variance estimation must account for the weighted nature of the sample, that requires robust variance estimation (46).

Results

From the 281 217 T2D patients recruited in DARWIN-T2D, 17 285 initiated GLMs. Among them, the patients that were undergoing Dapagliflozin (a SGLT2i drug) were 2 484, and patients who initiated a GLP-1RA medication were 2 247.

Follow-up data, which were collected between 3-12 month after baseline, were available for 830 patients in the Dapagliflozin group and for 811 in the GLP-1RA group.

The composite outcome (simultaneous reduction of hba1c, BW and SBP) was available for 473 patients who initiated Dapagliflozin and for 336 patients undergoing GLP-1RA.

The flowchart of the study is represented in Figure 3, extracted from (55).

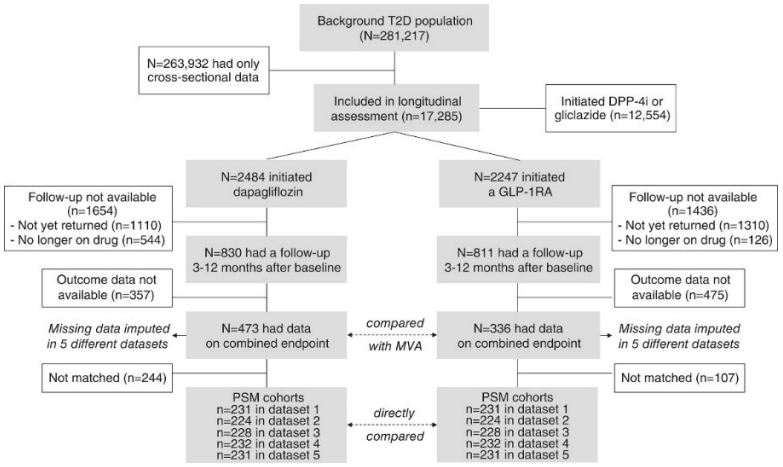


Figure 3: Study flowchart. MVA, multivariable adjustment. PSM, propensity score matching (55).

Baseline characteristics of the study participants are reported in Table 2. Some differences are underlined in the whole sample with complete data on the combined endpoint, before that PSM was performed, due

to the absence of randomization to the treatment groups. Such differences were in terms of age, diabetes duration, BMI, waist circumference, FPG, hba1c, EGFR, and associated therapy.

After PSM, 231 patients per group were retained (in the first imputed dataset, the other were similar and they are reported in the flowchart in Figure 3).

All the differences between the two treatment groups disappeared after that PSM was performed, and the baseline characteristics resulted well balanced, resembling a RCT setting.

Table 2: . Clinical characteristics of study subjects. Data are presented for the entire cohort before PSM and after PSM. For matched groups, data are shown for the first imputed dataset, whereas p-values and standardized difference (D) are shown for all imputed dataset pooled together. Only observed data are shown. BMI, body mass index. SBP, systolic blood pressure. BDP, diastolic blood pressure. FPG, fasting plasma glucose. HDL, high-density cholesterol. LDL, low-density cholesterol. eGFR, estimated glomerular filtration rate. UAER, urinary albumin excretion rate. ACEi, angiotensin converting enzyme inhibitors. ARBs, angiotensin receptor blockers.

	Before PSM				After PSM			
	Dapagliflozin	**GLP-1RA**	**p**	**D**	**Dapagliflozin**	**GLP-1RA**	**p**	**D**
Number	473	336	-	-	231	231	-	-
Age, years	59.6 ± 9.4	61.6 ± 9.2	0.003	0.212	60.5±9.1	60.4±9.2	0.899	0.030
Sex male, %	61.1%	54.5%	0.059	0.135	58.4	55.0	0.511	0.014
Diabetes duration, years	11.9 ± 8.1	9.8 ± 7.0	0.001	0.270	10.3±7.7	9.9±6.8	0.542	0.011
BMI, kg/m^2	33.4 ± 6.0	35.3 ± 5.5	<0.001	0.337	34.7±6.3	34.8±5.6	0.871	0.048
Waist circumference, cm	113.4 ± 13.2	117.6 ± 12.1	0.003	0.336	116.7±14.1	115.5±11.7	0.520	0.039
SBP, mm Hg	138.8 ± 18.2	140.6 ± 18.3	0.170	0.098	140.9±18.4	140.0±17.9	0.570	0.009
DBP, mm Hg	80.4 ± 10.4	80.5 ± 9.1	0.864	0.012	81.2±9.9	80.3±9.3	0.303	0.004
FPG, mg/dl	171.8 ± 51.3	152.3 ± 32.9	<0.001	0.453	158.9±47.4	153.3±34.3	0.171	0.020
HbA1c, %	8.6 ± 1.4	7.8 ± 0.8	<0.001	0.721	8.0±1.2	7.9±0.9	0.273	0.056
Total cholesterol, mg/dl	171.2 ± 36.4	171.3 ± 41.2	0.976	0.002	174.2±35.8	171.3±42.9	0.487	0.032
HDL cholesterol, mg/dl	45.8 ± 13.4	45.3 ± 11.8	0.622	0.041	46.8±13.4	45.6±12.3	0.371	0.016
Triglycerides, mg/dl	163.8 ± 99.9	164.6 ± 104.6	0.923	0.008	168.8±117.2	162.6±115.1	0.619	0.001
LDL cholesterol, mg/dl	93.3 ± 31.3	92.7 ± 35.3	0.838	0.017	94.5±31.5	93.4±37.4	0.770	0.032
eGFR, mg/min/1.73 m^2	89.7 ± 15.7	85.8 ± 17.5	0.006	0.232	86.0±16.1	88.7±17.0	0.136	0.009
UAER, mg/24h	105.0 ± 335.1	103.4 ± 273.0	0.955	0.005	83.5±241.7	103.1±526.3	0.700	0.023
Complications								
Microangiopathy, %	36.3	31.3	0.146	0.105	33.0	28.8	0.385	0.003
Macroangiopathy, %	31.9	32.6	0.853	0.014	34.0	31.6	0.677	0.019
Associated therapy								
Metformin, %	99.4	89.0	<0.001	0.454	98.3	96.5	0.384	0.015
Insulin, %	53.8	21.4	<0.001	0.709	30.9	29.4	0.815	0.015
GLM classes, median	2 (1-4)	2 (1-4)	1.000	0.000	2 (1-4)	2 (1-4)	1.000	0.000
Other therapies								
Anti-Platelet, %	45.7	42.3	0.368	0.068	44.8	42.0	0.634	0.080
Statin, %	64.5	62.0	0.488	0.052	56.1	61.8	0.277	0.017
ACE/ARBs, %	73.3	72.7	0.842	0.015	75.0	74.4	0.882	0.065
Beta blockers, %	31.9	32.0	0.978	0.002	33.0	30.0	0.568	0.021
Alpha blockers, %	7.1	9.0	0.363	0.070	7.1	5.9	0.596	0.049
Diuretics, %	10.7	13.0	0.346	0.071	11.3	12.6	0.667	0.012

Since outcome data were available only for a half of the cohort, differences between patients with observed and with missing outcomes were evaluated.

They significantly differed in terms of fasting glucose, total and LDL cholesterol, eGFR, and concomitant use of insulin and ACE inhibitors or angiotensin receptor blockers (Table 3). This leads thinking that the underlying missingness mechanism is not completely at random.

Table 3: Comparison between patients included in the composite outcome analysis and patients excluded from the analysis for missing outcome information. BMI, body mass index. SBP, systolic blood pressure. BDP, diastolic blood pressure. FPG, fasting plasma glucose. HDL, high-density cholesterol. LDL, low-density cholesterol. eGFR, estimated glomerular filtration rate. UAER, urinary albumin excretion rate. ACEi, angiotensin converting enzyme inhibitors. ARBs, angiotensin receptor blockers. CCB, calcium channel blockers.

	Excluded		**Included**		**Comparison**	
	% available	**Value**	**% available**	**Value**	**p**	**D**
Number		809		832		
Age, years	100.0	61.4±9.1	100.0	60.4±9.3	0.039	0.10
Sex male, %	100.0	57.0	100.0	58.3	0.584	0.03
Diabetes duration, years	100.0	11.4±7.8	100.0	11.0±7.7	0.259	0.06
BMI, kg/m^2	83.3	34.1±6.0	97.7	34.2±5.9	0.711	0.02
Waist circumference, cm	30.1	112.7±12.7	38.3	115.4±12.9	0.013	0.21
SBP, mm Hg	23.7	138.8±19.4	100.0	139.5±18.3	0.608	0.04
DBP, mm Hg	23.7	81.8±10.7	99.9	80.4±9.9	0.090	0.13
Fasting glucose, mg/dl	60.9	171.0±49.6	92.1	163.9±45.7	0.009	0.15
HbA1c, %	85.6	8.2±1.2	100.0	8.2±1.2	0.161	0.07
Total cholesterol, mg/dl	50.0	178.7±41.5	75.9	171.3±38.5	0.004	0.18
HDL cholesterol, mg/dl	48.1	45.1±12.7	74.5	45.6±12.7	0.514	0.04
Triglycerides, mg/dl	49.6	174.2±127.0	75.8	164.1±101.8	0.159	0.09
LDL cholesterol, mg/dl	46.3	99.3±33.9	72.9	93.0±33.0	0.004	0.19
eGFR, ml/min/1.73 m2	25.2	82.2±17.8	56.7	87.0±16.6	<0.001	0.27
UAER, mg/g	21.6	99.5±281.2	35.7	108.1±328.4	0.772	0.03
Associated therapy	99.9		99.9			
Insulin, %		91.2		95.2	0.001	0.16
Metformin, %		39.4		40.5	0.665	0.02
Other therapies	71.7		89.0			
Anti-platelet, %		44.3		44.3	0.991	0.00
Statin, %		61.4		63.5	0.451	0.04

ACEi/ARBs, %		66.5		73.1	0.010	0.14
CCB, %		22.7		25.0	0.337	0.05
Beta-blockers, %		30.0		31.9	0.441	0.04
Diuretics, %		10.3		11.7	0.421	0.04
Complications						
Microangiopathy, %	88.4	36.7	98.3	34.2	0.301	0.05
Macroangiopathy, %	80.3	29.5	88.4	32.2	0.279	0.06

The median follow-up was 5.9 months (IQR 4.0-6.5 months) in the Dapagliflozin group and it was 6.0 (IQR 4.4-6.6) months in the GLP-1RA group, without a statistically significant difference.

In Table 4 results about outcome analyses are reported, according to MVA and PSM approaches.

In the primary endpoint, no statistically significant results are obtained. In the multivariate analysis, the OR was 0.91 (95% CI, 0.64–1.30; P = 0.631) for Dapagliflozin vs GLP-1RA (Figure 4 A). The percentage of patients reaching a reduction in HbA1c greater than 0.5%, in BW greater than 2 kg and in SBP greater than 2 mm Hg in unadjusted and MVA analyses did not differ between groups (OR, 0.82; 95% CI,0.53–1.27; P = 0.397) (Figure 4 B).

Table 4: Percentages of patients achieving combined endpoints in the two groups. The 3 composite endpoints are shown and data are reported for the unadjusted analysis (percentages observed in the whole cohort), the multivariable adjustment (percentages estimated from regression models), and the propensity score matched analysis (percentages observed in matched groups). BW, body weight. SBP, systolic blood pressure. OR, odds ratio. Multivariable adjustment included the following variables: age, sex, diabetes duration, BMI, fasting plasma glucose, HbA1c, eGFR, concomitant use of metformin and insulin.

Combined endpoint	**Dapagliflozin**	**GLP-1RA**	**p**	**OR**
Any reduction in HbA1c, BW, and SBP, %				
Unadjusted	31.3	29.8	0.642	1.05 (0.85-1.30)
Multivariable adjustment	29.9	31.7	0.631	0.91 (0.64-1.30)
Propensity score matching (n=229/group)	30.3	30.2	0.760	0.93 (0.61-1.44)
ΔHbA1c>0.5%; ΔBW>2 kg; ΔSBP>2 mm Hg, %				
Unadjusted	16.9	17.3	0.897	0.98 (0.72-1.33)
Multivariable adjustment	16.0	18.6	0.397	0.82 (0.53-1.27)
Propensity score matching (n=229/group)	16.5	18.2	0.561	0.86 (0.53-1.41)
HbA1c≤7.0%; ΔBW≥3%; SBP <140 mm Hg, %				
Unadjusted	9.5	15.5	**0.010**	**0.61 (0.42-0.89)**
Multivariable adjustment	10.5	14.0	0.187	0.71 (0.44-1.15)
Propensity score matching (n=229/group)	12.6	17.7	0.183	0.70 (0.41-1.19)

A statistically significant result is achieved only evaluating a simultaneous reduction of hba1c ≤ 7.0 %; ΔBW ≥ 3 %; SBP < 140 MM HG %, in the unadjusted model, that has been reached by the 9.5% in the Dapagliflozin group and by the 15.5% in the GLP-1RA group (p = 0.01) with an OR = 0.61 (95% CI 0.42-0.89). However, this effect is not still significant in the multivariate analyses (Figure 4 C). In the MVA, HbA1c declined more significantly in the GLP-1RA group by 0.32 ± 0.07%; P < 0.001 (Figure 4D), whereas changes in BW and SBP are not different between groups (Figure 4 E, F).

PSM analyses lead to similar results, as reported in Table 4 and Figure 4. The proportion of patients that simultaneously decline in HbA1c, BW and SBP are not different between the two groups, with a OR= 0.93; 95% CI,0.61–1.44; P = 0.760 (Figure 4A). The OR for the composite end-point of reduction in HbA1c greater than 0.5%, in BW greater than 2kg and in SBP greater than 2 mm Hg was 0.86; 95% CI 0.53–1.41; P = 0.561 (Figure 4B). Furthermore, the proportion of patients that simultaneously reaching a HbA1c ≤ 7.0%, a BW loss of at least 3 Kg and a final SBP ≤ 140 mm Hg is lower in the Dapagliflozin group, but the result is not statistically significant (OR = 0.70; 95% CI 0.41-1.19; P =0.183) (Figure 4C). Finally, HbA1c declined more in the GLP-1RA group, by 0.29% (95% CI, −0.46; −0.12; P < 0.001) (Figure 4D), whereas changes in BW and SBP are not statistically different between Dapagliflozin and GLP-1RA groups (Figure 4E, F).

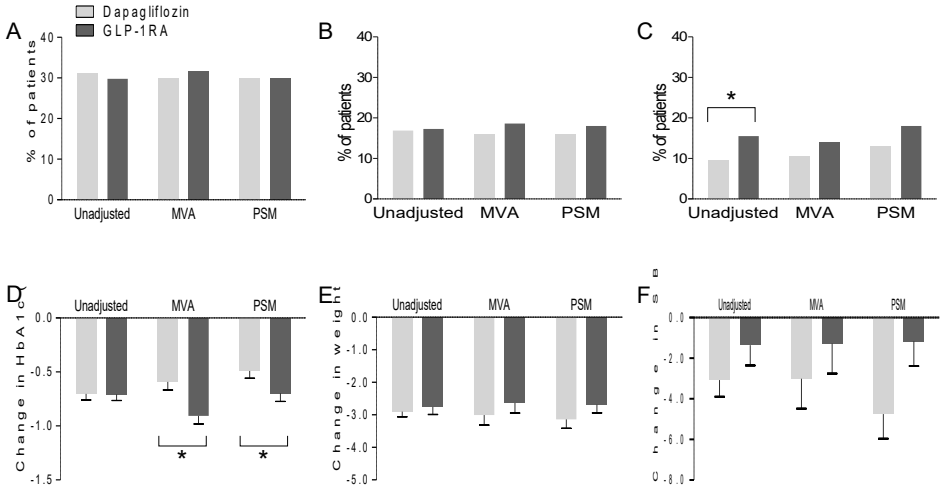


Figure 4: Extracted from (55). Comparative effectiveness on combined and individual endpoints. The proportion of patients in the unadjusted, multivariable adjusted (MVA), and propensity score matched (PSM) analyses attaining the primary combined endpoint of any reduction in HbA1c, body weight, and systolic blood pressure (A), the combined endpoint of a reduction of HbA1c >0.5%, body weight >2 kg, and systolic blood pressure >2 mm Hg (B) or the composite target of final HbA1c ≤7.0%, body weight loss ≥3%, and systolic blood pressure <140 mm Hg (C). Change from baseline to the end of follow-up in HbA1c (D), body weight (E), and systolic blood pressure (F) in the unadjusted, MVA, and PSM analyses. *p<0.05 for the indicated comparison. The histograms in panels D through F indicate mean and SEM.

We conducted also some sensitivity analyses, to test results stability.

First, a sensitivity analysis was performed including the prior GLM class number in the PS model, because line of therapy could influence the probability of receiving the treatment. However, we do not obtain any significant change in the results, if compared with the main outcome analyses (Table 5).

Table 5: Sensitivity analyses. The number of prior glucose lowering medication (GLM) classes was included in the propensity score (PS) model to perform PS matching (PSM). The pooled OR (with 95% C.I.) for each composite endpoint was obtained from the 5 imputed datasets and calculated for patients who received Dapagliflozin versus those who received GLP-1RA.

Combined endpoint	**Dapagliflozin**	**GLP-1RA**	**p**	**OR**
Unadjusted	N=473	N=336		
Any reduction in HbA1c, BW, and SBP, %	31.3	29.8	0.642	1.05 (0.85-1.30)
ΔHbA1c>0.5%; ΔBW>2 kg; ΔSBP>2 mm Hg, %	16.9	17.3	0.897	0.98 (0.72-1.33)
HbA1c≤7.0%; ΔBW≥3%; SBP <140 mm Hg, %	9.5	15.5	**0.010**	**0.61 (0.42-0.89)**
Multivariable adjustment	N=473	N=336		
Any reduction in HbA1c, BW, and SBP, %	29.9	31.7	0.631	0.91 (0.64-1.30)
ΔHbA1c>0.5%; ΔBW>2 kg; ΔSBP>2 mm Hg, %	16.0	18.6	0.397	0.82 (0.53-1.27)
HbA1c≤7.0%; ΔBW≥3%; SBP <140 mm Hg, %	10.5	14.0	0.187	0.71 (0.44-1.15)
Propensity score matching	N=229	N=229		
Any reduction in HbA1c, BW, and SBP, %	30.3	30.2	0.760	0.93 (0.61-1.44)
ΔHbA1c>0.5%; ΔBW>2 kg; ΔSBP>2 mm Hg, %	16.5	18.2	0.561	0.86 (0.53-1.41)
HbA1c≤7.0%; ΔBW≥3%; SBP <140 mm Hg, %	12.6	17.7	0.183	0.70 (0.41-1.19)

Secondly, in Table 6 are reported the results performing IPTW, with and without including the prior number of GLM classes in the PS model.

However, also in this case results did not change if compared with the main outcome analyses.

In Figure 5, balancing properties of PSM were reported. All the characteristics reached a Standardized Mean Difference (SMD) < 10%, so an optimal balancement between the Dapagliflozin and GLP-1RA groups was reached.

In the right part, the distribution of PS in the two groups is reported, showing a good overlap in baseline clinical characteristics between patients undergoing Dapagliflozin and patients under GLP-1RA.

Table 6: Sensitivity analyses. Inverse probability treatment weighting (IPTW) was used to estimate the average treatment effect with or without incorporation of the prior number of GLM classes in the PS. The pooled OR (with 95% C.I.) for each composite endpoint was obtained from the 5 imputed datasets and calculated for patients who received Dapagliflozin versus those who received GLP-1RA.

Outcome	PSM	IPTW
Any reduction in HbA1c, BW, and SBP		
Without prior GLM classes	0.93 (0.61-1.44)	0.93 (0.63-1.39)
Incorporating prior GLM classes	0.93 (0.60-1.44)	1.05 (0.73-1.52)
ΔHbA1c>0.5%; ΔBW>2 kg; ΔSBP>2 mm Hg		
Without prior GLM classes	0.86 (0.53-1.41)	0.85 (0.54-1.34)
Incorporating prior GLM classes	0.74 (0.40-1.35)	0.93 (0.61-1.43)
HbA1c≤7.0%; ΔBW≥3%; SBP <140 mm Hg		
Without prior GLM classes	0.70 (0.41-1.19)	0.67 (0.37-1.22)
Incorporating prior GLM classes	0.56 (0.29-1.07)	0.75 (0.37-1.55)

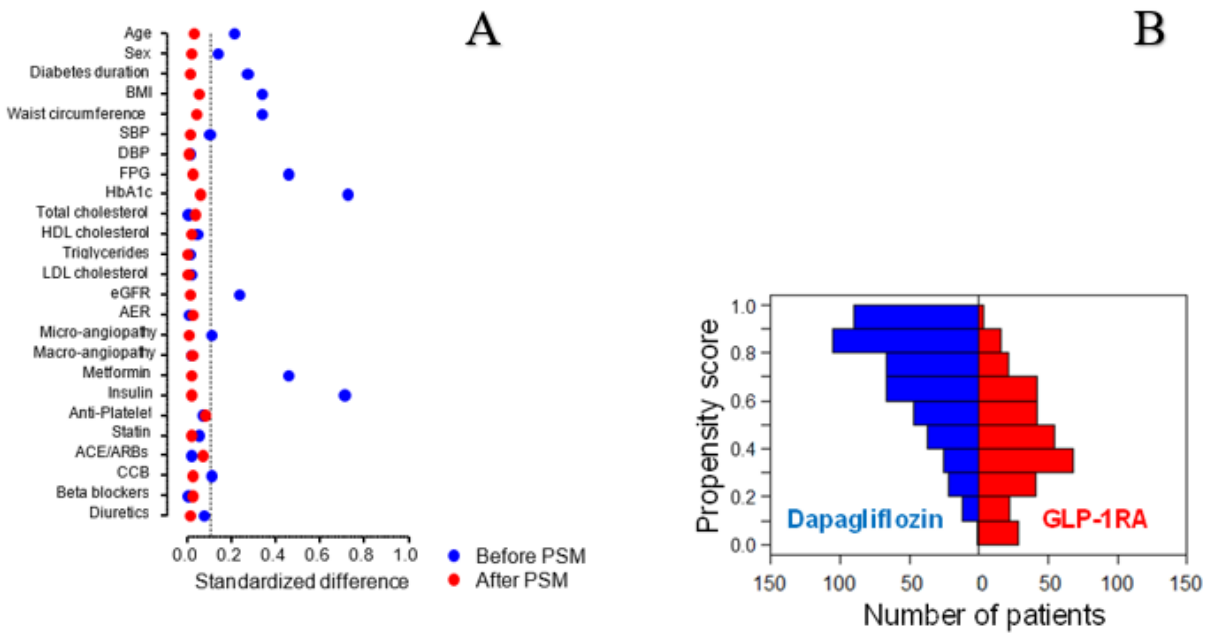


Figure 5: Extracted from (55). A. Rebalancing of patient characteristics after propensity score matching. The graph shows the standardized difference (STD) for each variable calculated in the dataset before (blue) and after (red) propensity score matching (PSM). A STD < 0.10 (dashed line) is indicative of a good match between groups. BMI, body mass index. SBP, systolic blood pressure. BDP, diastolic blood pressure. FPG, fasting plasma glucose. HDL, high-density cholesterol. LDL, low-density cholesterol. eGFR, estimated glomerular filtration rate. UAER, urinary albumin excretion rate. ACEi, angiotensin converting enzyme inhibitors. ARBs, angiotensin receptor blockers. CCB, calcium channel blockers. B. Common support between the two groups of patients. Common support refers to the overlap in clinical characteristics between the group of patients who received Dapagliflozin and the group of patients who received GLP-1RA. The graph represents the distribution of propensity scores in the two groups of treatment in the first imputed dataset.

Discussion

GLMs are of fundamental importance for secondary cardiovascular prevention of adverse events in T2D patients. However, different classes of GLMs exist, such as SGLT2i and GLP-1RAs. Both of them have favorable effects on the control of glycemic and extra-glycemic parameters, but they have not been analyzed in RCTs as combined endpoints, in terms of Hba1c, BW and SBP control.

In this study, we obtained a similar proportion of patients initiating Dapagliflozin, a SGLT2i medication, or a GLP-1RA who reached a simultaneous reduction in HbA1c, BW and SBP parameters.

More in detail, GLP-1RA resulted more effective than Dapagliflozin if we consider the reduction in HbA1c by almost 0.3% in both the MVA (95% CI, 0.2%–0.5%) and PSM (95% CI, 0.1%–0.5%) analyses (Figure 4D). This result is in line with the DURATION-8 trial (56).

Then, also when we considered specific thresholds in the composite outcome, we did not obtain significant differences between the two treatment groups. A possible explanation of this result, could be the fact that the effect of Dapagliflozin on BW and SBP is counterbalanced by the larger effects of GLP-1RA on HbA1c. Moreover, since the follow-up period of our study was relatively short (maximum 1 year), we cannot exclude that a difference in the treatment effect can be meet in a longer follow up.

Finally, we considered proportion of patients simultaneously reaching specific targets, obtaining a trend favorable to GLP-1RA, probably due to its greater glycemic control effect.

Even if in Figure 5 is showed a good overlap of the PS' support between the two treatment groups, important differences were highlighted in the baseline characteristics of the study participant.

To deal with these confounders, which is a typical problem that occurs when analyzing RWD, we applied different statistical approaches: MVA, PSM and IPTW. The first allows using data from all patients but it is based on a very strong assumption of linearity between covariates and outcomes. PSM simulates instead a quasi-experimental setting and makes no assumption about the relationships between variables that enter in the PS model and those in the outcome one. However, PSM restricts the analysis to matched patients, excluding a substantial part of data (about 40%), that could lead to biased results. Moreover, it is important to notice that the two different approaches give very similar results for all the endpoints considered.

Finally, IPTW analysis was performed as sensitivity analysis, to test the robustness of results. IPTW, differently from PSM, allows to retain all the patients included in the sample, reweighted for their probability to be assigned to the Dapagliflozin group given their baseline characteristics. Also in this case, results obtained through PSM and MVA were confirmed.

An important limitation of such statistical approaches is that they are not able to rule out residual confounding by unmeasured variables, such as diet and exercise habits, as well as patient preference, compliance and socio-economic status.

Furthermore, MVA and PS-based methods are highly sensitive to misspecification of both treatment and outcome models, which often occurs in observational context, so more robust methods are required. In fact, it has been shown that if the treatment and/or outcome model is misspecified, the OR estimate is biased in the direction of the conditional OR. This aspect it has been deepened in the following chapter.

Another heavy limitation of this study, is the presence of a very high percentage of missing data, another typical issue in RW studies. Missing data in covariates were handled with multiple imputation, thereby increasing the uncertainty of the estimates. For what concern outcome data, we decided not to impute them, led to exclusion of a big amount of patients from the analysis (about 50%), further limiting generalizability of the results. Also this aspect need further investigation, as pointed out in the next chapter.

In conclusion, this study shows that initiation on Dapagliflozin can be as effective as initiation on a GLP-1RA, in the simultaneous reduction of HbA1c, BW and SBP within routine specialist care. However, many issues related to RWD limitations, such as model misspecification and missing outcome data remained in this study open matters, that I have faced in the next chapter.

This chapter has been published as:
Similar effectiveness of dapagliflozin and GLP-1 receptor agonists concerning combined endpoints in routine clinical practice: A multicentre retrospective study. Fadini GP, Sciannameo V, Franzetti I, Bottigliengo D, D'Angelo P, Vinci C, Berchialla P, Arena S, Buzzetti R, Avogaro A; DARWIN-T2D network. Diabetes Obes Metab. 2019 Aug;21(8):1886-1894. doi: 10.1111/dom.13747. Epub 2019 May 8. PMID: 30985052

CHAPTER 3

TARGETED MAXIMUM LIKELIHOOD ESTIMATION OF TREATMENT EFFECTIVENESS UNDER OUTCOME DATA MISSINGNESS AND MODEL MISSPECIFICATION: A SIMULATION STUDY TO ASSESS RESULTS FROM THE DARWIN-T2D STUDY

Introduction

In RW studies the absence of randomization in the assignment of treatment, model misspecification, and missingness in both covariates and outcome data made the estimation of the marginal (i.e. at the population level) treatment effect very challenging.

When dealing with a dichotomous outcome, the most diffused approach is logistic regression (LR), which nevertheless requires lots of assumptions to be satisfied to be appropriately used. In particular, it is necessary to correctly specify the regression model, all the confounders must be measured, the observations should be independent from each other, it is not allowed multi-collinearity among the independent variables, and it is required linearity between the log odds and the independent variables. However, when dealing with RWD these assumptions often are not verified or they are not testable, leading to a highly biased estimate of the treatment effect of interest.

Furthermore, LR model approximates the estimate of marginal OR estimating the conditional OR (i.e. at the subject level). In this way, we are implicitly assuming homogeneity of the treatment effect in strata of subjects' observed covariates. In other words, it does not consider neither the potential treatment effect heterogeneity nor the presence of unmeasured confounders (57). In fact, in a observational study, conditional and marginal treatment effects coincide only if there was no unmeasured confounding, the true outcome model was known, and the outcome was continuous (46). Instead, if the outcome is dichotomous, in a observational setting, even without unmeasured confounding and even if the outcome model was correctly specified, the conditional and the marginal ORs do not coincide (46).

The marginal OR is often approximated with the conditional OR, for example estimated taking advantage from the potential outcome framework, through PS-based approaches introduced in the previous chapter. However, it has been shown that these methods often lead to biased estimates of the conditional ORs

(46,58), because PS-based methods are sensitive to the misspecification of both treatment and outcome models, they require lots of assumptions to be verified like the absence of unmeasured confounders, and the positivity assumption, and they suffer from missingness (59).

Lots of different methods have been developed to deal with missing data, both in the independent variables and in outcomes. The most diffused approach is the complete case (CC) method, i.e. only observations with no missing data were retrieved for the analysis (60). However, this approach leads to less precise and more biased estimates. An alternative to CC, is single or multiple imputation (MI), in which observed baseline variables are used to impute missing values, through regression-based approaches (61). Then, an alternative approach to both CC and MI is inverse probability weighting (IPW), which weights each subject with the inverse of the probability of being a CC. However, none of these approaches is statistically valid in general, and they can lead to serious bias in the treatment effect estimate(61).

In this study we focused mainly on missingness about dichotomous outcome data, but also on models misspecification, in a observational setting.

The risk of bias due to missingness depends on the reasons why data are missing, that are commonly classified as: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) (61). In Table 7, their main characteristics are summarized.

Table 7: Definitions and main features of missing completely at random (MCAR), missing at random (MAR), missing not at random (MNAR).

Missing completely at random (MCAR)
When subjects with missing data are a random sub-sample of the individuals participating in a study, the missing data are referred as MCAR (83) (82) (84). Only in this scenario, CC analysis yields unbiased estimates of the treatment effect (83) (82) (84). However, CC analysis is less efficient (i.e. imprecise) because not all the data are used (82).
Missing at random (MAR)
When missing data mechanism is not MCAR but it is related to some observed covariates, we referred to it as MAR data (83). In this case, performing a CC analysis may increase the bias in the estimate of the treatment effect (82). Rather observed data can be used as predictors of the missing outcome data, that can be imputed through regression-based approaches. The imprecision of the imputation process can be assessed performing multiple imputation, where multiple values are sampled from an estimated distribution and imputed (82). Hence, multiple data sets with different imputed outcomes are created. Then, each data set is analyzed separately, and results are pooled together by using standard techniques that take into account the variation between the imputed data sets.
Missing not at random (MNAR)
In this case, missingness depends by unobserved data or by the unobserved variable itself. With MNAR data, the aforementioned approaches to deal with missing data (i.e. CC, MI or IPW) are not suitable and there is no a universal method of handling this kind of data properly (82). With MNAR data, bias due to analyses based on MI may be bigger than the bias resulting from the CC analysis (83).

In general, the most used approaches to deal with missing data, such as CC and MI, yield to unbiased estimates only when MCAR or MAR is present. However, there are not rules which state how to correctly identify missingness mechanism of our data. When MNAR data are present, analyses result strongly biased if they are based on conventional statistical approaches, which instead make MCAR or MAR assumptions (62). Furthermore, the consequences of model misspecification, commonly diffused in RWD, are more relevant when MNAR data are present (63).

In recent years, double robust methods have been developed to handle with both model misspecification and missingness in outcome data (64). In particular, in this work we applied the Targeted Maximum Likelihood Estimator (TMLE) (65) and we compare it with other statistical approaches to estimate the marginal treatment effect under model misspecification and non-randomized treatment assignment, in particular when missing dichotomous outcome data are generated under MNAR or MAR mechanisms.

More details about double robust methods and TMLE are given in the next section.

Both real-world (DARWIN-T2D) and simulated data are analysed.

Material and Methods

In this study, we compared different approaches widely used to estimate marginal OR, both in RW and in a simulated setting. The methods applied were:

(i) covariate adjustment through logistic regression (LR),

(ii) PS adjustment,

(iii) PS matching (PSM),

(iv) Inverse Probability of Treatment Weighting (IPTW) via PS,

(v) TMLE.

The main focus of this study was about the mechanism generating missing data in the outcome, and we suppose it of type MNAR or MAR, that are the still less explored and rules on how to deal with them are still lacking. Furthermore, as showed in the previous chapter, we observed that in DARWIN-T2D patients with observed and missing outcome data differed in many covariates, suggesting a possible MNAR or MAR mechanism on outcome data.

LR and PS-based analyses were performed applying the CC approach. We do not perform MI because it has been shown that CC and MI lead to comparable results *(66,67)*.

More details about PS-based approaches are reported in the previous section, meanwhile TMLE is explained in the next paragraph.

Targeted Maximum Likelihood Estimator (TMLE)

In 2006 Mark J. van der Laan and Daniel Rubin introduced TMLE, an efficient, double robust (DR), semi-parametric, maximum likelihood estimator, which is based on G-computation (65). TMLE includes a secondary "targeting" step, which has the aim to optimize the bias-variance tradeoff for the parameter of interest (68).

G-computation algorithm was introduced in 1986 by Robins (69) to estimate causal effect in presence of time-dependent confounders affected by a time-varying exposure. In fact, in such scenario, traditional regression-based models typically fail. G-computation belongs to the generalized method (G-method) family (70), which includes the g-formula, marginal structural models, and structural nested models (71). Compared with standard regression-based methods, i.e. linear, logistic or Cox regressions, the G-methods provide consistent estimates of contrasts (i.e. ratios or differences) of average potential outcomes under less restrictive assumptions (72).

In fact, regression-based approaches rely on the strong assumption that the effect measure is constant across different levels of the confounders which are included as covariates in the model (73), excluding a priori a possible heterogeneity in the treatment effect. The g-formula allows instead to relax this hypothesis, thanks to a generalization of standardization.

G-computation relies on the estimation of the outcome mechanism, specified by E (Y|Z, **W**). Contrariwise, PS methods involve the estimation of the treatment mechanism, defined as P(Z=1|**W**). TMLE involves estimation of both the outcome and the treatment mechanisms, i.e. E (Y|Z, **W**) and P(Z=1|**W**) (Table 8) (68).

Table 8: G-computation, Propensity Score, TMLE

G-computation	E (Y|Z, **W**)	**TMLE**
Propensity Score	P(Z=1|**W**)	

TMLE is a DR semi-parametric method (45), which estimate treatment and outcome models taking advantages from machine learning (ML) approaches, which do not require strong assumptions on data distributions and allow to account for a large number of covariates, even with complex and non-linear relationships. Furthermore, it has been shown that the estimation of PS through ML approaches, in the case of binary treatment, outperforms simple logistic regression models with iterative variable selection (45).

DR estimation combines a model for the outcome with weighting to obtain an estimator that yields consistent estimates of the treatment effect if either the model for the outcome or the PS model is correct but not necessarily both (45). DR estimators do not require that the data-generating distributions are correctly identified, and are semi-parametric in this sense (74).

Between G-methods, we can perform standardization via g-formula and the IPW via the marginal structural modeling theory. However, if used individually, standardization requires that the outcome model is correctly specified, meanwhile IPW requires a correct specification of the exposure model. DR methods combine this two approaches into a single technique that has more relaxed assumptions. In fact, DR methods require a correct specification (in the case of parametric regression) of the outcome model or the exposure model, but not both.

TMLE can estimate many different statistical estimands of interest. In this work, our interest was about estimating ATE, i.e. the mean difference in outcomes between patients allocated to two different treatments, adjusting for confounders. ATE is defined as follows:

$$ATE = \Psi = E_W [E [Y |Z=1, \mathbf{W}] - E [Y|Z=0, \mathbf{W}]].$$

TMLE methodology is based on the counterfactual framework discussed in the previous chapter, which translates the problem of the estimation of the causal effect in a missing data problem. In fact, for each subject we can observe only one of the two potential outcomes.

Some assumptions are required (75):

- conditional exchangeability, i.e. there are no unmeasured confounders of the treatment effect on the outcome;
- positivity assumption, which means that if there are some strata of W in which no observations received the treatment Z=z, then we cannot compare the treatment effect at level z;
- consistency assumption, that means that the observed outcome is equal to the counterfactual outcome corresponding to the observed treatment.

More details about TMLE are reported in the following sections.

Step 1: Initial estimate of the outcome

The first step of TMLE involves the estimation of the expected value of the outcome using treatment and confounders as predictors, considering only observations with observed outcome, which we will indicate with $\Delta = 1$, with the following equation:

$$Q(Z, \boldsymbol{W}) = E\,(Y|Z, \boldsymbol{W}, \Delta = 1).$$

Formally, to estimate this conditional expectation it is possible to use any regression-based approach, however it is preferable to use Super Learner (SL) *(76)*, which combine flexible ML models, that allows to relax assumptions on the underlying data distributions. In fact, SL is an alternative statistical approach based on ML ensemble methods that finds the optimal combination of a collection of algorithms to minimize the cross validated risk. A mathematically proven theorem states that SL algorithm performs asymptotically as well as the oracle selector, i.e., the best candidate between learner algorithms inserted into SL *(77) (78)*. A representation of SL is given in Figure 6, extracted from the paper of M.J. van der Laan, in which it was introduced *(76)*.

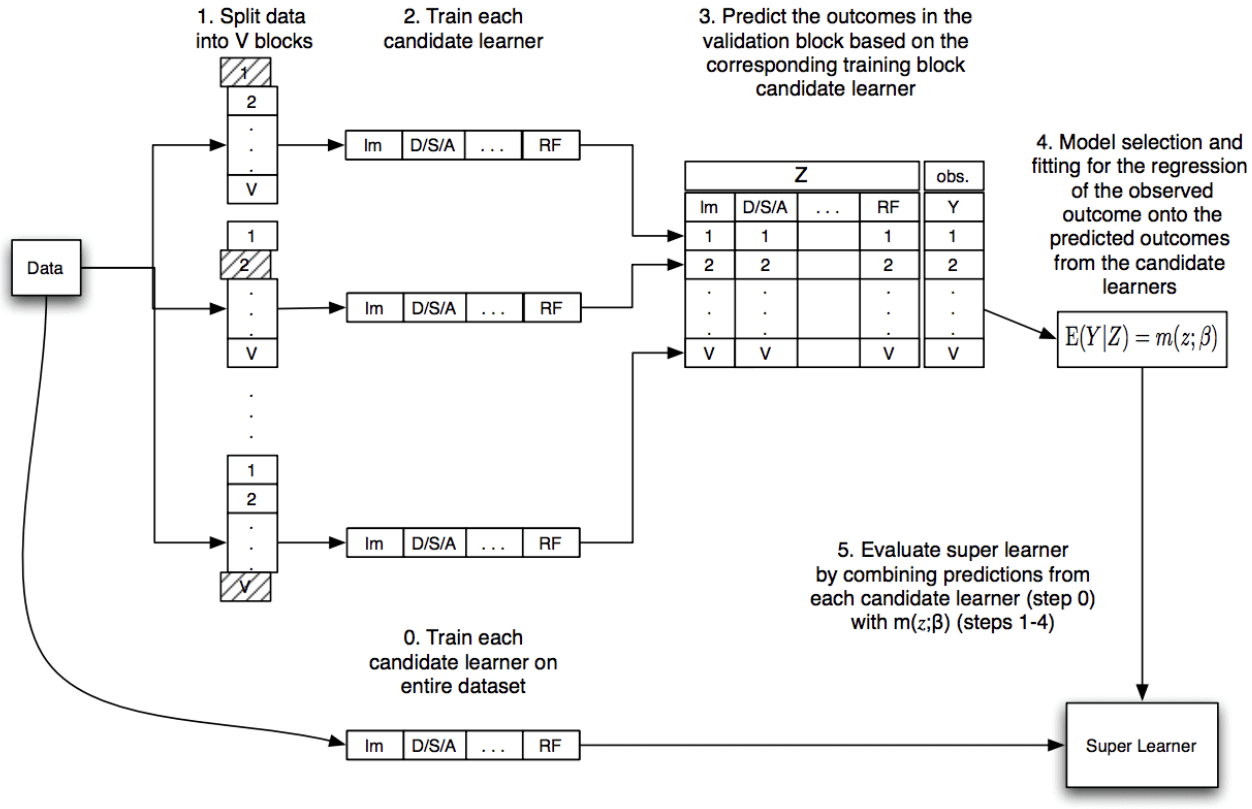


Figure 6: Flow diagram of SL, extracted from (67).

The first step of SL requires to specify a list of L base algorithms, called "base learners". Then, k-fold cross-validation is performed on each of these algorithms and a N x L matrix is obtained with the N cross-validated predicted values for each of the L learners. On this matrix, along with the original response matrix, a meta-learning algorithm is trained and it is used to generate predictions on the test set.

Then, for every observation we can estimate the outcome, under three different scenarios:

1. If every observation received the treatment they actually received

$$\hat{Q}(Z, \boldsymbol{W}) = \hat{E}[Y|Z, \boldsymbol{W}, \Delta = 1]$$

2. If every observation received the treatment, whether they actually did or not

$$\hat{Q}(1, \boldsymbol{W}) = \hat{E}[Y|Z = 1, \boldsymbol{W}, \Delta = 1]$$

3. If every observation received the control, whether they actually did or not

$$\hat{Q}(0, \boldsymbol{W}) = \hat{E}[Y|Z = 0, \boldsymbol{W}, \Delta = 1]$$

The average difference between $\hat{E}[Y|Z = 1, \boldsymbol{W}, \Delta = 1]$ and $\hat{E}[Y|Z = 0, \boldsymbol{W}, \Delta = 1]$ is a possible estimation of ATE, called standardization, g-formula estimation, or G-computation.

$$\widehat{ATE_{G-comp}} == \widehat{\Psi_{G-comp}} = \frac{1}{N} \sum_{i=1}^{N} (\hat{E}[Y|Z = 1, \boldsymbol{W}, \Delta = 1] - \hat{E}[Y|Z = 0, \boldsymbol{W}, \Delta = 1]).$$

Both G-computation and TMLE start with the same first step, which involves the estimation of the outcome mechanism and the potential outcomes. Then, G-computation computes ATE as the difference in the potential outcomes, meanwhile TMLE before computing ATE involves a second step, in which it incorporates information from the treatment assignment mechanism (68).

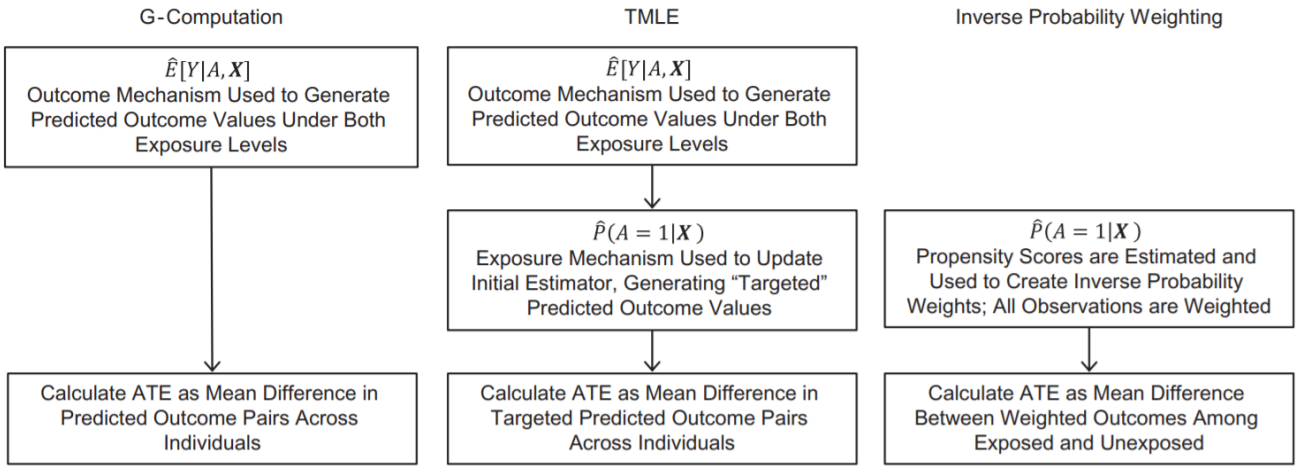


Figure 7: Commonalities and differences in the estimation sequence across 3 different estimators for the average treatment effect (ATE), extracted from (68).

Step 2: Estimate of the probability of treatment model

In the second step, the probability of receiving the treatment is estimated through SL, conditionally to the confounders. Mathematically:

$$g(\mathbf{W}) = P(Z=1|\mathbf{W}).$$

Such exposure mechanism is then used to update the initial estimate of $E(Y|Z,\mathbf{W}, \Delta = 1)$.

At this point, we have to compute the so called "clever covariates", that take advantage of the information contained in the PS model and in the missingness mechanism on the outcome model $P(\Delta = 1 \mid Z, \mathbf{W})$ as follows:

$$H(Z, W) = \frac{\Delta}{P(\Delta = 1 \mid Z,W)} \left(\frac{I(Z=1)}{P(Z = 1 \mid W)} - \frac{I(Z=0)}{P(Z = 0 \mid W)}\right),$$

were I() states for the indicator function. When this missingness mechanism on outcome is taken into account, we will refer to TMLE_MOD.

Step 3: Estimate the fluctuation parameter

Subsequently, the initial outcome regression model is updated, using information about the treatment mechanism obtained in step 2, to solve an estimating equation for the efficient influence fit. A logistic model is implemented:

$$logit\,(E(Y|Z,\boldsymbol{W},\Delta = 1) = logit\,(\hat{E}(Y|Z,\boldsymbol{W},\Delta = 1) + \varepsilon H(Z,\boldsymbol{W}),$$

Where the fluctuation parameter ε is estimated via a maximum likelihood approach.

Step 4: Update the initial estimates of the expected outcome

At this point, the initial estimates of the expected outcome are updated, using the inverse of the logit function, that we will indicate with *expit*. In this phase, TMLE modifies the initial estimate of $E(Y|Z,\mathbf{W})$ in order to get a less biased estimate of the target parameter.

1. Update the expected outcomes of all observations, given the treatment they actually received and their baseline confounders

$$\hat{E}^*[Y|\,Z, \mathbf{W}] = \text{expit}(\text{logit}(\hat{E}[Y|Z, \mathbf{W}]) + \hat{\varepsilon}\, H(Z, \mathbf{W})).$$

2. Update the expected outcomes, conditional on baseline confounders and everyone receiving the treatment

$$\hat{E}^*[Y|\,Z = 1, \mathbf{W}] = \text{expit}(\text{logit}(\hat{E}[Y|Z = 1, \mathbf{W}]) + \hat{\varepsilon}\, H(1, \mathbf{W})).$$

3. Update the expected outcomes, conditional on baseline confounders and no one receiving the treatment

$$\hat{E}^*[Y|\,Z = 0, \mathbf{W}] = \text{expit}(\text{logit}(\hat{E}[Y|Z = 0, \mathbf{W}]) + \hat{\varepsilon}\, H(0, \mathbf{W})).$$

Step 5: Compute the statistical estimands of interest and confidence errors

Now it is possible to compute the estimands of interest, that in our case is the ATE, as the mean difference in the updated outcome estimates under the two treatments:

$$\widehat{ATE_{TMLE}} = \widehat{\Psi_{TMLE}} = \frac{1}{N} \sum_{i=1}^{N} (\hat{E}^* [Y|Z = 1, \boldsymbol{W}] - \hat{E}^*[Y|Z = 0, \boldsymbol{W}]).$$

The efficient influence curve (IC) is then used to compute the Standard Error (SE) and the Wald-type 95% confidence interval (95% CI) (30, 37). More in detail, the IC is described by the following equation:

$$\widehat{IC} = (Y - \hat{E}^*[Y|Z, \boldsymbol{W}])\, H(Z, \boldsymbol{W}) + \hat{E}^*[Y|Z = 1, \boldsymbol{W}] - \hat{E}^*[Y|Z = 0, \boldsymbol{W}] - \widehat{ATE}.$$

Based on semiparametric and empirical processes theory the IC of a consistent and asymptotically linear estimator comes from the gradient of the pathwise derivative of the target parameter such that

$$\widehat{ATE} - \text{ATE} = \frac{1}{N} \sum_{i=1}^{N} IC_i - O_P(\tfrac{1}{N}).$$

Following the weak law of the large numbers, the O_p in the above equation converges to 0 at a rate of 1/N as the sample size (N) goes to infinity.

The IC is defined as a function of the observed data and the data-generating components that one can derive for a given model and target parameter that has mean 0 and finite variance. In sufficient large samples, the central limit theorem states that the variance of the estimator is thus the variance of the IC divided by N.

It follows that

$$\widehat{SE} = \sqrt{\frac{var(\widehat{IC})}{N}}$$

Where $var(\widehat{IC})$ is the sample variance of the estimated IC (79).

In this study, two different analyses were performed using TMLE. In the first (TMLE1), in SL were included only the learners included by default in the tmle() function of the *tmle* package in the R software (i.e. LR model with main terms only, LR model obtained from stepwise selection, LR model including interaction terms). In the second (TMLE2), in addition to the default learners, the following ones are also included: LR model with interaction terms obtained from stepwise procedures, generalized additive models (GAMs) (80), random forest (RF) (81) and recursive partitioning and regression trees (RPART) (82). TMLE analyses were performed on both CC data (TMLE CC) and considering the missingness mechanism on outcome data (TMLE MOD), using the IPW approach embedded in the TMLE itself, as described in step 2.

Real World case study: DARWIN-T2D

DARWIN-T2D was used as RWD. The main focus in this work is to compare patients which were undergoing Dapagliflozin (which is a SGLT2i drug) and patients that were initiated to a comparator GLM in the class of GLP-1RA, to compare the changes in glycemic efficacy parameters, as described more in detail in the previous chapter.

T2D patients initiated on Dapagliflozin or a GLP-1RA were compared to evaluate the proportion of patients with a simultaneous reduction in glycated haemoglobin (HbA1c) > 0.5%, body weight (BW) > 2 kg, and systolic blood pressure (SBP) > 2 mm Hg (55).

Missing covariate data were imputed through MICE algorithm (48), obtaining 5 imputed datasets. Only covariates with less than 40% missingness were included as predictors in the imputation process, including also the observed outcome values. However, the outcome variable was not imputed and we cannot exclude a priori a possible MNAR mechanism. In fact, as shown in the previous chapter, subjects with observed and missing outcome data significantly differ in terms of age, waist circumference, fasting glucose, total and LDL cholesterol, eGFR, insulin associated therapy and ACEi/ARBs therapy (55).

Then, PS-based model was estimated via LR approach in each imputed dataset, considering the following baseline covariates: age, sex, duration of diabetes, BW, BMI, FPG, HbA1c, SBP and DBP, total and HDL cholesterol, triglycerides, eGFR, insulin and metformin therapy, micro-angiopathy and macro-angiopathy.

Finally, outcome analyses were performed on each of the 5 imputed database, and results were pooled following Rubin's rules (83) and the within approach (61) .

More details about DARWIN-T2D data can be found in the previous chapters.

Simulation study

The most common approach to perform a simulation study in biomedical research is to assume that all random variables taken into account are conditionally independent. However, this is a very stringent assumption, which do not reflect reality.

An alternative and more realistic approach, is to construct a probabilistic model with conditional independence assumptions. More in detail, in this work data were simulated by defining a directed acyclic graph (DAG), as shown in Figure 8, using the *simcausal* R package (79).

The DAG was constructed to reflect the relationships between the main variables in the case study (DARWIN-T2D), i.e. sex (w_1), age at diagnosis (w_2), BMI (w_3), LDL cholesterol (w_4), insulin use (w_5) and macro-angiopathy (w_6). To establish these relationships from DARWIN-T2D, we used a Bayesian network (BN) to obtain the conditional probability distributions of the main variables. More details about the theory behind BNs can be found in Chapter 4. Peter-Clark stable algorithm with 100-fold bootstrap was applied for the structural learning of the BN (84). Finally, a more robust BN was obtained by averaging the 100 BNs obtained and considering only relationships between variables which were present in at least 95% of times (85).

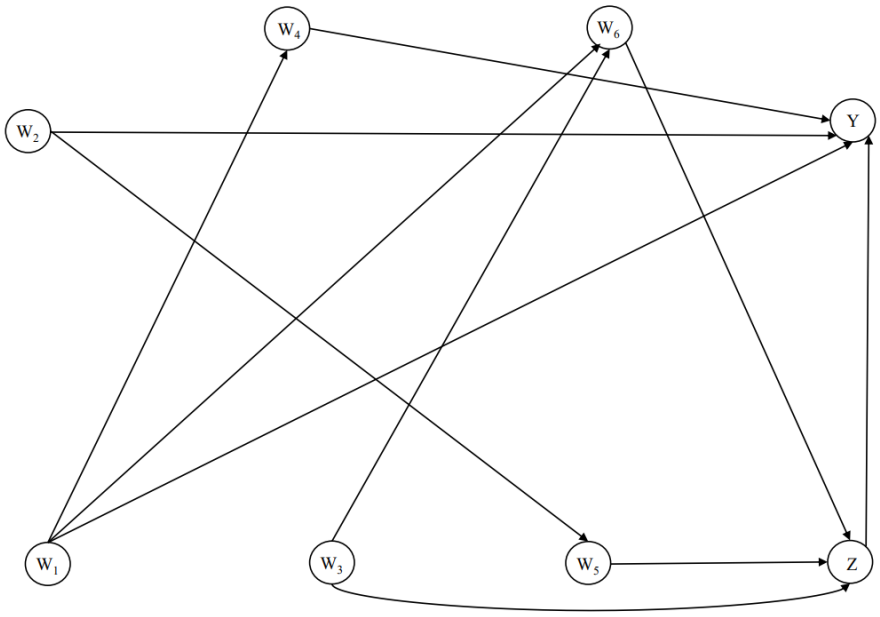


Figure 8: Direct acyclic graph (DAG) of the simulation scheme. **W** are the covariates, Y is a binary outcome, Z is a binary treatment.

Thus, following the dependencies between variables and collecting the summary statistics of the DARWIN-T2D variables, we performed the following simulation process:

$w_1 \sim$ Bernoulli (0.6);

$w_2 \sim$ Gaussian (mean = 60, sd = 8);

$w_3 \sim$ Gaussian (mean = 35, sd = 6);

$w_4 \sim$ Gaussian (mean = 90 if $w_1 = 1$ and mean = 97 if $w_1 = 0$, sd = 30),

$w_5 \sim$ Bernoulli (plogis (- 2 + 0.05 * w_2));

$w_6 \sim$ Bernoulli (plogis (-12 + 0.50 * w_3 - w_1));

Z ~ Bernoulli (plogis (-2 + 0.05 * w_3 – 0.20 * w_5 + 0.10 * w_6));

Y ~ Bernoulli (plogis (-3 + Z – 0.05 * w_2 + 0.05 * w_4 – 0.80 * w_1 – 0.20 * w_1 * w_2)), where plogis is the inverse logit function: 1/ log[p/(1p)].

Since in the study in the previous chapter, we observed that the subjects with observed and missing outcome differed in many covariates, we can exclude that a MCAR mechanism is present in the outcome data. For this reason, we simulated two different scenarios under MNAR mechanism with 20% and 40% of missing data percentage on the outcome. More in detail, when MNAR mechanism was considered, the probability of the missingness in Y depends on the Y value itself and, in particular, if Y = 1, we set the probability of missingness to 70%.

The MAR mechanism was also analyzed, because it is not possible to state which kind of missingness mechanism is present in observed data. In this case, for each subject, a weighted sum score is computed via a linear regression equation of each covariate with the same weight, excluding the outcome variable. Then, to each patient is associated a probability of having missing outcome data based on the weighted sum score, i.e., subjects with a higher weighted sum score have a higher probability of missing outcome data (86).

The *ampute* function in the *mice* R package was used to simulated missingness mechanisms (86).

In the simulation study, we performed the comparisons between the different methods in the situation that both the treatment and outcome models were misspecified, that is the most realistic situation. More in detail, in the treatment model Z, w_2 and w_6 were included as covariates, and w_3 and w_4 were included in the outcome model (Y):

$$Y \sim \alpha_0 Z + \alpha_1 w_3 + \alpha_2 w_4$$

$$Z \sim \alpha_3 w_2 + \alpha_4 w_6$$

The true marginal OR value was 1.66 and it was evaluated on 5 000 000 observations generated through the DAG in Figure 8.

Sensitivity analyses were performed to evaluate the importance of sample size. So, the analyses were performed with 1 000 and 5 000 observations. Overall, 1 000 simulations were performed to estimate the mean bias, the SE and the 95% nominal coverage intervals (95% NCI).

The bias was defined as the differences between the true marginal OR and the mean of the 1 000 ORs estimated.

Results

Real World case study: DARWIN-T2D

From the 281 217 patients with T2D collected in the DARWIN-T2D study, longitudinal data were available for 2 484 patients that were undergoing Dapagliflozin and for 2 247 subjects who initiated a GLP-1RA medication.

Follow-up data were collected for 830 patients in the Dapagliflozin group and 811 in the GLP-1RA arm. The composite outcome was available for 473 patients who initiated Dapagliflozin and for 336 patients undergoing GLP-1RA. Therefore, there was a high percentage of missing outcome data (49%).

PS matching was performed between 229 subjects in each group. More details about this analysis are reported in the previous chapter and in (55).

In Table 9, results about the estimate of the treatment effect of Dapagliflozin compared to GLP-1RA obtained through the different approaches are reported. LR and PS-based methods performed following the CC approach, yield similar results: they do not underline any difference between the two treatments. On the other hand, TMLE2, both following the CC and the MOD approaches, obtained statistically significant results. More in detail, TMLE2 (MOD) gives an estimate of OR = 1.35 with a statistically significant 95% CI (1.04 – 1.73). Dapagliflozin seems to be more effective than GLP-1RA in the simultaneous reduction of hba1c, BW and SBP.

Table 9: Results of the DARWIN T2D study. Dapagliflozin vs GLP 1RA. OR = odds ratio, 95% CI = 95% confidence interval, LR = logistic regression, PS = propensity score, IPTW = inverse probability of treatment weighting, TMLE = targeted maximum likelihood estimator, CC = complete case, MOD = Missing Outcome Data. TMLE1: main terms LR, stepwise forward and backward model selection, main terms LR and interaction terms; TMLE2: main terms LR, stepwise forward and backward model selection, main terms LR and interaction terms, stepwise forward and backward model selection with interaction, generalized additive models (GAM), random forest (RF) and recursive partitioning and regression trees (RPART).

Method	**OR (95% CI)**
LR	0.82 (0.53 – 1.27)
PS matching	0.86 (0.53 – 1.41)
PS covariate	0.81 (0.52 – 1.26)
IPTW	0.85 (0.54 – 1.34)
TMLE 1 (CC)	1.33 (0.91 – 1.96)
TMLE 2 (CC)	**1.53 (1.09 – 2.14)**
TMLE 1 (MOD)	1.34 (0.95 – 1.91)
TMLE 2 (MOD)	**1.35 (1.04 – 1.73)**

Then, in Table 10 we can analyse the contribution of each single learner selected by SL in TMLE2 (MOD). We can see that only the LR model with interaction terms was never selected by SL. On the other hand, the LR with stepwise and GAM model contribute 51% and 31% of the weight in the optimal predictor, respectively. RF was the algorithm with the highest weight on the prediction of treatment assignment (Z model) with 53%, followed by the LR with stepwise procedures (19%), and RPART (18%). Finally, the missingness mechanism on the outcome was mainly modelled by the RPART algorithm (70%), followed by GAM (21%) and, with a low contribution, by LR (9%).

Table 10: Coefficients of the algorithms selected by the super learner algorithm in TMLE2 (MOD) for the DARWIN T2D study

	Model		
SL algorithms	**Y**	**Propensity Score**	**Missingness mechanism on Y**
LR	0	0	0.09
Step	0.51	0.19	0
Step + interactions	0.12	0.10	0
LR + interactions	0	0	0
GAM	0.31	0	0.21
RF	0.03	0.53	0
RPART	0.03	0.18	0.70

Simulation study

Results of the simulation study are reported in Table 11. TMLE resulted as the approach with the smallest bias and the smallest SE, with every percentage of missingness on outcome (20% and 40%), and for each sample size tested (1 000 and 5 000).

Furthermore, results showed that including missingness mechanism on outcome data in TMLE (TMLE 2 MOD) improves the OR estimation if compared with the CC approach.

Between PS-based methods, IPTW resulted as the best approach, with the lower bias and SE, and the higher 95% NC.

LR, PS matching and PS included as a covariate in the regression model, had comparable performances.

When simple size increased, bias decreased in all the approaches, as expected.

Table 11: Results of the simulation study with different scenarios and the MNAR mechanism on the outcome. OR= odds ratio, 95% CI = 95% confidence interval, SE = standard error, 95% NC = 95% nominal coverage interval, MNAR = missing not at random, n = sample size, LR = logistic regression, PS = propensity score, IPTW = inverse probability of treatment weighting, TMLE = targeted maximum likelihood estimator, CC = complete case, MOD = Missing Outcome Data. TMLE1: main terms LR, stepwise forward and backward model selection, main terms LR and interaction terms; TMLE2: main terms LR, stepwise forward and backward model selection, main terms LR and interaction terms, stepwise forward and backward model selection with interaction terms, generalized additive models (GAM), random forest (RF) and recursive partitioning and regression trees (RPART).

	LR	PS matching	PS covariate	IPTW	TMLE1 (CC)	TMLE2 (CC)	TMLE1 (MOD)	TMLE2 (MOD
SCENARIO 1: 20% MNAR on Y, n = 1 000								
Mean OR	1.89	1.90	1.88	1.84	1.79	1.79	1.75	1.75
Mean Bias	0.23	0.25	0.23	0.18	0.13	0.13	0.09	0.09
SE	0.67	0.84	0.68	0.66	0.58	0.58	0.58	0.57
95% NC	94.5	93.9	94.0	94.5	94.5	92.4	93.9	91.6
SCENARIO 2: 40% MNAR on Y, n = 1 000								
Mean OR	2.06	2.07	2.05	2.00	1.97	1.98	1.92	1.94
Mean Bias	0.40	0.41	0.39	0.34	0.31	0.32	0.26	0.28
SE	1.30	1.31	1.30	1.23	1.20	1.18	1.15	1.17
95% NC	94.2	96.0	94.6	95.5	94.9	92.4	94.2	91.2
SCENARIO 3: 20% MNAR on Y, n = 5 000								
Mean OR	1.78	1.77	1.77	1.74	1.70	1.70	1.66	1.66
Mean Bias	0.12	0.11	0.11	0.08	0.04	0.04	0.002	0.006
SE	0.25	0.28	0.25	0.25	0.22	0.22	0.22	0.22
95% NC	92.0	93.5	92.7	94.6	95.6	94.2	94.6	93.4
SCENARIO 4: 40% MNAR on Y, n = 5 000								
Mean OR	1.78	1.79	1.78	1.74	1.73	1.73	1.68	1.68
Mean Bias	0.12	0.13	0.12	0.08	0.07	0.07	0.02	0.02
SE	0.37	0.43	0.37	0.36	0.35	0.35	0.35	0.35
95% NC	95.2	95.6	95.1	95.7	95.8	94.8	96.1	95.2

The same conclusions were achieved by simulating a MAR mechanism on the outcome, as reported in Table 12. However, in this case, differences between CC and MOD approaches when using TMLE are less evident, as expected. Also in this case, LR, PS matching and PS used as covariate in a regression model had comparable performances, and IPTW resulted the preferable PS-based approach.

There are no relevant differences in terms of the 95% NCI in the MNAR scenario; however, they are slightly higher for TMLE approaches in the MAR scenario reported in Table 12.

Table 12: Results of the simulation study with different scenarios and the MAR mechanism on the outcome. OR = odds ratio, 95% CI = 95% confidence interval, SE = standard error, 95% NC = 95% nominal coverage interval, MAR = missing at random, n = sample size, LR = logistic regression, PS = propensity score, IPTW = inverse probability of treatment weighting, TMLE = targeted maximum likelihood estimator, CC = complete case, MOD = missing outcome data. TMLE1: main terms LR, stepwise forward and backward model selection, main terms LR and interaction terms; TMLE2: main terms LR, stepwise forward and backward model selection, main terms LR and interaction terms, stepwise forward and backward model selection with interaction terms, generalized additive models (GAM), random forest (RF) and recursive partitioning and regression trees (RPART).

	LR	PS	PS	IPTW	TMLE1	TMLE2	TMLE1	TMLE2
SCENARIO 1: 20% MAR on Y, n = 1 000								
Mean OR	2.00	1.99	2.00	1.95	1.79	1.79	1.78	1.78
Mean Bias	0.35	0.33	0.34	0.29	0.13	0.13	0.12	0.12
SE	0.56	0.58	0.55	0.51	0.36	0.35	0.35	0.35
95% NC	87.7	91.3	88.0	90.7	94.0	92.9	94.5	92.0
SCENARIO 2: 40% MAR on Y, n = 1 000								
Mean OR	2.08	2.06	2.07	2.00	1.82	1.82	1.81	1.80
Mean Bias	0.42	0.40	0.42	0.34	0.16	0.16	0.15	0.14
SE	0.69	0.74	0.69	0.63	0.43	0.43	0.44	0.43
95% NC	88.8	91.2	88.8	92.1	94.1	92.2	93.8	92.0
SCENARIO 3: 20% MAR on Y, n = 5 000								
Mean OR	1.96	1.93	1.96	1.90	1.76	1.76	1.75	1.75
Mean Bias	0.30	0.27	0.30	0.24	0.10	0.10	0.09	0.09
SE	0.35	0.34	0.35	0.30	0.17	0.18	0.17	0.17
95% NC	60.2	70.8	61.2	72.5	90.4	89.1	90.4	89.3
SCENARIO 4: 40% MAR on Y, n = 5 000								
Mean OR	2.03	2.00	2.03	1.96	1.80	1.80	1.79	1.79
Mean Bias	0.37	0.34	0.38	0.30	0.14	0.14	0.13	0.13
SE	0.44	0.43	0.44	0.37	0.22	0.22	0.22	0.22
95% NC	56.6	71.3	56.4	72.0	88.2	87.2	88.6	88.4

Discussion

In RWD the absence of randomization, confounding and the high percentage of missing data both in dependent and independent variables, are challenging open issues that make necessary the application of more advanced ML approaches, which are able to overpass such criticisms, like for example TMLE and SL.

In this chapter, the performances of traditional methods (LR and PS-based approaches) are compared with those of TMLE, which is a more advanced ML method which takes advantage of SL, that address model misspecification and missingness in the outcome data including non-parametric models. In fact, in the previous chapters, we observed that DARWIN-T2D has a big amount of missing outcome data

(about 50%) and we showed that the missingness mechanism could be MNAR or MAR, because patients with observed and missing outcome data significantly differ in many characteristics. For this reason, we mainly focused in this analysis on MNAR and MAR missingness mechanisms on outcome. They are of particular interest, because they are the less deepened and studied in literature and it is still unclear how to deal with them, even if they suffer more of the consequences of model misspecification if compared with MCAR *(87)*.

In some studies *(88,89)*, the authors showed that TMLE implemented with SL has the lowest bias if compared with misspecified parametric model based methods (like LR or PS-based approaches). However, a comparison of how the mechanism of missingness in the outcome variable affects their performance is still lacking in literature.

In this chapter we performed a simulation study, resembling the DARWIN-T2D characteristics taking advantage by the BN theory. The simulation study confirmed that TMLE has the lowest bias and SE, even when a large amount of missing outcome data was present, both under MNAR and MAR mechanisms.

Additionally TMLE (MOD), i.e. the one that includes the model about missingness mechanism on outcome, showed a better performance than CC, confirming that CC has to be used only when MCAR or MAR mechanisms are present *(66)*.

The 95% NC intervals from the different methods that were applied to our analysis are similar when the outcome was affected by MNAR, while the 95% NC is higher for the TMLE when an MAR mechanism is considered.

In DARWIN-T2D, we obtained an opposite association when TMLE was applied, if compared with the estimates obtained via traditional methods, like LR or PS-based approaches. More in details, if we look at the result obtained via TMLE, we observed that Dapagliflozin simultaneously reduces Hba1c, BW and SBP, significantly more than GLP-1RAs. Since there are no RCTs providing background for this clinically relevant comparison, such results have a heavy interesting therapeutic implications for routine clinical practice.

Missingness patterns in RW settings may be driven by the characteristics of the different therapies being compared, thereby affecting the outcome comparison. We can try to explain the reversing of the OR by the fact that when the model is misspecified LR and PS-based methods gave a biased estimate of the marginal OR stretching in the direction of the conditional OR *(58)*. When the conditional and marginal treatment effects do not coincide and they are in opposite directions *(90)*; we refer to this situation as the

non-collapsibility of the OR *(91)*. Furthermore, because of the non-collapsibility of the OR, even a correctly specified LR model generally does not produce estimates of the marginal treatment effect *(79)*. The strength of TMLE is that algorithms included in SL aids to identify interactions between covariates and nonlinearities, which are not detected through traditional approaches and that could contribute to the change in direction of the OR. In fact, as sensitivity analysis we applied TMLE also without the intervention of the SL algorithm, but using the GLM approach only for the Y, Z and Δ models. In this case, we obtained a weaker OR with a non-statistically significant 95% CI (OR = 1.11; 95% CI 0.79 – 1.69).

However, TMLE has some limitations. For example, it is a very complex algorithm by a computational point of view, which is intensified by SL algorithm.

Furthermore, a limitation of the study is that only a few scenarios were considered in the simulation process, but this choice was justified by the aim of providing a simulation scheme as close as possible to DARWIN-T2D characteristics. In this view, we taken advantage of BN theory. However, we made some simulations also varying some settings, like the amount of missingness in the outcome or the sample size, to test stability and generalizability of our results, obtaining promising insights.

Results of this study suggest that in observational studies TMLE is able to simultaneously deal with misspecification and missingness on outcome data, even under MNAR (or MAR) scenarios. Furthermore, TMLE outperforms both the LR model and PS-based methods in terms of bias, SE and 95% NCI. In fact, traditional approaches require lots of assumptions to be satisfied and they are more suitable when MCAR or MAR mechanism on outcome are present but, from observed data, it is not possible to state which kind of missingness mechanism is underlying and we cannot a priori exclude the presence of a MNAR mechanism *(62)*, which amplifies the consequences of model misspecification.

So, the recommendation which arises from this study is to pay more attention to misspecification and to mechanism which is underlying the generation of missing outcome data and not a priori excluding MAR or MNAR schemes. Furthermore, it is advisable to use simulations which are tailored to the case study of interest, to identify which approach is the more adapt to that specific situation.

In conclusion, our study confirms that TMLE has appealing statistical properties. In fact, it is able to simultaneously deal with model misspecification through advanced ML algorithms used by SL and with missing outcome data, even under MNAR mechanism.

This chapter has been submitted as

Targeted maximum likelihood estimation of treatment effectiveness under outcome data missingness and model misspecification: a simulation study to assess results from the DARWIN-T2D study. Sciannameo V, Fadini GP, Bottigliengo D, Avogaro A, Baldi I, Gregori D, Berchialla P. American Journal of Epidemiology.

CHAPTER 4

ENROLMENT CRITERIA FOR DIABETES CARDIOVASCULAR OUTCOME TRIALS DO NOT INFORM ON GENERALIZABILITY TO CLINICAL PRACTICE: THE CASE OF GLUCAGON-LIKE PEPTIDE-1 RECEPTOR AGONISTS

Introduction

In the field of diabetes pharmacotherapy, before marketing authorization approval large cardiovascular safety trials (CVOTs) are needed to evaluate the effect of new glucose-lowering medications (GLMs) against comparators or placebo (10). Such CVOTs are designed mainly to show the superiority of the new treatment in the reduction of major adverse cardiovascular outcome events (MACE) in patients affected by T2D.

Lots of CVOTs, published mainly by The New England Journal of Medicine and by The Lancet, showed the efficacy of many drugs belonging to the class of glucagon-like peptide-1 receptor agonists (GLP-1RAs), in the prevention of cardiovascular complications. For example, the LEADER study showed the superiority of liraglutide, the SUSTAIN-6 study highlighted the efficacy of semaglutide, the HARMONY study pointed out the superiority of albiglutide and the REWIND study showed the efficacy of dulaglutide. All these drugs were compared with placebo, to evaluate the reduction of the rates of three-point MACE (cardiovascular death, non-fatal myocardial infarction or stroke) (92) (93) (38) (39) . All these prestigious studies shared the same conclusion: GLP-1RAs improve the cardiovascular outcomes of T2D patients (94).

However, the main limitation of such studies is that, to rapidly collect a sufficient number of cardiovascular events, patients enrolled in CVOTs are high-risk subjects. So, doubts arise about how much these patients are representative of the entire population affected by T2D (95). Furthermore, RCTs enroll a very selected group of patients, which are typically more motivated, compliant, and instructed in drugs use, free from co-morbidities and younger, if compared with the real world general population of T2D patients (10). On the other side, RWD well describe the patients that really may receive a drug (10). This lead to questions on how much CVOT results are generalizable to the routinely clinical practice setting and how much the differences between patients enrolled in RCTs and real world subjects have an impact on the generalizability of trial results.

Only a few studies have analyzed what proportion of T2D patients from various real world clinical care settings satisfy the I/E criteria of specific CVOTs (96–99). However, none of them analyzed what is the proportion of real-world patients corresponding to CVOT populations.

The aim of this study was to show how much the eligible population of patients differs from those of CVOTs, and calculate what proportion of patients from routine care would generate a true CVOT-like population.

Material and Methods

Data from the DARWIN-T2D study were used (see previous chapters for more details).

We extracted data about I/E criteria regarding the following CVOTs on GLP-1RAs, published during 2015-2019:

- **LEADER**

  The "Liraglutide Effect and Action in Diabetes: Evaluation of Cardiovascular Outcome Results" (LEADER) trial was initiated in 2010, to evaluate the cardiovascular effect of liraglutide (GLP-1RA) when added to standard care in T2D patients.

  LEADER is a double-blind trial, where T2D patients with high cardiovascular risk were randomly allocated to receive liraglutide or placebo. The primary composite outcome was the first occurrence of death from cardiovascular causes, nonfatal myocardial infarction, or nonfatal stroke. A total of 9 340 patients were included in the study, and the primary outcome occurred in significantly fewer patients in the liraglutide group than in the placebo one (hazard ratio, 0.87; 95% CI 0.78 to 0.97). More details about the LEADER trial can be found in (39).

- **SUSTAIN-6**

  The "Trial to Evaluate Cardiovascular and Other Long-term Outcomes with Semaglutide in Subjects with Type 2 Diabetes" (SUSTAIN-6) was conducted to assess the non-inferiority of semaglutide (GLP-1RA) with an extended half-life of approximately 1 week, in comparison with placebo, in terms of cardiovascular safety in T2D patients. In this trial, 3 297 T2D patients were randomly allocated to a standard-care regimen receiving once-weekly semaglutide (0.5 mg or 1.0 mg) or to placebo for 104 weeks. The primary composite outcome was the first occurrence of cardiovascular death, nonfatal myocardial infarction, or nonfatal stroke, and a hazard ratio of 0.74

was obtained, with a 95% CI from 0.58 to 0.95, so the non-inferiority of semaglutide was assessed. More details can be found in (38).

- **EXSCEL**

The "Exenatide Study of Cardiovascular Event Lowering" (EXSCEL) trial has the main aim of assessing the long-term cardiovascular safety and efficacy of exenatide, administered once weekly in addition to usual care, in T2D patients who had a wide range of cardiovascular risk.
14 752 T2D patients were randomly allocated to treatment or placebo group. In the treatment group, subcutaneous injections of extended-release exenatide at a dose of 2 mg were administered. The primary composite outcome was the first occurrence of death from cardiovascular causes, nonfatal myocardial infarction, or nonfatal stroke. A hazard ratio of 0.91, 95% CI from 0.83 to 1.00 was obtained. More details about this trial can be found in (37).

- **REWIND**

The "Researching Cardiovascular Events with a Weekly Incretin in Diabetes" (REWIND) trial was designed to assess whether the addition of dulaglutide (GLP-1RA) to the anti-hyperglycemic regimen of middle-aged and older T2D patients safely reduces the incidence of cardiovascular outcomes compared with placebo. REWIND is a multicenter, randomized, double-blind trial conducted in 371 sites in 24 countries. Subjects were T2D patients with high cardiovascular risk, in fact they were subjects with either a previous cardiovascular event or cardiovascular risk factors. The primary outcome was the first occurrence of the composite endpoint of non-fatal myocardial infarction, non-fatal stroke, or death from cardiovascular causes and a hazard ratio of 0.88 was obtained with a 95% CI 0.79 – 0.99. More details can be found in (92).

- **PIONEER-6**

The "Peptide Innovation for Early Diabetes Treatment" (PIONEER-6) trial was specifically designed to investigate if T2D patients treated with oral semaglutide, which registered an excess in the cardiovascular risk. PIONEER-6 is randomized, double-blind, placebo-controlled trial. A total of 3 183 patients were randomly assigned to receive oral semaglutide or placebo. The primary outcome was the first occurrence of a major adverse cardiovascular event (death from cardiovascular causes, nonfatal myocardial infarction, or nonfatal stroke), and a hazard ratio of 0.79, 95% CI from 0.57 to 1.11 was found. More details can be found in (100).

- **HARMONY**

  The main aim of the HARMONY trial is to determine the safety and efficacy of albiglutide (GLP-1RA) in preventing cardiovascular death, myocardial infarction, or stroke. HARMONY is a double-blinded, randomized, placebo controlled trial conducted in 610 sites in 28 countries. 9 463 participants were enrolled and randomly assigned to groups, with a ratio of 1:1. A hazard ratio of 0.78, 95% CI 0.68–0.90 was obtained in the analysis of the primary outcome, which indicated that albiglutide was superior to placebo in the protection against cardiovascular outcomes. More details about HARMONY can be found in (93).

We adapted their I/E criteria to the availability of data in DARWIN-T2D, as reported in the supplementary Table 1, at the end of this chapter.

Patients in DARWIN-T2D with missing data for key I/E criteria were excluded from the analysis.

We identify the T2D patients of DARWIN-T2D which could be eligible to be included in each CVOT considered, and we compared the average clinical characteristics.

Finally, we extracted from DARWIN-T2D the largest subgroup of patients with average clinical characteristics similar to those of patients enrolled in each CVOTs considered.

None specific tool was already available to this task, so we developed a novel strategy, which is described below.

Statistical analysis

Descriptive statistics were reported, reporting for continuous variables means and standard deviations (SDs), meanwhile categorical variables were described via frequencies and percentages.

To evaluate the similarity between characteristics of groups, standardized mean differences (SMD) were used for each variable considered. We defined a good balancement if a SMD<10% was achieved.

Sampling CVOT-like populations

No tool was available to detect the largest subgroup of T2D patients in DARWIN-T2D with clinical characteristics in average similar to those of T2D patients enrolled in the CVOTs considered.

So, in this work, we developed a new strategy, based on the Bayesian Network (BN) theory.

More in detail, a BN is a graphical probabilistic model which represents knowledge about an uncertain domain, that uses Bayesian inference for probability computations. BNs are represented as direct acyclic graph (DAG), where each node corresponds to a unique random variable and conditional probabilities between variables which are conditionally dependent were represented as edge (Figure 9) (101).

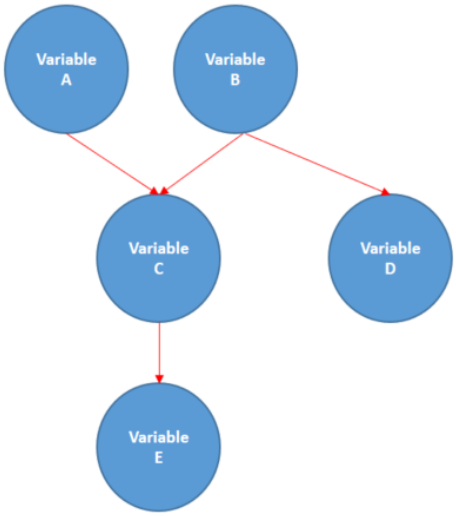


Figure 9. An example of DAG.

In the example shown in Figure 9, variables A and B are the so called "parents nodes" of the variable C, whereas the child of variable C is the variable E. The two variables A and B are marginally independent, but they become conditionally dependent, given variable C.

Each node is associated with a probability function that takes as input a particular set of values for the node's parent variables, and gives as output the probability of the variable represented by the node.

Following the paths specified in the BN, it is possible to obtain a factorized representation of the joint probability distribution by considering conditional dependences. If an edge exists between A and B, it means that P(B|A) is a factor in the joint probability distribution.

The joint distribution of a BN is equal to the product of P (node | parents(node)) for all nodes, expressed as follows:

$$P(X_1, \dots, X_n) = \prod_{i=1}^{n} P(X_i | X_1, \dots, X_{i-1}) = \prod_{i=1}^{n} P(X_i | Parents(X_i))$$

where *n* is the number of random variables taken into account.

More in details, in our study continuous variables in DARWIN-T2D were categorized into 5 classes, according to each CVOT summary statistics, making the assumption of normality of distributions.

Observations in DARWIN-T2D with missing data were deleted, because BN is more robust when dealing with complete data.

Then, one BN was constructed on DARWIN-T2D for each CVOT, because categorizations of continuous variables in DARWIN-T2D were made according to each CVOT's summary statistics. In this way, conditional probability distributions were obtained, reflecting the conditional dependencies among variables.

The variables included in the construction of BN were the following: age, sex, diabetes duration, hba1c, BMI, eGFR, SBP, DBP, heart failure, established cardiovascular events, cardiovascular risk factors.

The Peter-Clark stable algorithm with 100-fold bootstrap was employed for the structural learning of the BN (84). Then, we averaged 100 BNs learned, to obtain a more robust BN dealing with sampling variability, considering only relationships among variables obtained in at least 95% of times (85).

Finally, the set of probabilities for conditional nodes were computed as posterior estimations, whereas for unconditional nodes, probabilities were assigned by the computation of the ratio between CVOT and DARWIN-T2D frequencies, for each variable category, and then they were normalized to 1.

In this way, a final probability of inclusion in each CVOT for each patient in DARWIN-T2D was computed from the joint probability, which was decomposed into the product of conditional and unconditional probabilities through the BNs.

Subsequently, a random number was generated from a uniform distribution, between the maximum and the minimum value of the probability of inclusion, for each CVOT. Then, for every subject in DARWIN-T2D, if the probability of inclusion was greater than the half of the random number, he/she was included in the subsample of the DARWIN-T2D which is similar to the CVOT considered. Then, balancement of this group with CVOT was evaluated through SMD, and when all the SMDs were smaller than 10%, balancement was judged achieved. Elsewhere, if almost one variable resulted in a SMD>20%, starting from the variable with the higher SMD, 2% of patients with values in the tails of the distribution were sampled and removed from the DARWIN-T2D subsample.

We iteratively repeated this procedure until SMDs were all lower than 20%.

Finally, we joined together all of these balanced groups and, for each variable the same procedure was applied to reach a final SMD<10% for all the variables considered.

We performed a sensitivity analysis to assess the best thresholds of SMDs, and the choice of the double threshold 20% and 10% showed the best performances.

All of the analyses were performed using R version 3.5.0 (102).

Results

From the 281 217 patients with T2D collected in the DARWIN-T2D study, only those with complete data were retained for the analysis.

Among 130 380 patients with available data on GLMs, 6 699 (5.1%) were being treated with a GLP-1RA (73.8% liraglutide, 23.5% exeOW, 2.7% lixisenatide).

The numbers of patients in DARWIN-T2D who could be evaluated for CVOT eligibility were reported in Table 13. The greatest number was obtained for EXSCEL, meanwhile the smallest resulted when SUSTAIN-6 was taken into account, with only 98 725 patients evaluated for CVOT eligibility.

Table 13: Number of DARWIN-T2D patients evaluated for CVOT eligibility, the proportion of DARWIN-T2D patients after applying I/E criteria, and the DARWIN-T2D proportion of patients with CVOT-like characteristics.

	DARWIN-T2D		
CVOT	**Evaluated for CVOT eligibility**	**After applying I/E criteria, percentages of patients eligible for CVOTs**	**CVOT-like**
EXSCEL	124 164	13.4%	1%
PIONEER-6	116 553	34.1%	1.8%
HARMONY	107 040	9.5%	0%
LEADER	106 606	9.4%	1.2%
REWIND	105 074	35.8%	7.9%
SUSTAIN-6	98 725	10.1%	0.5%

After applying the I/E criteria, which are reported in Supplementary Table 1, the percentages of patients in DARWIN-T2D eligible for CVOTs ranged from 9.4% (LEADER) to 35.8% (REWIND). Such data were reported in Table 13.

Clinical characteristics of patients treated with GLP-1RA and of those eligible for CVOTs are reported in Table 14.

In Figure 10 A, we can see that the average clinical characteristics of patients eligible for CVOTs following the I/E criteria are different from the average features of patients who composed the CVOT trials, showing SMDs in general greater than 10% (white dots).

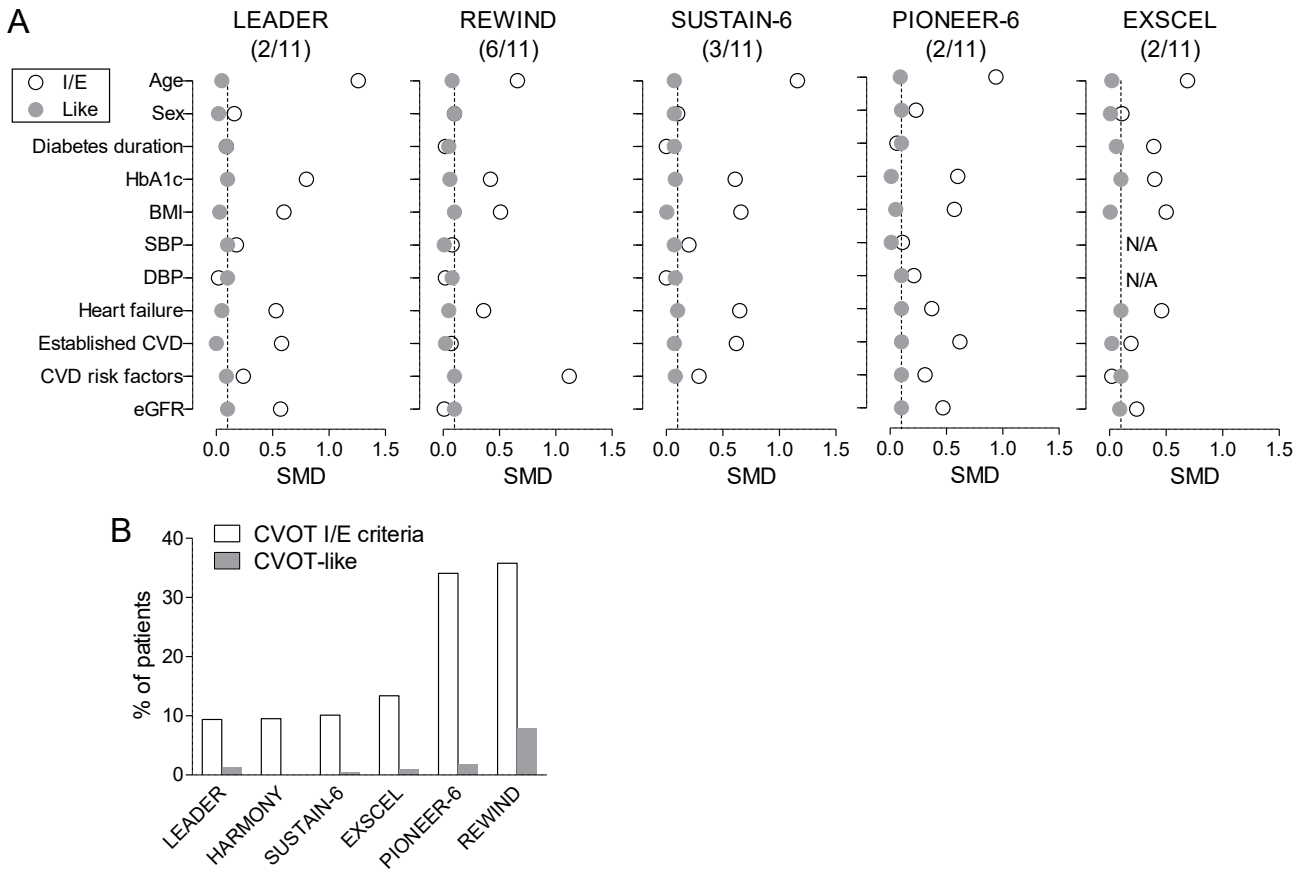


Figure 10: Real-world patients and CVOTs. A) For each CVOT, the panels show the absolute standardized mean difference (SMD) between the actual trial population (retrieved from respective publications) and real-world patients selected based on inclusion/exclusion criteria (I/E) or for being CVOT-like (Like). In each plot, a dashed line indicates the SMD threshold of 0.1, indicating good balance. Fractions in brackets refer to the number of key clinical characteristics that are matched between real-world patients selected by I/E and trial characteristics. By design, all characteristics were balanced between CVOT-like patients and the respective CVOT population. B) Proportion of real-world patients eligible for each CVOT based on I/E or sampled for being CVOT-like.

As reported in Table 15, patients in CVOTs were in general younger if compared with patients selected from the real-world database. Furthermore, despite 80%–100% of patients in LEADER and SUSTAIN-6 having established cardiovascular disease, applying the I/E criteria to the DARWIN-T2D yielded patients with a 70%–80% prevalence of micro-angiopathy (mostly chronic kidney disease) and a lower prevalence of macro-angiopathy (40%–50%). Imbalance of other clinical characteristics are reported in Table 15.

Out of 11 key clinical variables that we considered, eligible patients matched the trial characteristics with an absolute SMD smaller than 10% for just two or three variables, with the notable exception of REWIND. DARWIN-T2D patients eligible for REWIND were matched with the REWIND population for 6/11 variables.

Patients who were already on a GLP-1RA showed different clinical characteristics when compared with both those satisfying CVOT I/E criteria and those actually enrolled in CVOTs.

We then evaluated what proportion of real-world patients would constitute a population of individuals with key average characteristics similar to those enrolled in CVOTs.

BNs obtained on DARWIN-T2D, following the summary statistics of CVOTs are reported in Figure 11. They were used to compute the joint probability of inclusion for each DARWIN-T2D patients in each CVOT. The largest datasets of real-world patients yielding CVOT-like populations were 0.5% for SUSTAIN-6, 1.0% for EXSCEL, 1.2% for LEADER, 1.8% for PIONEER-6 and 7.9% for REWIND, as reported in Table 13. We were not able to obtain a dataset of DARWIN-T2D patients who would match the population of the HARMONY study (Figure 10B).

Table 14: Clinical characteristics of patients treated with GLP-1RA and of those eligible for CVOTs.

	GLP-1RA users in DARWIN-T2D	LEADER	SUSTAIN-6	EXSCEL	REWIND	PIONEER-6	HARMONY
Number	6699	10061	9942	16544	37574	39726	10208
Percentage[a]	5.1	9.4	10.1	13.4	35.8	34.1	9.5
Age, years	61.7±9.5	74.2±8.4	74.2±8.5	70.8±8.6	70.8±7.1	73.7±8.3	73.6±9.1
Sex male, %	54.9	56.5	55.8	67.3	59.0	57.6	68.4
Diabetes duration, years	11.3±7.5	13.6±9.1	13.6±9.1	15.4±9.8	10.7±8.2	14.3±10.0	17.9±10.
Active smoke, %	19.8	13.5	13.6	17.1	14.3	14.0	16.1
Body mass index, kg/m^2	34.8±6.2	29.1±5.1	29.2±5.2	29.1±4.9	29.8±4.7	29.3±5.2	29.1±4.9
Waist circumference, cm	114.7±13.6	103.9±12.2	104.1±12.4	104.2±11.9	104.5±11.3	104.5±12.6	104.6±11
Systolic blood pressure, mm Hg	138.3±18.3	139.1±18.6	139.2±18.7	137.5±18.4	138.4±17.9	138.1±18.7	137.5±18
Diastolic blood pressure, mm Hg	80.3±9.9	76.9±9.3	77.0±9.4	76.5±9.2	77.8±9.1	76.3±9.5	75.8±9.2
Heart rate, bpm	78.8±11.9	73.3±11.9	73.5±11.8	71.7±11.6	73.3±11.8	72.8±11.8	70.2±11.
Fasting plasma glucose, mg/dl	151.2±42.7	151.3±36.9	155.6±42.1	150.2±42.2	136.8±33.7	146.4±47.7	160.4±50
HbA1c, %	7.5±1.1	7.8±0.6	8.0±1.0	7.6±0.8	6.9±0.9	7.4±1.3	8.1±1.0
Total cholesterol, mg/dl	169.3±37.5	168.2±38.2	169.0±38.8	160.5±38.3	170.0±37.4	166.8±38.8	158.2±39
HDL cholesterol, mg/dl	45.9±12.6	48.7±13.6	48.4±13.5	47.1±13.6	49.8±13.8	49.0±14.5	45.7±13.
Triglycerides, mg/dl	160.2±87.5	141.5±74.2	144.2±76.5	140.9±85.3	134.6±71.7	137.4±77.6	146.9±88
LDL cholesterol, mg/dl	91.9±32.3	91.4±32.2	91.9±32.7	85.3±31.7	93.3±32.1	90.5±32.6	83.1±32.
eGFR, ml/min/1.73 m^2	87.7±24.1	68.3±21.5	68.5±21.9	73.3±21.6	76.2±19.4	66.7±20.5	68.9±22.
Albumin excretion rate, mg/g	51.5±147.7	59.5±98.7	61.2±109.5	42.5±108.7	45.3±66.0	57.1±155.0	43.4±132
GLMs, %							
Insulin	24.9	25.7	27.8	41.0	14.6	43.3	56.4
Metformin	85.9	75.2	73.8	67.8	83.4	62.1	59.5
Sulphonylurea / repaglinide	26.4	52.9	52.0	28.7	32.1	28.0	30.9
Acarbose	2.0	3.5	2.7	1.8	2.2	2.6	2.4
Pioglitazione	9.0	5.8	3.5	3.7	4.0	3.7	3.1
DPP-4 inhibitors	0.2	0.0	0.0	27.6	0.0	0.0	28.0
GLP-1RA	100.0	0.0	0.0	0.0	0.0	0.0	0.0
SGLT-2 inhibitors	0.7	5.7	5.4	4.5	3.4	3.6	5.4

Other therapies, %							
Anti-platelet agents	46.4	58.5	58.5	74.6	51.3	60.9	84.0
Statin	62.9	64.3	63.9	76.1	63.4	64.8	79.4
Renin-angiotensin system blockers	74.5	71.2	71.3	74.0	70.7	72.1	75.9
Calcium channel blockers	25.6	27.0	27.1	28.6	25.7	27.4	29.4
Beta-blockers	31.5	36.4	36.6	44.5	32.7	36.9	49.7
Diuretics	15.8	21.4	21.6	23.7	15.0	25.2	30.5
Complications, %							
Chronic kidney disease	10.0	40.7	40.7	29.1	20.9	44.9	37.4
Albuminuria >30 mg/g	37.3	59.0	59.7	33.5	40.7	57.6	32.1
Retinopathy	15.6	16.1	16.6	24.9	11.4	17.9	31.2
Peripheral neuropathy	14.8	21.2	21.8	25.9	17.2	23.5	30.4
Atherosclerosis obliterans	12.4	27.2	27.7	48.3	13.8	26.1	60.9
Peripheral revascularization	1.2	3.0	3.0	5.2	1.3	2.8	6.4
Diabetic foot	7.6	13.0	13.6	15.9	10.0	12.4	19.3
Stroke / Transient ischemic attack	2.2	9.5	9.8	11.3	4.8	9.3	14.4
Carotid atherosclerosis	39.1	47.5	47.6	51.4	42.1	45.3	54.7
Ischemic heart disease	8.2	20.9	20.9	44.2	11.7	21.0	56.7
Coronary revascularization	6.0	13.5	13.5	29.9	7.5	13.6	37.9
Micro-angiopathy	43.1	85.4	85.6	61.6	56.5	87.1	69.0
Macroangiopathy	30.4	52.3	52.5	79.8	37.2	50.3	98.2

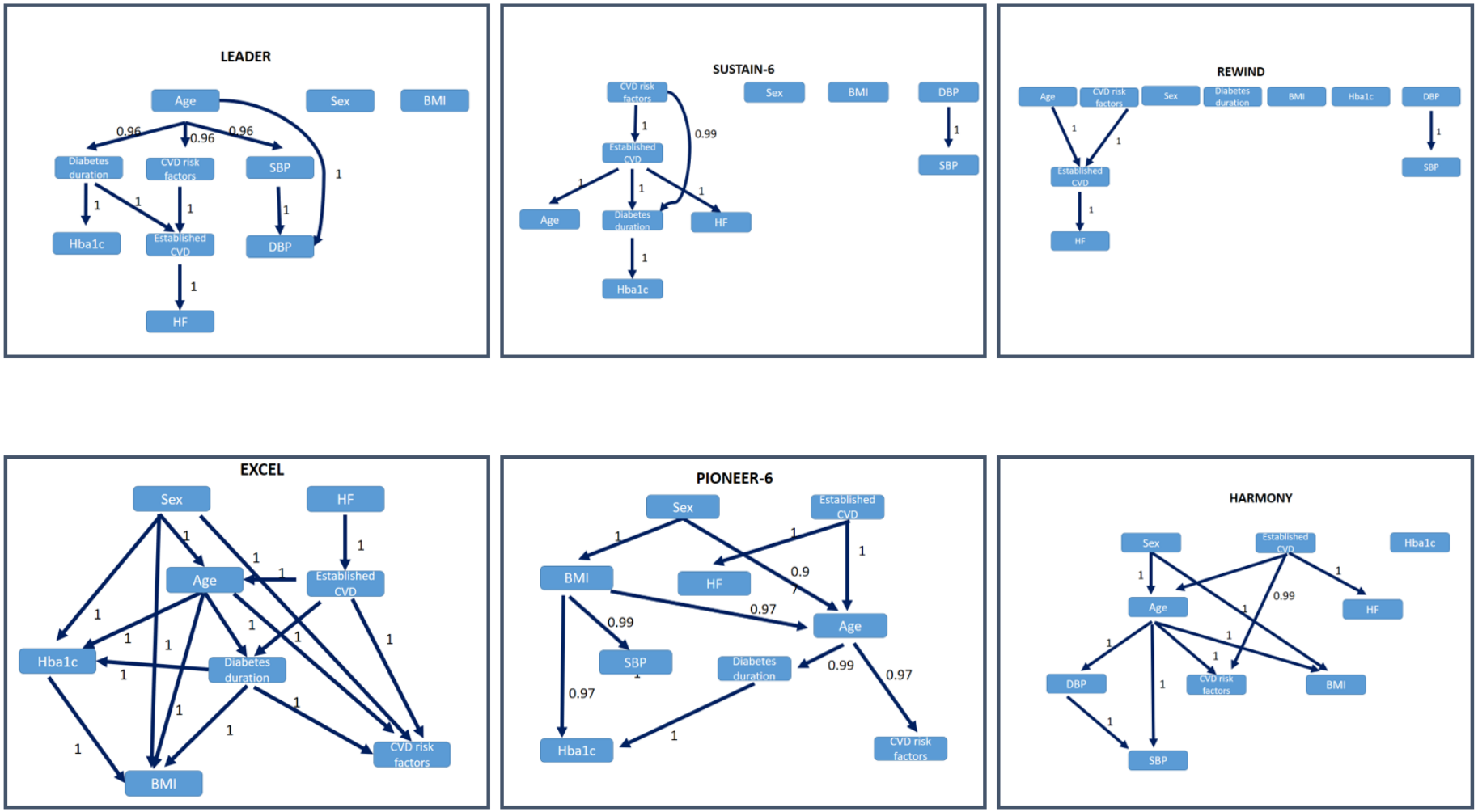


Figure 11: DAGs obtained through BNs in DARWIN-T2D data, based on the summary statistics of the corresponding CVOT reported in the title.

Table 15: Key clinical characteristics of real-world patients compared to CVOT patients. For each CVOT, we show the average clinical characteristics extracted from the respective publications, the characteristics of real-world patients who would be recruited into the CVOT based on inclusion / exclusion (I/E) criteria, and the characteristics of real-world patients sampled for being CVOT-like (Like). For both subgroups of real-world patients, we calculated the absolute standardized mean difference (SMD) as a measure of balance between groups. a SMD≤0.10 is conventionally considered indicative of a good balance. BMI, body mass index. SBP, systolic blood pressure. DBP, diastolic blood pressure. CVD, cardiovascular disease. eGFR, estimated glomerular filtration rate. N/A, not available. Established CVD and CVD risk factors are defined as described in each trial publication and slightly modified as illustrated in table S1.

Variable	**LEADER**	**I/E**	**SMD**	**Like**	**SMD**
Number	9340	10061		1132	
Age, years	64.3 (7.2)	74.2 (8.4)	1.26	64.6 (7.6)	0.05
Sex male, %	64.2	56.5	0.16	65.0	0.02
Diabetes duration	12.8 (8.0)	13.6 (9.1)	0.09	13.5 (8.4)	0.09
HbA1c, %	8.7 (1.5)	7.8 (0.6)	0.80	8.5 (0.8)	0.10
BMI, kg/m^2	32.5 (6.3)	29.1 (5.1)	0.60	32.7 (5.8)	0.03
SBP, mm Hg	135.9 (17.7)	139.1 (18.6)	0.18	137.8 (18.7)	0.10
DBP, mm Hg	77.1 (10.2)	76.9 (9.3)	0.02	78.2 (9.3)	0.10
Heart failure, %	17.9	2.5	0.53	16.0	0.05
Established CVD, %	81.4	55.7	0.58	81.4	0.001
CVD risk factors, %	18.7	28.7	0.24	22.3	0.09
eGFR, ml/min/1.73 m^2	80.4 (21.0)	68.3 (21.5)	0.57	78.0 (26.5)	0.10
Variable	**REWIND**	**I/E**	**SMD**	**Like**	**SMD**
Number	9901	37574		7280	
Age, years	66.2 (6.5)	70.8 (7.1)	0.66	66.7 (6.2)	0.08
Sex male, %	53.9	59.0	0.10	59.3	0.10
Diabetes duration	10.5 (7.2)	10.7 (8.2)	0.02	10.8 (7.1)	0.05
HbA1c, %	7.3 (1.1)	6.9 (0.9)	0.42	7.4 (1.2)	0.06
BMI, kg/m^2	32.3 (5.7)	29.8 (4.7)	0.51	31.9 (5.3)	0.10
SBP, mm Hg	137.0 (17.0)	138.4 (18.0)	0.08	137.2 (15.5)	0.01
DBP, mm Hg	78.0 (9.9)	77.8 (9.1)	0.02	78.7 (8.1)	0.08
Heart failure, %	8.7	1.0	0.36	7.3	0.05
Established CVD, %	31.4	28.2	0.07	30.6	0.02
CVD risk factors, %	68.6	19.9	1.12	63.8	0.10
eGFR, ml/min/1.73 m^2	75.0 (22.1)	75.2 (21.2)	0.009	77.4 (22.2)	0.10
Variable	**SUSTAIN-6**	**I/E**	**SMD**	**Like**	**SMD**
Number	3297	9942		476	
Age, years	64.6 (7.4)	74.2 (8.5)	1.16	65.1 (6.7)	0.07
Sex male, %	60.7	55.8	0.10	64.1	0.07
Diabetes duration	13.9 (8.1)	13.6 (9.1)	0.00	14.5 (6.8)	0.07
HbA1c, %	8.7 (1.5)	8.0 (1.0)	0.61	8.6 (0.7)	0.08
BMI, kg/m^2	32.8 (6.2)	29.2 (5.2)	0.66	32.8 (5.9)	0.004
SBP, mm Hg	135.6 (17.2)	139.2 (18.7)	0.20	136.8 (17.5)	0.07
DBP, mm Hg	77.0 (10.0)	77.0 (9.4)	0.00	77.8 (9.7)	0.08
Heart failure, %	23.6	2.6	0.65	19.5	0.10
Established CVD, %	83.0	55.7	0.62	80.5	0.07
CVD risk factors, %	17.0	29.0	0.29	14.1	0.08
eGFR, ml/min/1.73 m^2	N/A	N/A	N/A	N/A	N/A

Variable	PIONEER-6	I/E	SMD	Like	SMD
Number	3183	39726		1663	
Age, years	66.0 (7.0)	73.7 (8.3)	0.94	66.6 (7.2)	0.09
Sex male, %	68.4	57.6	0.23	72.9	0.10
Diabetes duration	14.9 (8.5)	14.3 (10.0)	0.06	14.0 (8.6)	0.10
HbA1c, %	8.2 (1.6)	7.4 (1.3)	0.60	8.2 (0.7)	0.01
BMI, kg/m^2	32.3 (6.5)	29.3 (5.2)	0.57	32.0 (3.7)	0.05
SBP, mm Hg	136.0 (18.0)	138.1 (18.7)	0.11	135.7 (14.0)	0.01
DBP, mm Hg	74.0 (21.0)	76.3 (9.5)	0.21	77.0 (8.0)	0.10
Heart failure, %	12.2	2.7	0.37	8.7	0.10
Established CVD, %	84.7	58.1	0.62	80.7	0.10
CVD risk factors, %	15.3	27.8	0.31	19.1	0.10
eGFR, ml/min/1.73 m^2	76.0 (10.0)	66.2 (21.3)	0.47	71.6 (26.0)	0.10
Variable	**EXSCEL**	**I/E**	**SMD**	**Like**	**SMD**
Number	14752	16544		915	
Age, years	62.0 (16.3)	70.8 (8.6)	0.69	62.3 (5.6)	0.02
Sex male, %	62.0	67.3	0.11	62.4	0.008
Diabetes duration	12.0 (7.4)	15.4 (9.8)	0.39	11.5 (6.9)	0.06
HbA1c, %	8.0 (1.2)	7.6 (0.8)	0.40	7.9 (0.7)	0.10
BMI, kg/m^2	31.8 (5.9)	29.1 (4.9)	0.50	31.8 (5.9)	0.006
SBP, mm Hg	N/A	N/A	N/A	N/A	N/A
DBP, mm Hg	N/A	N/A	N/A	N/A	N/A
Heart failure, %	16.2	2.9	0.46	12.6	0.10
Established CVD, %	73.1	64.5	0.19	72.3	0.02
CVD risk factors, %	26.9	27.9	0.02	22.3	0.10
eGFR, ml/min/1.73 m^2	76.3 (22.9)	70.5 (25.5)	0.24	78.4 (31.5)	0.09
Variable	**HARMONY**	**I/E**	**SMD**	**Like**	**SMD**
Number	9463	10208			
Age, years	64.1 (8.7)	73.6 (9.1)	1.07	N/A	N/A
Sex male, %	69.0	68.4	0.01	N/A	N/A
Diabetes duration	14.1 (8.7)	17.9 (10.3)	0.40	N/A	N/A
HbA1c, %	8.7 (1.5)	8.1 (1.0)	0.47	N/A	N/A
BMI, kg/m^2	32.3 (5.9)	29.1 (4.9)	0.59	N/A	N/A
SBP, mm Hg	134.7 (16.5)	137.5 (18.6)	0.16	N/A	N/A
DBP, mm Hg	76.8 (10.1)	75.8 (9.2)	0.10	N/A	N/A
Heart failure, %	20.0	4.3	0.50	N/A	N/A
Established CVD, %	100.0	85.4	0.58	N/A	N/A
CVD risk factors, %	0.0	33.6	1.00	N/A	N/A
eGFR, ml/min/1.73 m^2	79.0 (25.5)	68.9 (22.1)	0.42	N/A	N/A

Discussion

Real-World T2D patients are characterized by clinical features that often significantly differ from those of the patients enrolled in CVOTs. At our knowledge, this is the first time that proportions of RW T2D patients who have characteristics similar to those enrolled in CVOTs are computed.

We observed that such proportions are very small, ranging from 0.5% to 7.9%.

Many CVOTs have shown that some GLP-1RAs have the capacity to reduce the rate of adverse cardiovascular outcomes in T2D patients (94). However, doubts about the generalizability of such findings arise, because the populations analysed in CVOTs are in general very different from RW T2D patients. In fact, often T2D patients enrolled in CVOTs have higher risk of development of cardiovascular events if compared with the RW T2D population, to reduce the time needed to observe the CVD outcome of interest. The representativeness of the RW T2D population is an aspect rarely investigated in CVOT designs, but it deserves more attention, to understand how much results obtained in CVOTs could be generalize to the RW population seen in the clinical practice routine.

Prior studies examined the proportions of patients from RW databases eligible for CVOTs on GLP-1RAs and SGLT-2is, but only applying the I/E criteria. For example, Boye et colleagues (98) reported the proportions of US adult T2D patients which have similar characteristics to patients enrolled in LEADER, SUSTAIN-6, ESCEL and REWIND, obtaining results similar to those obtained in our analysis. Considering the fact that CVOTs are mainly conducted in US, such small differences between the proportions obtained by Boye and those obtained in our study, suggest that geographical and cultural factors may have a very small impact. Nicolucci et al (96) observed that RW T2D patients eligible for CVOTs on SGLT-2is are different to trial populations in many instances.

However, no study computed the proportions of RW patients which constitute CVOT-like populations.

If we look at the GLP-1RA CVOTs, we found many differences between the eligible RW population and those enrolled in CVOTs. In this study, we found a high proportion of patients in DARWIN-T2D eligible for PIONEER-6, which reflects enrolment criteria that, different to those of EXSCEL, lacked constraint on the ratio between patients with established cardiovascular disease and those with multiple cardiovascular risk factors. However, the PIONEER-6 eligible subgroup was imbalanced if compared with the true PIONEER-6 population.

We therefore examined what proportion of patients from DARWIN-T2D would generate CVOT-like populations, developing a novel approach based on the BN theory, to sample patients from a large RW dataset based on given average clinical characteristics.

In this way, we found that the greatest subset of patients with CVOT-like characteristics was much smaller than the proportion of eligible patients obtained through I/E criteria.

Furthermore, REWIND was confirmed as the CVOT mostly represented within the T2D population, even if only 7.9% of patients in DARWIN-T2D database were REWIND-like. On the contrary, the apparently large generalizability of PIONEER-6 based on I/E criteria was not confirmed.

DARWIN-T2D database has some limitations, which are inherent of observational studies. For example, there is a big amount of missing data, both in dependent and independent variables (about 50%), that can potentially affect our analysis. Under-reporting is then another issue typical in the RW context, where data are collected for clinical purposes and not to conduct medical research.

In conclusion, our study confirms that CVOT populations are extremely specific and that they are poorly represented by RW T2D patients. Furthermore, such results suggest that generalizability of trial populations to clinical practice should not be based on I/E criteria only, which can lead to misleading conclusions. Observational studies are needed to complement CVOTs findings and they are of fundamental importance to evaluate effectiveness of medications in a RW context.

This chapter was published as

Enrolment criteria for diabetes cardiovascular outcome trials do not inform on generalizability to clinical practice: The case of glucagon-like peptide-1 receptor agonists. Sciannameo V, Berchialla P, Orsi E, Lamacchia O, Morano S, Querci F, Consoli A, Avogaro A, Fadini GP; DARWIN-T2D study. Diabetes Obes Metab. 2020 May;22(5):817-827. doi: 10.1111/dom.13962. Epub 2020 Feb 6. PMID: 31943710

Supplementary Table 1: I/E criteria and application to the DARWIN-T2D database.

LEADER

Inclusion criteria	**Applied**	**Note or reason for not applying**
Type 2 diabetes	X	
HbA1c ≤9.5%	X	
Anti-diabetic drug naïve or treated with one or more oral anti-diabetic drugs or treated with human NPH insulin or long-acting insulin analogue or premixed insulin, alone or in combination with OAD(s)	X	
HbA1c ≥7.0%	X	
Prior cardiovascular disease cohort: age ≥50 and ≥1 of the following criteria: Prior MI; Prior stroke or TIA; Prior coronary, carotid or peripheral arterial revascularization; >50% stenosis of coronary, carotid, or lower extremity arteries; History of symptomatic CHD documented by positive exercise stress test or any cardiac; imaging or unstable angina with ECG changes; Asymptomatic cardiac ischemia documented by positive nuclear imaging test, exercise test or dobutamine stress echo; Chronic heart failure NYHA class II-III; Chronic renal failure:; eGFR <60 mL/min/1.73m^2 (Modification of Diet in Renal Disease formula); eGFR <60 mL/min (Cockcroft-Gault formula)	X	Data on stress or imaging tests not available. CKD-EPI eGFR was used
No Prior cardiovascular disease group: Age ≥60 y and ≥1 of the following criteria: Microalbuminuria or proteinuria; Hypertension and left ventricular hypertrophy by ECG or imaging; Left ventricular systolic or diastolic dysfunction by imaging; Ankle-brachial index <0.9	X	No data on diastolic dysfunction or ABI<0.9
Exclusion criteria	**Applied**	**Note or reason for not applying**
Type 1 diabetes	X	
Calcitonin ≥50 ng/L		No data on calcitonin concentrations
Use of a GLP-1 receptor agonist (exenatide, liraglutide or other) or pramlintide or any DPP-4 inhibitor within the 3 months prior to screening	X	Information on ongoing therapy used because no timing information was available

Use of insulin other than human NPH insulin or long-acting insulin analogue or premixed insulin within 3 months prior to screening.	X	All patients using rapid-acting insulin were excluded
Acute decompensation of glycemic control	X	Patients with FPG >400 mg/dl were excluded
Acute coronary or cerebrovascular event in the previous 14 days		No info available on the timing of cardiovascular events
Currently planned coronary, carotid, or peripheral artery revascularization		No info on timing of events nor on plans to revascularize
Chronic heart failure (NYHA class IV)	X	
Current continuous renal replacement therapy	X	eGFR<15 ml/min/1.73 m^2
End-stage liver disease		ALT≥3.0 × normal used in place
History of solid organ transplant or awaiting solid organ transplant		No info available
Malignant neoplasm		No info available
Family or personal history of multiple endocrine neoplasia type 2 or familial medullary thyroid carcinoma		No info available
Personal history of non-familial medullary thyroid carcinoma		No info available

SUSTAIN-6

Inclusion criteria	**Applied**	**Note or reason for not applying**
Men and women with type 2 diabetes	X	
HbA1c ≥7.0% at screening	X	
Antidiabetic drug naïve, or treated with one or two oral antidiabetic drug(s), or treated with human Neutral Protamine Hagedorn (NPH) insulin or long-acting insulin analogue or pre-mixed insulin, both types of insulin either alone or in combination with one or two oral antidiabetic drug(s)	X	All patients using rapid-acting insulin were excluded
Age ≥50 years at screening and clinical evidence of cardiovascular disease: prior myocardial infarction; prior stroke or prior transient ischemic attack; prior coronary, carotid or peripheral arterial revascularization; more than 50% stenosis on angiography or imaging of coronary, carotid or lower extremities arteries; history of symptomatic coronary heart disease documented by e.g. positive exercise stress test or any cardiac imaging or unstable angina with ECG changes; asymptomatic cardiac ischemia documented by positive nuclear imaging test or exercise test or stress echo or any cardiac imaging; chronic heart failure New York Heart Association (NYHA) class II-III; chronic	X	Data on stress or imaging tests not available. CKD-EPI eGFR was used

renal impairment, documented (prior to screening) by estimated glomerular filtration rate below 60 ml/min/1.73 m^2 per MDRD		
Or Age ≥60 years at screening and subclinical evidence of cardiovascular disease: persistent microalbuminuria or proteinuria; hypertension and left ventricular hypertrophy by electrocardiogram or imaging; left ventricular systolic or diastolic dysfunction by imaging; ankle/brachial index less than 0.9	X	No data on diastolic dysfunction; diagnosis of peripheral arterial disease used in place of ABI<0.9
Exclusion criteria	**Applied**	**Note or reason for not applying**
Type 1 diabetes	X	
Use of other glucagon-like peptide-1 receptor agonist or pramlintide within 90 days prior to screening	X	Info on ongoing therapy used because no timing information was available
Use of any dipeptidyl peptidase-4 inhibitor within 30 days prior to screening	X	Info on ongoing therapy used because no timing information was available
Treatment with insulin, other than basal and pre-mixed insulin, within 90 days prior to screening (except for short-term use)	X	All patients using rapid-acting insulin were excluded
Acute decompensation of glycemic control requiring immediate intensification of treatment to prevent acute complications of diabetes (e.g. diabetes ketoacidosis) within 90 days prior to screening	X	Patients with FPG >400 mg/dl were excluded
History of chronic pancreatitis or idiopathic acute pancreatitis		No info available
Acute coronary or cerebrovascular event within 90 days prior to randomization		No info on timing of prior cardiovascular events
Currently planned coronary, carotid or peripheral artery revascularization		No info on the plan to revascularize
Chronic heart failure New York Heart Association class IV	X	
Chronic hemodialysis or chronic peritoneal dialysis	X	eGFR<15 ml/min/1.73 m^2
End-stage liver disease, defined as the presence of acute or chronic liver disease and recent history of one or more of the following: ascites, encephalopathy, variceal bleeding, bilirubin ≥2.0 mg/dl, albumin level ≤3.5 g/dl, prothrombin time ≥4 seconds prolonged, international normalized ratio ≥1.7 or prior liver transplant		ALT≥3.0 × normal used in place
A prior solid organ transplant or awaiting solid organ transplant		No info available
Diagnosis of malignant neoplasm in the previous 5 years (except basal cell skin cancer or squamous cell skin cancer)		No info available

Personal or family history of multiple endocrine neoplasia type 2 (MEN2) or familial medullary thyroid carcinoma		No info available
Personal history of non-familial medullary thyroid carcinoma		No info available
Screening calcitonin ≥50 ng/l		No info available
Any acute condition or exacerbation of chronic condition that would in the investigator's opinion interfere with the initial trial visit schedule and procedures		Does not apply in real-world
Known or suspected hypersensitivity to trial products or related product		Does not apply in real-world
Known use of non-prescribed narcotics or illicit drugs		No info available
Previous participation in this trial. Participation is defined as randomized		Does not apply in real-world
Simultaneous participation in any other clinical trial of an investigational agent. Participation in a clinical trial with investigational stent(s) is allowed		Does not apply in real-world
Receipt of any investigational medicinal product within 30 days prior to screening (Visit 1) or according to local requirements, if longer		Does not apply in real-world
Brazil: receipt of any investigational drug within one year prior to screening visit (Visit 1), unless there is a direct benefit to the research patient at the investigator's discretion		Does not apply in real-world
Any other factor likely to limit protocol compliance or reporting of adverse event at the discretion of the investigator		Does not apply in real-world
Female of childbearing potential who is pregnant, breast-feeding or intends to become pregnant or is not using an adequate contraceptive method (adequate contraceptive measure as required by local regulation or practice)	X	All females of childbearing potential excluded.

EXSCEL

Inclusion criteria	**Applied**	**Note or reason for not applying**
Patient has type 2 diabetes mellitus	X	
Patient will be able to see a usual care provider at least twice a year		Does not apply in real-world
Patient has an HbA1c of ≥6.5% and ≤10.0% and is currently using one of the following treatment regimens: - Treatment with up to three (i.e. 0 – 3) oral AHAs (concomitant use of DPP-4 inhibitors is permitted)	X	

- Insulin therapy, either alone or in combination with up to two (i.e. 0 – 2) oral AHAs (use of basal and prandial insulins is permitted in any combination of individual or premixed insulins)		
Patients with any level of CV risk and meeting all other inclusion criteria may be enrolled. Recruitment will be constrained such that approximately 30% will not have had a prior CV event and 70% will have had a prior CV event.	X	A random 30% sample of patients without prior CV event was selected
A prior CV event is defined as at least one of the following: - History of a major clinical manifestation of coronary artery disease i.e. myocardial infarction, surgical or percutaneous (balloon and/or stent) coronary revascularization procedure, or coronary angiography showing at least one stenosis ≥50% in a major epicardial artery or branch vessel - Ischemic cerebrovascular disease, including: History of ischemic stroke; strokes not known to be hemorrhagic will be allowed as part of this criterion; transient ischemic attacks (TIAs) are not included; History of carotid arterial disease as documented by ≥50% stenosis documented by carotid ultrasound, magnetic resonance imaging (MRI), or angiography, with or without symptoms of neurologic deficit - Atherosclerotic peripheral arterial disease, as documented by objective evidence such as amputation due to vascular disease, current symptoms of intermittent claudication confirmed by an ankle-brachial pressure index or toe-brachial pressure index less than 0.9, or history of surgical or percutaneous revascularization procedure	X	Info on imaging and ABI not available
Female patients must not be breast feeding and agree to use an effective method of contraception or must not otherwise be at risk of becoming pregnant	X	No info available on contraception All women with childbearing potential were excluded
Patient understands the trial procedures, alternative treatments available, the risks involved with the trial, and voluntarily agrees to participate by providing written informed consent		Does not apply to the real world
Patient agrees to provide permission to obtain all medical records necessary for complete data ascertainment during the follow-up period, and agrees to communication between the trial site and the usual care provider in order to facilitate routine care		Does not apply to the real world

Patient is 18 years or older at enrolment	X	
Exclusion criteria	**Applied**	**Note or reason for not applying**
Patient has a diagnosis of type 1 diabetes mellitus, or a history of ketoacidosis	X	No information on ketoacidosis available
Patient has a history of (≥2 episodes) of severe hypoglycemia within 12 months of enrolment		No information on hypoglycaemia available
Patient has ever been treated with an approved or investigational GLP-1 receptor agonist e.g., BYETTA (exenatide), BYDUREON (EQW), VICTOZA (liraglutide), LYXUMIA (lixisenatide), albiglutide, taspoglutide or dulaglutide	X	All patient already on GLP-1RA were excluded
Patient is enrolled in another experimental protocol which involves the use of an investigational drug or device, or an intervention that would interfere with the conduct of the trial		Does not apply to the real world
Patient has a planned or anticipated revascularization procedure		No information available on planned revascularization
Pregnancy or planned pregnancy during the trial period		No information on pregnancy, pregnancy plan or contraception
Patient has medical history that indicates a life expectancy <2 years or might limit the individual's ability to take trial treatments for the duration of the trial	X	Patients aged 85 or older were excluded
Patient has a history or current evidence of any condition, therapy, laboratory abnormality, or other circumstance which, in the opinion of the investigator or coordinator, might pose an unacceptable risk to the patient, confound the results of the trial e.g. if patient cannot comply with requirements of the trial, or likely to interfere with the patient's participation for the full duration of the trial		Does not apply to the real world
Patient has end-stage renal disease or an estimated glomerular filtration rate (eGFR) derived from serum creatinine (using the simple MDRD-4 formula) of <30mL/min/1.73m^2	X	CKD-EPI was used
Patient has a known allergy or intolerance to exenatide		Does not apply to the real world
Patient has a history of gastroparesis		Information not available
Personal or family history of medullary thyroid cancer or MEN2 (Multiple Endocrine Neoplasia Type 2) or calcitonin level of >40ng/L at baseline		Information not available
Patient has previously been randomized in EXSCEL		Does not apply to the real world
Patient has a history of pancreatitis		Information not available
Is an employee of Amylin Pharmaceuticals, LLC, Bristol-Myers Squibb Company, or AstraZeneca.		Does not apply to the real world

REWIND

Inclusion criteria	Applied	Note or reason for not applying
Type 2 diabetes	X	
HbA1c ≤9.5%	X	
Stable dose of 0, 1 or 2 oral glucose-lowering drugs ± basal insulin for ≥3 months	X	No information on the prior regimen and dose
BMI ≥ 23 kg/m^2	X	
If age ≥50 years, at least 1 of: prior MI; prior ischaemic stroke; coronary revascularization ≥2 years earlier; carotid or peripheral revascularization ≥2 months earlier; unstable angina hospitalization; image proven myocardial ischaemia; or percutaneous coronary intervention	X	No information on imaging nor on the timing
If age ≥55 years, any of the above or at least 1 of: documented myocardial ischaemia by stress test or imaging; >50% coronary, carotid or lower extremity artery stenosis; ankle–brachial index <0.9; eGFR persistently <60 mL/min/1.73 m2; hypertension with left ventricular hypertrophy; or persistent albuminuria	X	No information on imaging, stress tests, ABI.
If age ≥ 60 years, any of the above or at least 2 of: any tobacco use; use of lipid-modifying therapy or a documented untreated LDL cholesterol ≥3.4 mmol/L (130 mg/dL) within the past 6 months; HDL cholesterol <1.0 mmol/L (40 mg/dL) for men and <1.3 mmol/L (50 mg/dL) for women or triglycerides ≥2.3 mmol/L (200 mg/dL) within the past 6 months; use of ≥1 blood pressure drug or untreated systolic blood pressure ≥ 140 mm Hg or diastolic blood pressure ≥ 95 mm Hg; or waist-to-hip ratio >1.0 (men) and >0.8 (women)	X	Waist-hip ratio substituted with waist circumference
Run-in adherence to study drug = 100%		Does not apply in real-world
Signed informed consent		Does not apply in real-world
Exclusion criteria	**Applied**	**Note or reason for not applying**
Uncontrolled diabetes	X	Defined as FPG>400 mg/dl
Severe hypoglycaemia in preceding year		No info on hypoglycamias
Coronary or cerebrovascular event in preceding 2 months or plans to revascularize		No info on timing of events nor on plans to revascularize
eGFR <15 mL/min/1.73 m^2 or on dialysis	X	
Gastric bypass or emptying abnormality		Missing information
Prior pancreatitis/concordant symptoms		Missing information
Liver disease or ALT ≥3.0 × normal	X	ALT≥3.0 × normal used
Family history of/or C-cell hyperplasia or medullary thyroid cancer or MEN 2A or 2B or calcitonin value ≥20 pg/mL		Missing information

Unwilling to stop GLP-1 receptor agonist or DPP-4 inhibitor or weight loss drug	X	Patients on GLP-1RA or DPP-4i were excluded
Cancer within prior 5 years		Missing information
Pregnant or not using reliable birth control	X	All females of childbearing potential excluded.
Life expectancy <1 year	X	Patients aged 85 or older were excluded

PIONEER-6

Inclusion criteria	**Applied**	**Note or reason for not applying**
Informed consent		Does not apply to the real world
Male or female diagnosed with type 2 diabetes	X	
Age ≥50 years at screening and at least one of the following conditions: prior myocardial infarction; prior stroke or transient ischaemic attack; prior coronary, carotid or peripheral arterial revascularization; >50% stenosis on angiography or imaging of coronary, carotid or lower extremity arteries; history of symptomatic coronary heart disease documented by e.g. positive exercise stress test or any cardiac imaging or unstable angina pectoris with electrocardiogram changes; asymptomatic cardiac ischaemia documented by positive nuclear imaging test or exercise test or stress echo or any cardiac imaging; chronic heart failure New York Heart Association (NYHA) class II-III; moderate renal impairment (estimated glomerular filtration rate [eGFR] 30–59 mL/min/1.73 m^2)	X	Results of imaging and stress test not available; NYHA class not available; CKD-EPI equation used.
Or Age ≥60 years at screening and at least one of the following risk factors: microalbuminuria or proteinuria; hypertension and left ventricular hypertrophy by electrocardiogram or imaging; left ventricular systolic or diastolic dysfunction by imaging; ankle–brachial index <0.9	X	No information on diastolic dysfunction; diagnosis of peripheral arterial disease was used in place of ABI<0.9
Exclusion criteria	**Applied**	**Note or reason for not applying**
Current or previous (within 90 days prior to screening) treatment with any GLP-1 receptor agonist, DPP-4 inhibitor or pramlintide	X	Only current users excluded
Family or personal history of multiple endocrine neoplasia type 2 or medullary thyroid carcinoma		No information available
History of pancreatitis (acute or chronic)		No information available
History of major surgical procedures involving the stomach potentially affecting absorption of trial product (e.g. subtotal and total gastrectomy, sleeve gastrectomy, gastric bypass surgery)		No information available

Subjects presently classified as being in NYHA class IV heart failure	X	
Planned coronary, carotid or peripheral artery revascularisation known on the day of screening		No info on planned revascularization available
Any of the following: myocardial infarction, stroke or hospitalisation for unstable angina or transient ischaemic attack within the past 60 days prior to screening		No information on the timing of prior cardiovascular events available
Chronic or intermittent haemodialysis or peritoneal dialysis or severe renal impairment (corresponding to eGFR <30 mL/min/1.73 m2)	X	eGFR <30 mL/min/1.73 m^2
History or presence of malignant neoplasms within the last 5 years (except basal and squamous cell skin cancer and carcinoma in situ)		No information available
History of diabetic ketoacidosis		No information available
Proliferative retinopathy or maculopathy requiring acute treatment. Verified by fundus photography or dilated fundoscopy performed within 90 days prior to screening or within the period between screening and randomisation	X	All patients with proliferative retinopathy and macular edema were excluded
Female who is pregnant, breast-feeding or intends to become pregnant or is of childbearing potential and not using adequate contraceptive methods	X	All women of childbearing age excluded
Known or suspected hypersensitivity to the trial product or related products		Does not apply to the real world
Previous participation in this trial		Does not apply to the real world
Receipt of any investigational medicinal product within 90 days before screening. For Brazil only: Participation in other clinical trials within one year prior to screening unless there was a direct benefit to the research subject at the investigator's discretion		Does not apply to the real world
Participation in another clinical trial of an investigational medicinal product. Participation in a clinical trial which evaluate stent(s) was allowed		Does not apply to the real world
Any disorder, which in the investigator's opinion might jeopardise the patient's safety or compliance with the protocol		Does not apply to the real world

HARMONY

Inclusion criteria	**Applied**	**Note or reason for not applying**
Men or women at least 40 years old with a diagnosis of type 2 diabetes.	X	
Established cardiovascular disease, including at least 1 of the following:	X	Results of imaging and stress test not available; ABI not available.

- Coronary artery disease with EITHER of the following: Documented history of spontaneous myocardial infarction, at least 30 days prior to Screening; Documented coronary artery disease (CAD) ≥ 50% stenosis in 1 or more major epicardial coronary arteries, determined by invasive angiography, or history of surgical or percutaneous (balloon and/or stent) coronary revascularization procedure (at least 30 days prior to Screening for percutaneous procedures and at least 5 years prior to Screening for coronary artery bypass graft (CABG)). - Cerebrovascular disease – ANY of the following: Documented history of ischaemic stroke, at least 90 days prior to study entry; Carotid arterial disease with 50% stenosis documented by carotid ultrasound, magnetic resonance imaging or angiography, with or without symptoms of neurologic deficit; Carotid vascular procedure (e.g. stenting or surgical revascularisation), at least 30 days prior to Screening; - Peripheral arterial disease (PAD) with EITHER of the following: intermittent claudication and ankle:brachial index < 0.9 in at least one ankle; prior non-traumatic amputation, or peripheral vascular procedure (e.g. stenting or surgical revascularisation), due to peripheral arterial ischaemia.		
HbA1c >7.0% (53 mmol/mol) based on the most recent documented laboratory assessment measured no more than 6 months prior to randomization. Local laboratory HbA1c values taken as part of usual care are permitted	X	
Female: subject is eligible to participate if she is not pregnant (as confirmed by a negative urine human chorionic gonadotrophin (hCG) test for females of reproductive potential only), not breastfeeding, and at least one of the following conditions applies…	X	All women of childbearing age were excluded
Able and willing to provide informed consent.		Does not apply to the real world
Exclusion criteria	**Applied**	**Note or reason for not applying**
eGFR calculated using MDRD formula <30mL/min/1.73m2 (based on the most recent documented serum creatinine laboratory assessment measured no more than 6 months prior to randomization. Local laboratory creatinine values taken as part of usual care are permitted) or renal replacement therapy.	X	CKD-EPI equation was used; no timing available

Use of a GLP-1 receptor agonist at Screening.	X	Ongoing use of GLP-1RA
Severe gastroparesis requiring therapy within 6 months prior to Screening.		No information available
History of pancreatitis or considered clinically at significant risk of developing pancreatitis during the course of the study (e.g. due to symptomatic gallstones, excess alcohol use).		No information available
Personal or family history of medullary carcinoma of the thyroid or subject with MEN-2. Personal history of pancreatic neuroendocrine tumours. In the opinion of the investigator, the subject has a medical history which might affect his / her ability to remain in the study for its entire duration, or which might limit management, such as life expectancy of <5 years (e.g. due to active malignancy).		No information available
Subject has a medical history which in the opinion of the investigator might limit the individual's ability to take trial treatments for the duration of the study or to otherwise complete the study.		Does not apply to the real world
Breastfeeding, pregnancy, or planning a pregnancy during the course of the study. Pregnancy test will be required in women of child bearing potential. Women who have undergone a sterilisation procedure or who are clearly post-menopausal will not be required to undergo pregnancy testing. Women who have developed spontaneous secondary amenorrhoea of 12 months or more where post-menopausal status is in doubt, a blood sample where FSH >40MU/ml and oestrodiol <40 pg/mL (<140 pmol/L) are simultaneously measured will be considered confirmatory.	X	All women of childbearing age were excluded
Known allergy to any GLP-1 receptor agonist or excipients of albiglutide.		Does not apply to the real world
Use of another investigational product within 30 days or according to local regulations, or currently enrolled in a study of an investigational device.		Does not apply to the real world
Any other reason the investigator deems the subject to be unsuitable for the study		Does not apply to the real world

CHAPTER 5

TRANSPOSITION OF CARDIOVASCULAR OUTCOME TRIAL EFFECTS TO THE REAL-WORLD POPULATION OF PATIENTS WITH TYPE 2 DIABETES

Introduction

As it has been discussed in the previous chapter, many cardiovascular outcome trials (CVOTs) have been performed to demonstrate the safety of glucose lowering medications (GLMs) administered to T2D patients to protect them against cardiovascular events, which are more frequent in T2D patients if compared with the general population (95,103).

Many of these CVOTs showed lower rates of cardiovascular events among patients randomized to receive the active GLM added to the standard care if compared to those randomized to assume placebo or any comparator (104).

However, to rapidly collect a sufficient number of cardiovascular events, such trials enroll T2D patients with higher cardiovascular risks, if compared with the general population of T2D patients, which can really receive a GLM in the real world clinical practice (104).

Furthermore, in the previous chapter we showed that only a small proportion of RW T2D population would satisfy the enrolment criteria of CVOTs, and even smaller proportions have CVOT-like characteristics (98,105).

Consequently, it was raised an intense debate on whether results of CVOTs can be transferred to the general real-world population of T2D patients (106,107). So, after that we had assessed the generalizability of CVOTs to the general T2D population, successively it naturally arises the question about the transposition of the CVOTs effects to the real world T2D population, which has different characteristics from that of the trials.

In this chapter, we transposed the effects of GLP-1RA or SGLT2i registered in some CVOTs to the general T2D population of patients recruited in DARWIN-T2D, which were followed under routine specialist care and which could potentially be prescribed such medications in the RW routinely clinical practice.

Transposition and statistical analysis

The most diffused setting in which transposition of trial effect is performed in literature is when individual level data for both the trial and the target population are available. In such situation, which is the gold-standard, patients in the trial are weighted by the probability to meet the inclusion criteria and an outcome analysis is performed with the weighted trial data (108) (Table 16).

If instead individual-level data are available for the trial, and aggregated data are present for the target population, there exist four different approaches, which are resumed in Table 16 (109).

Contrariwise, in this study we disposed of individual-level data for DARWIN-T2D, which is the target population, and aggregated data for CVOTs. In such situation, no method has already been developed (Table 16) (109). So, we cannot apply the golden standard approach described above, i.e. weighting using simulated individual data or weighting using the method of moments, because it requires individual-level data for the trial, but we used the strata specific trial estimates to transpose the trial effect to the target population (110).

More in detail, we implemented a modified post-stratification approach using aggregated data from CVOTs and individual-level data from DARWIN-T2D. More specifically, we implemented an inverse approach compared to method 3 in Table 16, which was described previously (97, 98).

The method that we implemented to transpose the effect of GLMs on the prevention of cardiovascular outcomes obtained in CVOTs to DARWIN-T2D, is represented in the flow-chart in Figure 12.

First, patients from DARWIN-T2D with missing data were excluded because to date no method has been validated to apply the transposition approach to multiple imputed datasets.

Then, for each CVOT considered, we categorized continuous variables in DARWIN-T2D according to the stratum-specific HR estimates reported in the CVOTs' publications. Then, we collected the sub-group specific estimates from CVOTs and we used proportions of the categorized variables in DARWIN-T2D to compute the transposed treatment effect for the target population, by weighting the average of the stratum-specific treatment effects according to proportions of a given characteristic in the target population.

Table 16: Table modified from Hong 2019 (109). Description of Different Methods for Generalizing a Randomized Clinical Trial's Results to a Target Population.

Data availability in CVOT	**Target population**	
	Individual data	***Aggregate data***
Individual data	Such situation is the **Gold-standard**. The recommendation is the weighting approach using individual data. First, probabilities of being eligible for CVOT are computed via regression approaches and CVOT participants are reweighted to reflect the patient characteristics in the target population. Finally, outcome analyses are performed using weighted individual-level data of the trial.	**Method 1**. Weighting using simulated individual data. To simulate individual data based on aggregate data for target population and using gold-standard weighting method to perform the outcome analysis. **Method 2**. Weighting using the method of moments. To use the methods of moments to estimate the weights and then estimating treatment effect within the weighted trial's individual data. **Method 3**. Post-stratification. To compute weighted treatment effect estimate by reweighting subgroup-specific treatment effects in CVOT based on the distribution of a given effect modifier in the target population.
Aggregate data	**NONE**	**Method 4.** Expected absolute risk reduction. To multiply the observed outcome risk in the target population who were unexposed to treatment by the relative treatment effect in CVOT to obtain the expected risk in the target population if they were exposed to placebo. Next, calculating the expected absolute treatment effect in target population by subtracting the risk in the target population who were unexposed to treatment from the expected risk if the target population were exposed to treatment.

For example, let us consider the "gender" variable. In the REWIND trial 46% of participants were female, instead in DARWIN-T2D they were 44%. In REWIND, the stratum-specific HR estimates were 0.85 (95% C.I. 0.71–1.02) for females and 0.90 (95% C.I. 0.79–1.04) for males. Then, the weighted HR estimate for the variable "gender" is computed by the formula

$$\text{Transposed HR} = \exp\left(\frac{\sum \ln(HR_i) * p_i}{\sum p_i}\right), \quad i = 1,2$$

where p_i are the proportion in DARWIN-T2D in the level i of the variable. In the example above, the transposed HR was obtained as $\frac{\ln(0.85)*0.44+\ln(0.90)*0.56}{0.44+0.56}$ = - 0.13, that exponentiated leads to a HR = 0.88 (see Supplementary Table 1).

Then, standard deviations across strata were pooled together to obtain the 95% confidence interval.

Such calculation was performed for one characteristic at a time.

Finally, the unweighted average of the estimated transposed treatment effect of each characteristic was used to summarize the post-stratification estimates of treatment effect.

Analyses were performed using R version 3.5.2.

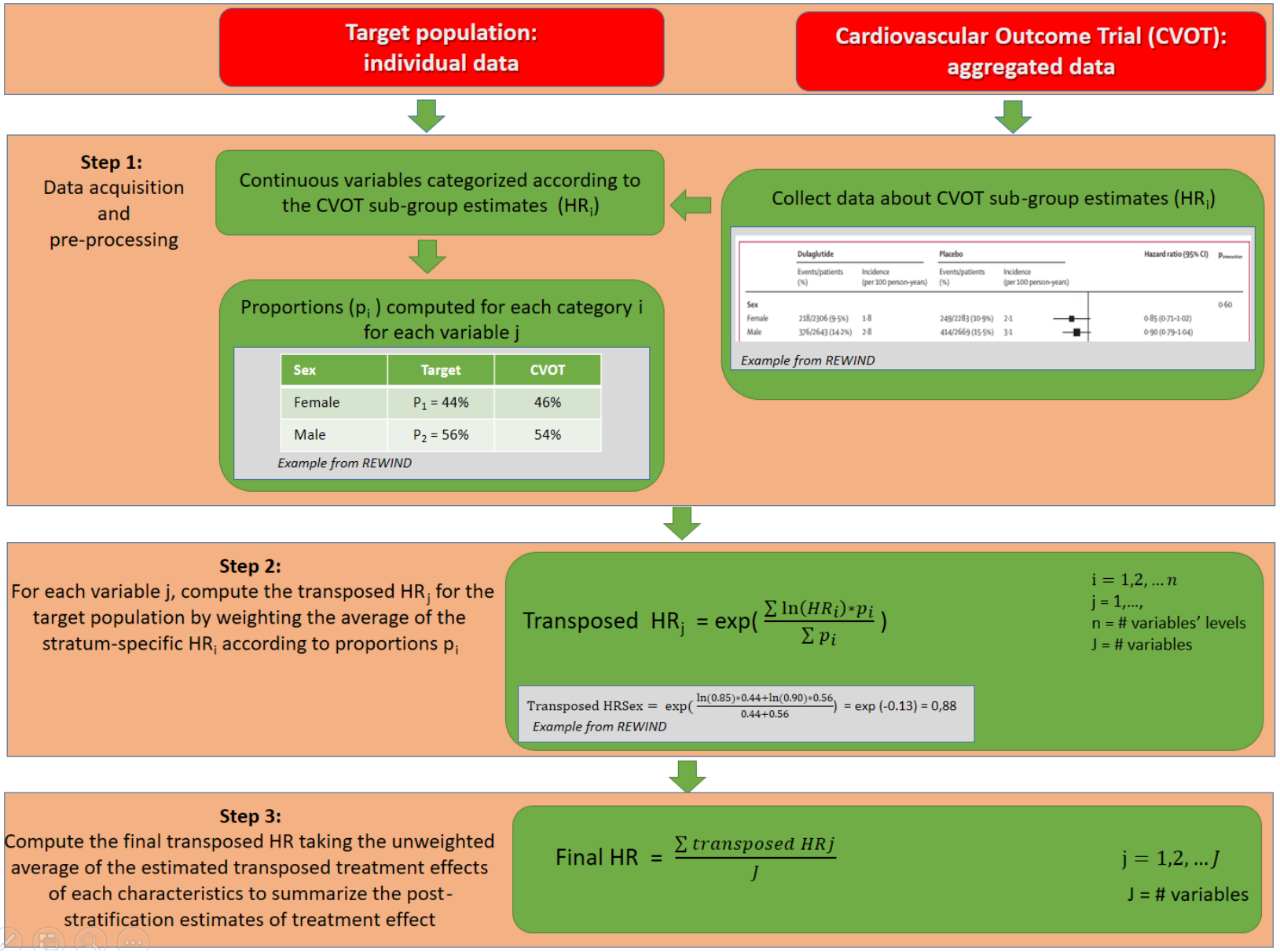


Figure 12: A flow-chart of the transposition method. The figure illustrates the 3-step procedure used to transpose a cardiovascular outcome trial (CVOT) result to the target population. An example from the REWIND study is described in the text.

Selection of CVOTs

Since the transposition approach that we used requires stratum-specific estimates of the treatment effect, we selected CVOTs which reported hazard ratios for the primary outcome (the 3-point major adverse cardiovascular events, which is a composite of non-fatal myocardial infarction, non-fatal stroke, or cardiovascular death, identified by the acronyms 3P-MACE) stratified by sub-groups of patients based on clinical characteristics of the trial population.

CVOTs were selected based on literature search, performed by an endocrinologist clinician, and the choose was based on whether key information were available. The search string which was used is ("cardiovascular" AND "outcome" AND "randomized" AND "trial" AND "type 2 diabetes").

The procedure of selection of CVOTs usable for this analysis lead to a list of 9 studies.

- EMPA-REG

EMPA-REG is a randomized, placebo-controlled CVOT designed with the main aim of studying the effects of empagliflozin (SGLT2i), in addition to standard care, on cardiovascular morbidity and mortality in T2D patients which have a high cardiovascular risk.

Patients were randomized to three different groups: treated with empagliflozin 10 mg, treated with empagliflozin 25 mg, or treated with placebo (double blind) superimposed upon the standard of care.

A total of 7 020 participants entered the study.

The primary composite outcome was death from cardiovascular causes, nonfatal myocardial infarction, or nonfatal stroke. A hazard ratio in the empagliflozin group of 0.86 was obtained with a significant 95% confidence interval, ranging from 0.74 to 0.99.

More details about EMPA-REG can be found in Zinman et al 2015 (111) and in Zinman et al 2014 (112).

- TECOS

In the "Trial Evaluating Cardiovascular Outcomes with Sitagliptin" (TECOS), the study group analyzes the long-term effects on cardiovascular events of adding sitagliptin (DPP-4), to usual care in T2D patients with cardiovascular diseases.

TECOS is a randomized, double-blind, placebo controlled study, including 14 671 patients affected by T2D.

The primary cardiovascular outcome was a composite of cardiovascular death, nonfatal myocardial infarction, nonfatal stroke, or hospitalization for unstable angina. A hazard ratio of 0.98 was obtained,

with a non-significant 95% CI (0.88 to 1.09). So, in this trial authors found that adding sitagliptin to usual care in T2D patients with established cardiovascular diseases, did not increase the risk of major adverse cardiovascular events, hospitalization for heart failure, or other adverse cardiovascular events.
More detail about TECOS can be found in Green et al 2015 (113).

- SAVOR-TIMI

The "Saxagliptin Assessment of Vascular Outcomes Recorded in Patients with Diabetes Mellitus (SAVOR)– Thrombolysis in Myocardial Infarction (TIMI)" trial was designed with the main aim of studying the cardiovascular safety and efficacy of saxagliptin (DPP-4) inhibitor. 16 492 T2D patients were randomly assigned to the saxagliptin or to the placebo groups.
Patients recruited in the trial were T2D patients at high risk of cardiovascular events.
The primary end point was a composite of cardiovascular death, myocardial infarction, or ischemic stroke. A hazard ratio of 1.00 was obtained, with 95% confidence interval from 0.89 to 1.12. So the conclusion of SAVOR-TIMI trial was that DPP-4 inhibition with saxagliptin did not change the rate of ischemic events, even if the rate of hospitalization for heart failure was increased.
In conclusion, although saxagliptin improves glycemic control, other drugs are necessary to control the cardiovascular risk associated to T2D.
More details about SAVOR-TIMI trial can be found in (114).

- DECLARE

The "Dapagliflozin Effect on Cardiovascular Event" (DECLARE) trial was designed to study the cardiovascular safety profile of dapagliflozin, a selective inhibitor of sodium–glucose cotransporter 2 (SGLT2i) which promotes glycosuria in T2D patients.
17 160 subjects affected y T2D were included in this trial. The selected patients were at high risk for atherosclerotic cardiovascular disease, and they were randomized to receive either dapagliflozin or placebo, with a ratio of 1:1.
The primary safety outcome was a composite of major adverse cardiovascular events (MACE), defined as cardiovascular death, myocardial infarction, or ischemic stroke.
Dapagliflozin did not result in a lower rate of MACE (hazard ratio, 0.93; 95% CI, 0.84 to 1.03) but did result in a lower rate of cardiovascular death or hospitalization for heart failure (hazard ratio, 0.83; 95% CI, 0.73 to 0.95).

In conclusion, the DECLARE CVOT found that in T2D patients who were at risk for atherosclerotic cardiovascular disease, adding Dapagliflozin to standard care did not result in a higher or lower rate of MACE than placebo, but resulted in a lower rate of cardiovascular death or hospitalization for heart failure.
More details about DECLARE trial can be found in Wiviott et al 2018 (35).

The other studies which were selected (SUSTAIN-6, LEADER, EXCEL, REWIND, PIONEER-6), have already been described in the previous chapter.
Other CVOTs have been excluded due to loss of information. For example, CANVAS trial was excluded because in the publication, the stratum-specific effect estimates were reported without the numbers of patients in each stratum (36). Furthermore, we excluded the HARMONY study because albiglutide has never become clinically available.

Target population

We used DARWIN-T2D data as target real world population, because in Italy GLP-1RA and SGLT2i can be prescribed only by diabetes specialists.
DARWIN-T2D includes patients with T2D, which are an unselected population of adults with T2D which attend Italian diabetes clinics, that represents about 20% of the entire population with T2D in Italy (103, 104).
More details about DARWIN-T2D dataset can be found in the previous chapters and in Fadini et al (10).

Results

The number of variables for whom stratified effect estimates were reported in publications, ranged from a maximum of 28 for EMPA-REG (111) to a minimum of 6 for DECLARE (35). More details on which variables were used to transpose the treatment effects to the target population (DARWIN-T2D) are reported in Table 17. Since different variables were used to the transpositions, we cannot compare results across the different CVOTs taken into account.

Table 17: Post-stratification variables. For each cardiovascular outcome trial, we report which variables were used for post-stratification transposition to the target population. BMI, body mass index. CVD, cardiovascular disease. PAD, peripheral arterial disease. MI, myocardial infarction. eGFR, estimated glomerular filtration rate. DPP-4, dipeptidyl peptidase-4. RAS, renin angiotensin system.

	EMPA-REG	**TECOS**	**SAVOR-TIMI**	**SUSTAIN-6**	**LEADER**	**EXSCEL**	**REWIND**	**PIONEER-6**	**DECLARE**
Duration of diabetes		X	X	X	X	X	X		X
Age	X	X	X	X	X	X	X	X	X
Sex	X	X	X	X	X	X	X	X	
HbA1c	X	X	X	X	X	X	X	X	
BMI	X	X	X	X	X	X	X	X	
Body weight			X						
Systolic blood pressure	X	X							
Diastolic blood pressure	X	X							
Established CVD	X			X	X	X	X	X	
Prior MI or Stroke	X			X			X	X	
PAD	X								
Previous MI	X								X
Heart failure		X	X	X	X	X			X
CVD risk factors	X			X	X			X	X
Only cerebrovascular disease	X								
eGFR	X	X		X	X	X		X	X
Urinary albumin/creatinine ratio	X		X						
Anti-diabetic therapy					X				
Insulin	X	X	X	X		X			
Metfromin	X	X	X						
Sulphonylurea	X	X	X						
Thiazolidinediones	X	X	X						
DPP-4i	X					X			
Anti-hypertive therapy	X		X						
RAS blockers		X	X						
Calcium channel blockers	X	X							
Beta blockers	X	X							
Diuretics	X	X	X						
Aspirin	X								
Statin	X	X	X						
Europe	X	X	X	X	X	X	X		
Ethnicity	X			X	X				
White	X	X	X	X	X	X	X	X	
Number of variables	**28**	**20**	**18**	**14**	**13**	**12**	**9**	**9**	**6**

After excluding from DARWIN-T2D patients with missing data in key variables, we obtained a database composed by 139 726 patients. In Table 18, a comparison between patients in DARWIN-T2D and in the CVOTs (n = 95,816) was reported, from which arises that CVOT population is younger, with a shorter diabetes duration, is more obese, and has a two to threefold greater prevalence of cardiovascular disease, reflected by more frequent use of cardiovascular medications. Among GLMs, patients enrolled in CVOTs had more frequent use of sulphonylurea and insulin. On average, only 41.9% patients enrolled in the selected trials were recruited in Europe and 75.0% were white.

Table 18: Clinical characteristics. Data are presented as mean (SD) for continuous variables or as percentage for categorical variables. The number of patients with available information for each variable is shown for both populations.

	Target population		**CVOTs**	
Variable	**Number**	**Value**	**Number**	**Value**
Duration of diabetes, years	139,700	12.1 (9.4)	71,636	11.7
Age, years	139,708	68.8 (11.2)	95,816	64.4
Sex male, %	139,726	57.1	95,816	66.4
HbA1c, %	132,717	7.3 (1.3)	95,816	8.0
BMI, kg/m^2	126,994	29.6 (5.5)	95,816	31.6
Body weight, kg	128,431	80.8 (17.1)	39,332	89.1
Systolic blood pressure, mm Hg	104,305	137.2 (18.4)	64,572	135.6
Diastolic blood pressure, mm Hg	104,226	77.5 (9.5)	47,412	77.3
Established CVD, %	139,726	28.9	95,816	67.5
PAD, %	139,726	6.0	53,603	14.4
Previous MI, %	97,074	11.7	50,820	38.8
Heart failure, %	139,726	1.4	92,633	13.5
eGFR, ml/min/1.73 m^2	113,593	75.7 (24.5)	83,179	76.7
Albumin creatinine ratio, mg/g	113,775	22.6	41,064	1.4
Glucose-lowering therapy, %	139,726	93.3	95,816	95.0
Insulin, %	130,380	33.5	95,816	39.5
Metformin, %	130,380	71.3	95,816	77.3
Sulphonylurea, %	130,380	27.5	95,816	42.5
Thiazolidinediones, %	130,380	5.0	78,656	4.1
DPP-4 inhibitors, %	130,080	23.3	95,816	16.3
SGLT-2 inhibitors, %	130,080	4.4	95,816	13.8
GLP-1 receptor agonists, %	130,080	5.1	95,816	21.1
Anti-hypertensive therapy, %	117,632	80.1	37,592	92.3
RAS blockers, %	117,632	67.0	92,633	79.2
Calcium channel blockers, %	117,632	25.1	49,080	32.8
Beta blockers, %	117,632	31.5	92,633	56.7
Diuretics, %	117,632	19.2	52,263	41.9
Statin, %	117,632	61.1	95,816	75.3
Aspirin, %	117,632	50.6	95,816	68.3

The substantial difference between the CVOT and the target population was expected and it was analyzed in the previous chapter, and it gives the rationale for performing the transposition analysis to a RW setting.

In Figure 13, we can see results about transposition analyses, comparing results obtained in CVOTs and those obtained after transposition analysis to DARWIN-T2D.

After transposition, the estimated HR showed a protective effect for LEADER (39), SUSTAIN-6 (38), REWIND (92) and DECLARE (35) (co-primary endpoint of cardiovascular death or hospitalization for heart failure).

The HR for 3P-MACE inpatients randomized to empagliflozin in EMPA-REG was 0.86 (95% C.I. 0.74–0.99) and changed to 0.88 (95% C.I. 0.74–1.03) when transposed to DARWIN-T2D.

For each CVOT, subgroup-weighted mean of stratum-specific estimates from CVOTs are given in supplementary tables at the end of the chapter.

The effect on 3p-MACE observed in EXSCEL (37), PIONEER-6 (100) and DECLARE (35) was not significant in the CVOT and remained so after transposition.

As expected, the transposed estimate of DPP-4i effects using stratum-specific data from TECOS (113) or SAVOR-TIMI (114) yielded neutral results also in DARWIN-T2D.

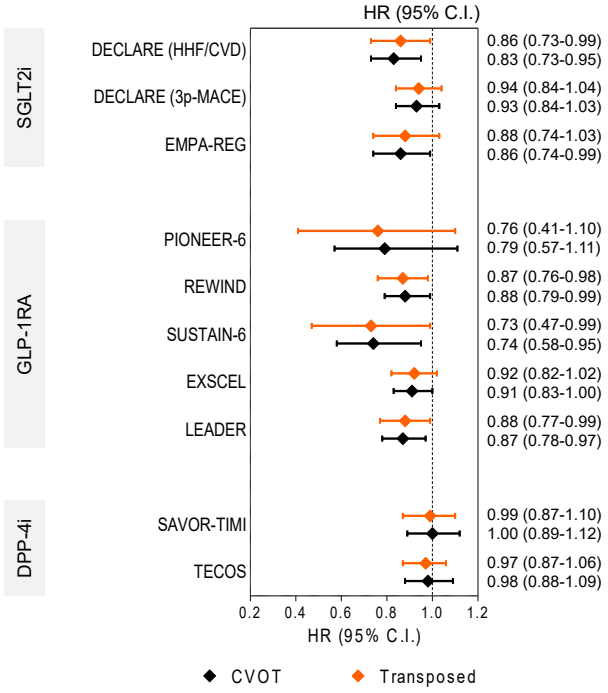


Figure 13: Comparison between observed and transposed effects. The Forest plot reports hazard ratios and 95% confidence intervals (C.I.) for 3-point major adverse cardiovascular events (3P-MACE) and the second co-primary endpoint in DECLARE in the original cardiovascular outcome trial (CVOTs, black) and after transposition to the target population (red). HHF, hospitalization for heat failure; CVD, cardiovascular death. Image extracted from (117)

Discussion

In the previous chapter, we showed that patients in DARWIN-T2D database, which reflect RW patients with T2D that are seen in routine clinical practice, significantly differ from populations included in CVOTs. The same conclusion has been reached in many other publications, even if in these works the authors only used the I/E criteria (88, 95, 96). We developed instead a sample procedure to detect the largest subgroup in DARWIN-T2D of patients with clinical characteristics in average similar to patients enrolled in CVOTs, and even smaller proportions were detected.

Despite this, the transposition analysis that we performed in this chapter, showed that most of the significant results obtained in CVOTs are applicable to the RW T2D population (DARWIN-T2D).

The doubts about transportability to the general T2D population of the protective effect of some GLM drugs arise because CVOTs were conducted on very selected patients, with high risk of developing cardiovascular events (105). However, stratified analyses performed in CVOTs showed that there are not significantly effect's modifiers, suggesting a clinical transferability of the findings. Nonetheless, several trends of interaction and a few nominally significant interactions between the active treatment and stratification variables may yield overall significant effects when transposed to a much different target population.

Prior of this work, there is none attempt to transpose results obtained in CVOTs to a RW population, with quantitative estimates.

The post-stratification transposition approach that we applied is in general used when individual-level data are available for CVOT, and aggregated-level data are available for the target population (109). However, we applied this approach to individual-level data in the RW population, and aggregated-data for the CVOTs. The gold standard approach presented in (109), requires individual-level data from both CVOTs and the RW population to compute probabilities of being sampled in the trial and to reweight trial participants to reflect the target population of patient characteristics. However, accessing individual data of multiple CVOTs sponsored by different companies can be harnessed by conflicts of interest. Alternative methods, as the one we used, are subjected to biases and based on some critical assumptions. Specifically, this approach requires only categorical variables and it is effective only when a small number of variables are taken into account (108). Furthermore, conditional dependencies among variables are not considered, and only one variable at a time is taken into account, making the strong assumption of no correlation between them, which is not realistic.

Another limitation of this approach is that transposed results are heavily influenced by proportions of the target population in each stratum (109), but the high percentage of missing data in DARWIN-T2D could lead to biased results.

In general, our results confirm the superiority of active GLM drugs versus placebo for cardiovascular protection obtained in CVOTs, even when transposed to a RW population with different characteristics. This was true for LEADER, SUSTAIN-6, REWIND and DECLARE.

Results of EMPA-REG were instead not confirmed. In fact, results lost the statistical significance after transposition, probably due to the presence of heterogeneity observed in subgroups of patients stratified by age and baseline HbA1c (111), the 2:1 ratio between patients randomized to active treatment (empagliflozin) and those randomized to the placebo group, which yields small numbers of patients in some strata, and the large number of variables (n=28) that composed strata used for transposition. In fact, as sensitivity analysis, we transposed the EMPA-REG result with the same 6 variables which were used to transpose DECLARE. In this case, we obtained a HR of 0.85 (95% C.I. 0.70–0.99) for the RW population, which is still significant. However, the fact that fully transposed HR for EMPA-REG lost the statistical significance does not imply that EMPA-REG results are less generalizable to the target population than other CVOT's, because the observed and transposed HRs were however quite similar.

It is important to notice that we transposed the CVOT drug's effects as if all T2D patients included in DARWIN-T2D could receive that medication. We did not apply CVOT I/E criteria, because our aim was to estimate the treatment effect in an unselected target population.

In conclusion, despite the limitations described above, we provide the first quantitative estimate about the cardiovascular protection by diabetes drugs investigated in CVOTs, and we showed that they could be applied to a very different and highly heterogeneous population of patients with T2D seen in routine care.

This chapter has been published as
Transposition of cardiovascular outcome trial effects to the real-world population of patients with type 2 diabetes. Sciannameo V, Berchialla P, Avogaro A, Fadini GP; DARWIN-T2D Network. Cardiovasc Diabetol (2021) 20:103 https://doi.org/10.1186/s12933-021-01300-y

Supplementary material

Supplementary Table 2: REWIND. CVOT cardiovascular outcome trial, CVD cardiovascular disease, BMI Body Mass Index, MI myocardial infarction, HR hazard ratio, low CI and high CI refer to the low and high 95% confidence interval (CI) limit, respectively.

Characteristics	**CVOT estimates**			**Proportions**		**Sub-group weighted mean of stratum-specific CVOT estimates**		
	HR	*low CI*	*high CI*	*Darwin*	*CVOT*	*HR*	*low CI*	*high*
Age < 66 yrs	0.92	0.78	1.09	0.34	0.53	0.88	0.77	0.99
Age ≥ 66 yrs	0.86	0.74	1.00	0.65	0.47			
Sex: Female	0.85	0.71	1.02	0.44	0.46	0.88	0.76	0.99
Sex: Male	0.90	0.79	1.04	0.56	0.54			
Duration diabetes < 5 yrs	0.84	0.66	1.06	0.27	0.24	0.88	0.77	0.99
Duration diabetes in [5,10]	0.89	0.73	1.09	0.22	0.30			
Duration diabetes > 10 yrs	0.90	0.77	1.06	0.51	0.46			
CVD: Yes	0.87	0.74	1.02	0.20	0.31	0.87	0.75	0.99
CVD: No	0.87	0.74	1.02	0.80	0.63			
Hba1c < 7.2 %	0.90	0.76	1.06	0.29	0.47	0.87	0.76	0.98
Hba1c ≥ 7.2 %	0.86	0.74	1.00	0.65	0.53			
BMI < 32 kg/m^2	0.94	0.81	1.09	0.34	0.54	0.90	0.80	1.01
BMI ≥ 32 kg/m^2	0.82	0.69	0.96	0.14	0.46			
Region: Europe	0.77	0.65	0.90	1.00	0.44	0.77	0.67	0.87
MI or Stroke: Yes	0.79	0.66	0.96	0.09	0.21	0.92	0.80	1.04
MI or Stroke: No	0.93	0.81	1.07	0.91	0.79			
Race: White	0.90	0.79	1.02	1.00	0.76	0.90	0.79	1.01
CVOT estimate 0.88 (0.79-0.99)						Transposed: 0.87 (0.76-0.98)		

Supplementary Table 3: SUSTAIN-6. CVOT cardiovascular outcome trial, CVD cardiovascular disease, BMI Body Mass Index, MI myocardial infarction, HR hazard ratio, low CI and high CI refer to the low and high 95% confidence interval (CI) limit, respectively.

Characteristics	**CVOT estimates**			**Proportions**		**Sub-group weighted mean of stratum-specific CVOT estimates**		
	HR	*low CI*	*high CI*	*Darwin*	*CVOT*	*HR*	*low CI*	*high CI*
Sex: Female	0.84	0.54	1.31	0.44	0.39	0.75	0.49	1.00
Sex: Male	0.68	0.50	0.92	0.56	0.61			
Age < 65 yrs	0.74	0.52	1.05	0.32	0.39	0.73	0.49	0.96
Age ≥ 65 yrs	0.72	0.51	1.02	0.67	0.48			
BMI < 30 kg/m^2	0.58	0.39	0.87	0.29	0.36	0.67	0.42	0.93
BMI ≥ 30 kg/m^2	0.84	0.61	1.16	0.20	0.64			
Hba1c ≤ 8.5 %	0.72	0.50	1.03	0.45	0.56	0.72	0.47	0.97
Hba1c > 8.5 %	0.74	0.52	1.04	0.07	0.44			
Duration diabetes ≤ 10 yrs	0.73	0.48	1.12	0.54	0.35	0.73	0.48	0.98
Duration diabetes > 10 yrs	0.73	0.54	0.99	0.47	0.65			
Egfr < 60 ml/min/1.73 m^2	0.84	0.57	1.25	0.12	0.28	0.72	0.46	0.96
Egfr ≥ 60 ml/min/1.73 m^2	0.67	0.48	0.92	0.33	0.72			
Insulin: No	0.52	0.33	0.81	0.31	0.42	0.65	0.38	0.93
Insulin: Yes	1.02	0.64	1.62	0.16	0.32			
CVD: Yes	0.72	0.55	0.93	0.20	0.83	0.72	0.49	0.95
CVD risk factors: Yes	1.00	0.41	2.46	0.10	0.17	1.00	0.63	1.37
Heart Failure: No	0.64	0.48	0.86	0.98	0.83	0.64	0.39	0.90
Heart Failure: Yes	1.03	0.64	1.66	0.02	0.17			
MI or stroke = No	0.70	0.47	1.04	0.91	0.59	0.70	0.44	0.97
MI or stroke = Yes	0.76	0.55	1.05	0.09	0.41			
Region; Europe	0.62	0.34	1.13	1.00	0.19	0.62	0.36	0.88
Race: White	0.76	0.58	1.00	1.00	0.83	0.76	0.51	1.01
Ethnicity: Not Hispanic or Latinos	0.74	0.57	0.96	1.00	0.85	0.74	0.50	0.98
	CVOT estimate: 0.74 (0.58-0.95)					Transposed: 0.73 (0.47-0.99)		

Supplementary Table 4: DECLARE HHF/CVD. HHF hospitalization for heat failure, CVOT cardiovascular outcome trial, MI myocardial infarction, HR hazard ratio, low CI and high CI refer to the low and high 95% confidence interval (CI) limit, respectively.

Characteristics	**CVOT estimates**			**Proportions**		**Sub-group weighted mean of stratum-specific CVOT estimates**		
	HR	*low CI*	*high CI*	*Darwin*	*CVOT*	*HR*	*low CI*	*high*
Heart Failure: Yes	0.79	0.63	0.99	0.02	0.10	0.84	0.64	1.04
Heart Failure: No	0.84	0.72	0.99	0.98	0.90			
Age < 65 yrs	0.88	0.72	1.07	0.32	0.54	0.86	0.69	1.04
Age in [65; 75] yrs	0.77	0.63	0.94	0.31	0.40			
Age ≥ 75 yrs	0.94	0.65	1.36	0.36	0.06			
Egfr < 60 ml/min/1.73 m²	0.78	0.55	1.09	0.12	0.07	0.83	0.66	1.01
Egfr in [60; 90) ml/min/1.73	0.79	0.66	0.95	0.20	0.45			
Egfr ≥ 90 ml/min/1.73 m²	0.96	0.77	1.19	0.13	0.48			
Duration diabetes < 5	1.08	0.87	1.35	0.32	0.22	0.94	0.81	1.07
Duration diabetes in [5;10)	1.02	0.83	1.25	0.22	0.28			
Duration diabetes in [10; 15)	0.94	0.77	1.15	0.17	0.23			
Duration diabetes in [15; 20)	0.92	0.71	1.18	0.12	0.14			
Duration diabetes ≥20 yrs	0.67	0.52	0.86	0.18	0.13			
Previous MI: No	0.85	0.72	1.00	0.56	0.79	0.85	0.67	1.02
Previous MI: Yes	0.81	0.65	1.00	0.07	0.21			
	CVOT estimate: 0.83 (0.73-0.95)					Transposed: 0.86 (0.73-0.99)		

Supplementary Table 5: DECLARE MACE. MACE 3-point major adverse cardiovascular events, CVOT cardiovascular outcome trial, MI myocardial infarction, HR hazard ratio, low CI and high CI refer to the low and high 95% confidence interval (CI) limit, respectively.

Characteristics	**CVOT estimates**			**Proportions**		**Sub-group weighted mean of stratum-specific CVOT estimates**		
	HR	*low CI*	*high CI*	*Darwin*	*CVOT*	*HR*	*low CI*	*hig*
Heart Failure: Yes	1.01	0.81	1.27	0.02	0.10	0.92	0.80	1.0
Hear Failure: No	0.92	0.82	1.02	0.98	0.90			
Age < 65 yrs	0.93	0.81	1.08	0.32	0.54	0.91	0.77	1.0
Age in [65; 75) yrs	0.97	0.83	1.13	0.31	0.40			
Age ≥ 75 yrs	0.84	0.61	1.15	0.36	0.06			
Egfr < 60 ml/min/1.73 m²	0.92	0.69	1.23	0.12	0.07	0.94	0.81	1.0
Egfr in [60; 90) ml/min/1.73 m²	0.95	0.82	1.09	0.20	0.45			
Egfr ≥ 90 ml/min/1.73 m²	0.94	0.80	1.10	0.13	0.48			
Duration diabetes < 5 yrs	1.08	0.87	1.35	0.32	0.22	0.94	0.81	1.0
Duration diabetes in [5;10) yrs	1.02	0.83	1.25	0.22	0.28			
Duration diabetes in [10; 15)	0.94	0.77	1.15	0.17	0.23			
Duration diabetes in [15;20) yrs	0.92	0.71	1.18	0.12	0.14			
Duration diabetes ≥ 20 yrs	0.67	0.52	0.86	0.18	0.13			
Previous MI: No	1.00	0.88	1.13	0.56	0.79	0.98	0.85	1.1
Previous MI: Yes	0.84	0.72	0.99	0.07	0.21			
	CVOT estimate : 0.93 (0.84-1.03)					Transposed: 0.94 (0.84-1.04)		

Supplementary Table 6: EMPA-REG. CVOT cardiovascular outcome trial, CVD cardiovascular disease, BMI Body Mass Index, MI myocardial infarction, SBP systolic Blood Pressure, DBP Diastolic Blood Pressure, HR hazard ratio, low CI and high CI refer to the low and high 95% confidence interval (CI) limit, respectively.

Characteristics	**CVOT estimates**			**Proportions**		**Sub-group weighted mean of stratum-specific CVOT estimates**		
	HR	*low*	*high CI*	*Darwin*	*CV*	*HR*	*low CI*	*high*
Age < 65 yrs	1.04	0.84	1.29	0.32	0.5	0.80	0.68	0.93
Age ≥ 65 yrs	0.71	0.59	0.87	0.67	0.4			
Angiotensing converting: No	0.77	0.56	1.07	0.14	0.1	0.84	0.71	0.97
Angiotensing converting: Yes	0.88	0.75	1.04	0.27	0.8			
Antihypertensive: No	0.94	0.45	1.95	0.08	0.0	0.87	0.74	0.99
Antihypertensive: Yes	0.85	0.73	0.99	0.34	0.9			
Acetylsalicyclic acid: No	0.80	0.57	1.12	0.22	0.1	0.83	0.71	0.95
Acetylsalicyclic acid: Yes	0.87	0.74	1.02	0.20	0.8			
Beta blockers: No	0.90	0.70	1.17	0.29	0.3	0.88	0.75	1.01
Beta blockers: Yes	0.83	0.70	1.00	0.13	0.6			
BMI < 30 kg/m^2	0.74	0.60	0.91	0.28	0.4	0.83	0.71	0.96
BMI ≥ 30 kg/m^2	0.98	0.80	1.21	0.20	0.5			
Calcium channel blockers : No	0.87	0.73	1.05	0.31	0.6	0.86	0.73	0.99
Calcium channel blockers: Yes	0.83	0.65	1.06	0.10	0.3			
Cerebrovascular disease	1.15	0.74	1.78	0.02	0.1	1.15	1.01	1.29
CVD risk factors	0.79	0.61	1.04	0.10	0.1	0.79	0.69	0.89
Diuretics: No	0.83	0.67	1.02	0.34	0.5	0.84	0.72	0.96
Diuretics: Yes	0.88	0.71	1.07	0.08	0.4			
Dpp4: No	0.81	0.70	0.95	0.36	0.8	0.90	0.77	1.03
Dpp4: Yes	1.27	0.81	1.98	0.11	0.1			
Egfr < 60 ml/min/1.73 m^2	0.88	0.69	1.13	0.12	0.2	0.88	0.75	1.00
Egfr in [60; 90) ml/min/1.73	0.76	0.61	0.94	0.20	0.5			
Egfr ≥ 90 ml/min/1.73 m^2	1.10	0.77	1.57	0.13	0.2			
Ethnicity: Not	0.91	0.77	1.07	1.00	0.8	0.91	0.79	1.03
Region: Europe	1.02	0.81	1.28	1.00	0.4	1.02	0.90	1.14
Hba1c < 8.5 %	0.76	0.64	0.90	0.44	0.6	0.81	0.69	0.93
Hba1c ≥ 8.5 %	1.14	0.86	1.50	0.08	0.3			
Insulin: No	0.79	0.64	0.97	0.31	0.5	0.83	0.71	0.95
Insulin: Yes	0.93	0.75	1.13	0.16	0.4			
Metformin: No	0.72	0.56	0.94	0.13	0.2	0.86	0.73	0.98
Metformin: Yes	0.92	0.77	1.10	0.33	0.7			
MI or stroke: No	0.88	0.66	1.18	0.91	0.3	0.88	0.75	1.01
MI or stroke: Yes	0.84	0.71	1.00	0.09	0.6			
peripheral artery disease	0.94	0.47	1.88	0.06	0.0	0.94	0.77	1.11
SBP ≥140 mmHg and/or	0.83	0.66	1.03	0.20	0.3	0.83	0.72	0.94
SBP <140 mmHg and	0.89	0.73	1.08	0.20	0.6	0.89	0.76	1.02
Sex: Male	0.87	0.73	1.02	0.56	0.7	0.85	0.73	0.97
Sex: Female	0.83	0.62	1.11	0.44	0.2			
Statins: No	0.79	0.59	1.07	0.16	0.2	0.84	072	0.97

Statins: Yes	0.88	0.74	1.04	0.26	0.7			
Sulfonylurea: No	0.85	0.70	1.02	0.34	0.5	0.86	0.73	0.98
sulfonylurea : Yes	0.87	0.69	1.11	0.13	0.4			
Thiazolidinediones: No	0.85	0.73	0.98	0.44	0.9	0.86	0.74	0.98
Thiazolidinediones: Yes	1.13	0.55	2.31	0.02	0.0			
Albumin creatinine ratio < 30	0.89	0.72	1.10	0.29	0.5	0.89	0.76	1.02
Albumin creatinine ratio 30-	0.89	0.69	1.16	0.16	0.2			
Albumin creatinine ratio >300	0.69	0.49	0.96	0.00	0.1			
Race: White	0.88	0.74	1.04	1.00	0.7	0.88	0.76	1.00
	CVOT estimate : 0.86 (0.74-0.99)					Transposed: 0.88 (0.74-1.03)		

Supplementary Table 7: LEADER. CVOT cardiovascular outcome trial, CVD cardiovascular disease, BMI Body Mass Index, HR hazard ratio, low CI and high CI refer to the low and high 95% confidence interval (CI) limit, respectively.

Characteristics	**CVOT estimates**			**Proportions**		**Sub-group weighted mean of stratum-specific CVOT estimates**		
	HR	*low CI*	*high CI*	*Darwin*	*CVO*	*HR*	*low CI*	*high CI*
Sex: Female	0.88	0.72	1.08	0.44	0.36	0.87	0.76	0.98
Sex: Male	0.86	0.75	0.98	0.56	0.64			
Age < 60 yrs	0.78	0.62	0.97	0.20	0.25	0.87	0.77	0.98
Age ≥ 60 yrs	0.90	0.79	1.02	0.79	0.75			
BMI < 30 kg/m^2	0.96	0.81	1.15	0.29	0.38	0.90	0.79	1.01
BMI ≥ 30 kg/m^2	0.82	0.71	0.94	0.20	0.62			
Hba1c ≤ 8.3 %	0.89	0.76	1.05	0.43	0.51	0.88	0.77	0.99
Hba1c > 8.3 %	0.84	0.72	0.98	0.09	0.49			
Duration diabetes ≤ 11 yrs	0.82	0.70	0.97	0.58	0.47	0.85	0.75	0.96
Duration diabetes > 11 yrs	0.90	0.78	1.04	0.42	0.52			
CVD: Yes	0.83	0.74	0.93	0.20	0.81	0.83	0.73	0.93
CVD Risk factors: Yes	1.20	0.86	1.67	0.10	0.19	1.20	1.06	1.34
Egfr < 60 ml/min/1.73 m^2	0.69	0.57	0.85	0.12	0.23	0.86	0.76	0.98
Egfr ≥ 60 ml/min/1.73 m^2	0.94	0.83	1.07	0.33	0.77			
Heart Failure: No	0.85	0.76	0.96	0.98	0.86	0.85	0.74	0.96
Heart Failure: Yes	0.94	0.72	1.21	0.02	0.14			
Antidiabetic therapy: 1 oral	0.75	0.58	0.98	0.17	0.19	0.83	0.73	0.94
Antidiabetic therapy: more than	0.95	0.78	1.16	0.14	0.32			
Antidiabetic therapy: Insulin	0.89	0.74	1.06	0.09	0.37			
Antidiabetic therapy: Insulin	0.86	0.63	1.17	0.07	0.08			
Antidiabetic therapy: None	0.73	0.42	1.25	0.04	0.04			
Region: Europe	0.82	0.68	0.98	1.00	0.35	0.82	0.72	0.92
Race: White	0.90	0.80	1.02	1.00	0.77	0.90	0.79	1.01
Ethnicity: Non-Hispanic	0.89	0.79	1.00	1.00	0.88	0.89	0.78	1.00
	CVOT estimate : 0.87 (0.78-0.97)					Transposed: 0.88 (0.77-0.99)		

Supplementary Table 8: PIONEER-6. CVOT cardiovascular outcome trial, CVD cardiovascular disease, BMI Body Mass Index, MI myocardial infarction, HR hazard ratio, low CI and high CI refer to the low and high 95% confidence interval (CI) limit, respectively.

Characteristics	**CVOT estimates**			**Proportions**		**Sub-group weighted mean of stratum-specific CVOT estimates**		
	HR	*low CI*	*high CI*	*Darwin*	*CVOT*	*HR*	*low CI*	*high CI*
Age < 65 yrs	0.51	0.29	0.90	0.26	0.42	0.85	0.66	1.05
Age ≥ 65 yrs	1.04	0.68	1.59	0.67	0.58			
Egfr < 60 ml/min/1.73 m^2	0.74	0.41	1.33	0.12	0.27	0.79	0.60	0.98
Egfr ≥ 60 ml/min/1.73 m^2	0.81	0.54	1.22	0.33	0.73			
Sex: Female	1.16	0.54	2.51	0.44	0.32	0.89	0.69	1.09
Sex: Male	0.72	0.50	1.05	0.56	0.68			
CVD: Yes	0.83	0.58	1.17	0.20	0.85	0.83	0.65	1.01
CVD risk factors: Yes	0.51	0.15	1.68	0.10	0.15	0.51	0.25	0.77
Hba1c < 8.6 %	0.81	0.53	1.24	0.45	0.67	0.80	0.61	0.99
Hba1c ≥ 8.6 %	0.73	0.42	1.26	0.07	0.32			
BMI < 31 kg/m^2	0.61	0.36	1.03	0.32	0.40	0.71	0.52	0.90
BMI ≥ 31 kg/m^2	0.95	0.61	1.48	0.16	0.60			
Race: White	0.83	0.56	1.23	1.00	0.72	0.83	0.64	1.02
MI or stroke: Yes	0.97	0.64	1.49	0.09	0.45	0.62	0.41	0.82
MI or stroke: No	0.59	0.34	1.03	0.91	0.54			
	CVOT estimate : 0.79 (0.57-1.11)					Transposed: 0.76 (0.41-1.10)		

Supplementary Table 9: TECOS. CVOT cardiovascular outcome trial, CVD cardiovascular disease, BMI Body Mass Index, MI myocardial infarction, SBP Systolic Blood Pressure, DBP Diastolic Blood Pressure, HR hazard ratio, low CI and high CI refer to the low and high 95% confidence interval (CI) limit, respectively.

Characteristics	**CVOT estimates**			**Proportions**		**Sub-group weighted mean of stratum-specific CVOT estimates**		
	HR	*low CI*	*high CI*	*Darwin*	*CVOT*	*HR*	*low CI*	*high CI*
Age < 65 yrs	0.95	0.82	1.11	0.32	0.45	0.99	0.87	1.11
Age ≥ 65 yrs	1.01	0.90	1.15	0.67	0.53			
Sex: Male	0.99	0.88	1.10	0.56	0.71	0.97	0.86	1.09
Sex: Female	0.95	0.78	1.15	0.44	0.29			
Race: White	0.97	0.87	1.08	1.00	0.68	0.97	0.86	1.08
Region: Europe	0.95	0.73	1.23	1.00	0.14	0.95	0.83	1.07
Duration diabetes < 5 yrs	0.99	0.78	1.26	0.27	0.19	0.99	0.87	1.11
Duration diabetes is [5; 15)	0.89	0.78	1.02	0.40	0.51			
Duration diabetes ≥ 15 yrs	1.12	0.95	1.32	0.32	0.29			
Sulfonylurea: Yes	0.99	0.86	1.14	0.13	0.45	0.98	0.86	1.09
Sulfonylurea: No	0.97	0.85	1.10	0.34	0.55			
Metformin: Yes	0.96	0.83	1.04	0.33	0.82	1.01	0.91	1.10
Metformin: No	1.13	0.93	1.38	0.13	0.18			
Thiazolidinedione: Yes	0.86	0.49	1.49	0.02	0.03	0.97	0.86	1.09
Thiazolidinedione: No	0.98	0.89	1.08	0.44	0.97			
Insulin: Yes	1.01	0.85	1.21	0.16	0.23	0.98	0.86	1.10
Insulin: No	0.96	0.89	1.08	0.31	0.77			
Heart Failure: Yes	0.97	0.80	1.17	0.02	0.18	0.99	0.88	1.10
Heart Failure: No	0.99	0.88	1.10	0.98	0.82			
Hba1c < 7.2 %	0.95	0.83	1.09	0.29	0.52	0.97	0.86	1.09
Hba1c ≥ 7.2 %	1.00	0.88	1.14	0.23	0.48			
Egfr < 60 ml/min/1.73 m^2	0.92	0.78	1.10	0.12	0.23	0.98	0.86	1.10
Egfr ≥ 60 ml/min/1.73 m^2	1.00	0.89	1.12	0.33	0.76			
SBP < 140 mmHg	0.96	0.85	1.09	0.20	0.60	0.98	0.86	1.10
SBP in [140; 160) mmHg	1.03	0.87	1.23	0.13	0.31			
SBP ≥ 160 mmHg	0.92	0.70	1.23	0.06	0.09			
DBP <90 mmHg	0.98	0.88	1.09	0.34	0.85	0.97	0.85	1.09
DBP in [90; 100) mmHg	1.08	0.84	1.40	0.04	0.13			
DBP ≥ 100 mmHg	0.51	0.25	1.02	0.01	0.02			
BMI < 30 kg/m^2	1.08	0.95	1.24	0.29	0.53	0.99	0.88	1.11
BMI ≥ 30	0.88	0.76	1.01	0.20	0.46			
Statins: Yes	0.98	0.88	1.10	0.26	0.80	0.97	0.85	1.09
Statins: No	0.96	0.79	1.16	0.16	0.20			
ACE inhibitors: Yes	1.00	0.90	1.11	0.27	0.79	0.96	0.85	1.08
ACE inhibitors: No	0.89	0.71	1.11	0.14	0.21			
Diuretics: Yes	0.96	0.84	1.09	0.08	0.41	1.00	0.89	1.11
Diuretics: No	1.01	0.88	1.15	0.34	0.59			
Calcium channel blockers: Yes	0.93	0.79	1.09	0.10	0.34	0.99	0.88	1.10
Calcium channel blockers: No	1.01	0.89	1.13	0.31	0.66			
Beta blockers: Yes	0.96	0.85	1.07	0.13	0.64	1.01	0.90	1.13
Beta blockers: No	1.04	0.87	1.23	0.29	0.36			
	CVOT estimate: 0.98 (0.88-1.09)					Transposed: 0.97 (0.87-1.06)		

Supplementary Table 10: SAVOR-TIMI. CVOT cardiovascular outcome trial, BMI Body Mass Index, HR hazard ratio, Low CI and high CI refer to the low and high 95% confidence interval (CI) limit, respectively.

Characteristics	CVOT estimates			Proportions		Sub-group weighted mean of stratum-specific CVOT estimates		
	HR	*low CI*	*high CI*	*Darwin*	*CVOT*	*HR*	*low*	*high CI*
Egfr < 30 ml/min/1.73 m²	0.83	0.49	1.39	0.02	0.02	1.00	0.85	1.15
Egfr in [30;50) ml/min/1.73 m²	1.02	0.79	1.30	0.05	0.14			
Egfr ≥ 50 ml/min/1.73 m²	1.01	0.88	1.15	0.37	0.84			
Sex: Male	1.01	0.89	1.16	0.56	0.67	0.99	0.84	1.14
Sex: Female	0.97	0.78	1.20	0.44	0.33			
Race: White	0.98	0.86	1.11	1.00	0.75	0.98	0.84	1.12
Age < 75 yrs	1.01	0.89	1.15	0.63	0.86	0.99	0.84	1.14
Age ≥ 75 yrs	0.96	0.75	1.22	0.36	0.14			
Region: Europe	0.96	0.81	1.13	1.00	0.42	0.96	0.82	1.10
BMI < 30 kg/m²	1.01	0.86	1.19	0.28	0.46	1.00	0.85	1.15
BMI ≥ 30 kg/m²	0.99	0.85	1.16	0.20	0.53			
Heart Failure: Yes	1.13	0.89	1.43	0.02	0.13	0.97	0.83	1.12
2Heart Failure: No	0.97	0.85	1.10	0.98	0.87			
Duration diabetes < 5 yrs	1.07	0.82	1.40	0.27	0.24	1.01	0.86	1.16
Duration diabetes in [5; 10) yrs	1.04	0.81	1.33	0.22	0.24			
Duration diabetes in [10;15) yrs	0.94	0.74	1.19	0.18	0.21			
Duration diabetes in [15;20) yrs	1.06	0.79	1.41	0.12	0.13			
Duration diabetes ≥20 yrs	0.93	0.74	1.17	0.20	0.18			
Hba1c < 7 %	1.01	0.78	1.31	0.24	0.25	1.01	0.86	1.15
Hba1c in[7; 8) %	0.98	0.80	1.20	0.16	0.33			
Hba1c in [8; 9) %	1.09	0.85	1.39	0.07	0.19			
Hba1c ≥ 9 %	0.95	0.77	1.18	0.05	0.21			
Insulin: Yes	1.03	0.88	1.20	0.16	0.41	0.98	0.84	1.13
Insulin: No	0.96	0.82	1.13	0.31	0.59			
Sulfonylurea: Yes	0.95	0.79	1.14	0.13	0.40	1.01	0.86	1.15
Sulfonylurea: No	1.03	0.90	1.19	0.34	0.60			
Metformin: Yes	0.97	0.84	1.13	0.33	0.70	0.99	0.84	1.15
Metformin: No	1.05	0.88	1.25	0.13	0.30			
Thiazolidinedione: Yes	0.59	0.33	1.04	0.02	0.06	0.99	0.84	1.14
Thiazolidinedione: No	1.02	0.91	1.15	0.44	0.94			
Micro-albumin creatinine ratio < 30 mg/g	1.07	0.90	1.27	0.29	0.59	1.01	0.85	1.16
Micro-albumin creatinine ratio in [30;	0.90	0.74	1.09	0.16	0.27			
Micro-albumin creatinine ratio ≥ 300	0.88	0.68	1.13	0.00	0.10			
Ethnicity: Not hispanic	0.97	0.86	1.10	1.00	0.79	0.97	0.83	1.11
Weight < 80 Kg	1.10	0.91	1.33	0.25	0.36	1.05	0.88	1.17
Weight ≥ 80 Kg	0.95	0.83	1.09	0.24	0.64			
Hypertension: Yes	0.97	0.86	1.10	0.34	0.82	1.00	0.85	1.15
Hypertension: No	1.14	0.87	1.51	0.08	0.18			
Statins: Yes	0.99	0.87	1.12	0.26	0.78	1.01	0.87	1.15
Statins: No	1.04	0.80	1.34	0.16	0.22			
ACEi/ARB: Yes	0.98	0.86	1.11	0.27	0.79	1.01	0.87	1.16
ACEi/ARB: No	1.08	0.85	1.38	0.14	0.21			
Diuretics: Yes	1.02	0.88	1.18	0.08	0.44	0.99	0.84	1.14
Diuretics: No	0.98	0.82	1.16	0.34	0.56			
	CVOT estimate : 1.00 (0.89-1.12)					Transposed: 0.99 (0.87-1.10)		

Supplementary Table 11: EXCEL. CVOT cardiovascular outcome trial, CVD cardiovascular disease, BMI Body Mass Index, HR hazard ratio, low CI and high CI refer to the low and high 95% confidence interval (CI) limit, respectively.

Characteristics	**CVOT estimates**			**Proportions**		**Sub-group weighted mean of stratum-specific CVOT estimates**		
	HR	*low CI*	*high CI*	*Darwin*	*CVOT*	*HR*	*low CI*	*high CI*
Age < 65 yrs	1.05	0.92	1.21	0.32	0.60	0.87	0.75	0.99
Age ≥ 65 yrs	0.80	0.71	0.91	0.67	0.40			
Sex: Male	0.94	0.84	1.05	0.56	0.62	0.90	0.79	1.02
Sex: Female	0.86	0.73	1.03	0.44	0.38			
Race: White	0.95	0.85	1.05	1.00	0.76	0.95	0.84	1.06
Region: Europe	1.00	0.87	1.15	1.00	0.46	1.00	0.88	1.12
Duration diabetes < 5 yrs	0.70	0.50	0.97	0.27	0.14	0.87	0.75	0.99
Duration diabetes in [5;15) yrs	0.98	0.85	1.12	0.40	0.49			
Duration diabetes ≥ 15 yrs	0.90	0.79	1.04	0.32	0.37			
Anti-hyperglycemic oral agent	0.93	0.84	1.04	0.39	0.85	0.85	0.75	1.00
Anti-hyperglycemic oral agent	0.84	0.69	1.03	0.61	0.15			
Insulin: Yes	0.89	0.78	1.00	0.16	0.46	0.93	0.81	1.05
Insulin: No	0.95	0.83	1.10	0.31	0.54			
Dpp4: Yes	1.08	0.84	1.39	0.11	0.15	0.93	0.81	1.05
Dpp4: No	0.89	0.80	0.99	0.36	0.85			
Heart Failure: Yes	0.97	0.81	1.16	0.02	0.16	0.90	0.79	1.01
Heart Failure: No	0.90	0.81	1.00	0.98	0.84			
Hba1c < 8 %	0.91	0.80	1.05	0.40	0.49	0.91	0.79	1.03
Hba1c ≥ 8%	0.91	0.80	1.04	0.12	0.51			
Egfr < 60 ml/min/1.73 m^2	1.01	0.86	1.19	0.12	0.22	0.90	0.77	1.02
Egfr ≥ 60 ml/min/1.73 m^2	0.86	0.77	0.97	0.33	0.78			
BMI < 30 kg/m^2	0.94	0.79	1.10	0.28	0.36	0.92	0.81	1.03
BMI ≥ 30 kg/m^2	0.89	0.79	1.00	0.20	0.63			
CVD: Yes	0.90	0.82	1.00	0.20	0.73	0.97	0.84	1.10
CVD: No	0.99	0.77	1.28	0.80	0.27			
	CVOT estimate: 0.92 (0.82-1.02)					Transposed: 0.91 (0.83-1.00)		

CHAPTER 6

DEEP LEARNING FOR PREDICTING URGENT HOSPITALIZATIONS IN ELDERLY POPULATION USING ADMINISTRATIVE ELECTRONIC HEALTH RECORDS

Introduction

The World Health Organization (WHO) pointed out, in a report published in 2011 (118), that the population with more than 65 years is constantly growing from an estimated 524 million in 2010 to nearly 1.5 billion in 2050. This could heavily affect healthcare system and increasing social costs in the future. Often, elderly people are simultaneously affected by at least two chronic morbidities (multi-morbidity), which means that lots of different drugs are used (poly-pharmacy) by the same individual in the same time, that increases the complexity of managing such kind of patients. Furthermore, it is well known from literature that multi-morbidity and poly-pharmacy are risk factors for urgent hospitalizations and worsening of the quality of life of elderly people (119–121). Consequently, it is a priority to prevent adverse outcomes by early warnings, understanding the pattern of health trajectories over time, defined as the dynamic course of the health status of an individual described as a succession of healthcare events, like medication prescriptions, diagnoses registered during hospitalizations, and so on.

Healthcare trajectories of elderly population can be reconstructed from Healthcare Administrative Databases (HADs), as described in Chapter 1.

Even if HADs were originally born for administrative purposes, in the last decades they begun to be used to do epidemiological and medical researches too. The greatest part of the studies conducted on HADs, perform analyses based on traditional regression-based approaches, such as logistic regression (122), support vector machines (123) or random forest (124). However, these approaches are not suitable when high dimensional data are available, they are not able to manage irregular time intervals between events and they are based on very strong assumptions, often hard to verify (125).

Furthermore, in the last decades a personalized medicine perspective is being developed, and advanced ML techniques help to reach this aim.

ML and Deep Learning (DL) approaches (126–128) are able to learn compact representations of personalized healthcare trajectories of elderly population, taking advantage from the big amount of data available from HADs, such as medication prescriptions and hospitalizations' diagnoses.

For example, Nguyen et colleagues (129) used a Convolutional Neural Network (CNN) to predict the probability of readmission, constructing a DL model called Deepr (Deep record). They reconstructed healthcare trajectories using diagnoses, clinical procedures and medications extracted directly from hospitals' Electronic Health Records (EHRs).

Pham et al (130) used instead the Long-Short Term Memory (LSTM) algorithm (131) to construct the DeepCare model to predict the onset of diabetes, using EHRs, and taking advantage from the LSTM's attention mechanism.

In a similar way, Choi et colleagues used a Recurrent Neural Network (RNN) called RETAIN to predict heart failure from EHRs (132).

Successively, Li et al (133) introduced the use of Transformer architecture, i.e. a natural language processing (NLP) approach, to predict future diagnoses based on the healthcare trajectory defined as a succession of previous diagnoses. The model that was born from this application was named BEHRT, which is the union of "BERT", i.e. the transformer algorithm developed by Google in 2018 (134), applied to EHRs.

Then, very recently, also Rasmy et al (135) applied BERT to structured EHRs, proposing the so called MED-BERT. They used data about hospitalizations' diagnoses, and they used BERT to predict heart failure in patients with diabetes and pancreatic cancers on three different cohorts extracted from EHR databases in the United States.

The aim of this study was instead to analyze the healthcare trajectories of elderly population in the Piedmont region (Northern Italy) extracted from medication prescriptions and hospitalizations' diagnoses from HADs, to predict urgent hospitalizations 3 months in advance.

The previous studies applied DL approaches to EHRs, meanwhile we applied BERT to HADs, using both medications prescriptions and hospitalization diagnoses, and not only the second ones like in the majority of the previously published works.

Furthermore, at our knowledge, this is the first time medication prescriptions made by general physicians, and not drugs prescribed during hospitalizations, were considered in a similar setting.

Additionally, we tried to understand which type of information, between medication prescriptions, hospitalization diagnoses and demographics data, are more informative to predict a future urgent hospitalization.

Material and Methods

Data source

Data were extracted from the Piedmont Longitudinal Study (PLS), which is a study built through record-linkage of census data with administrative ones (i.e. hospital discharges, drugs prescriptions, outpatients cares, and so on) of the inhabitants of Piedmont region, in the North-West of Italy.

In this study we considered only subjects aged at least 65 years at the 1st of January 2015, thus resulting in a cohort of 1 159 141 people.

Data about hospital admissions and drug prescriptions occurred between 1st of January 2015 and 31 December 2018 were collected. Then, age, gender and educational level, used as proxy of socio economic status (SES), were also gathered from the 2011 census.

Method: Deep Learning

Artificial Intelligence, Machine Learning, Deep Learning

In the classical programming we are used to have rules and data as input, and we obtain answers as output. In ML the inputs are instead data and answers, and the output are rules (Figure 14).

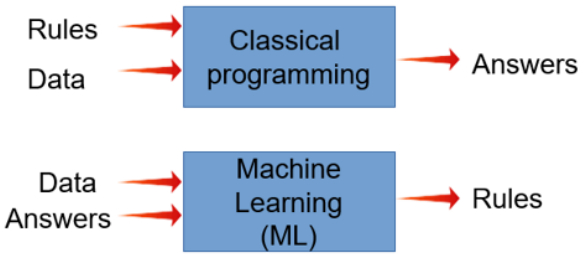


Figure 14: Classical programming vs Machine Learning

ML is a bub-field of Artificial Intelligence (AI), which comprises any technique which enables computers to mimic human behavior. More in detail, ML are AI techniques that give computers the ability to learn from data by training, without being explicitly programmed to do so (Figure 15Figure 15) (136).

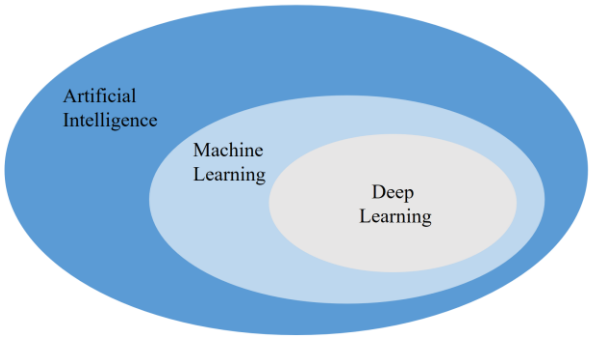


Figure 15: Artificial Intelligence, Machine Learning and Deep Learning

ML requires three ingredients:

1) Input data;
2) Examples of the expected output;
3) A way to measure the error made by the algorithm, i.e. the distance between the algorithm's output and the expected one. This error is then used as a feedback signal to adjust the rules that the algorithm uses. This is the mechanism through the algorithm learns rules between input and output data.

In other words, the ML algorithm is exposed to examples of pairs of input and output data, from which it learns the optimal transformation of input data into meaningful output (136).

In other words, let us consider the example reported in Figure 16. Let's suppose that we want an algorithm which is able to correctly classify the point's color, given its coordinates (x, y). In this case, inputs are the coordinates of the points, and the outcome is the color of the points. The error could be given by the percentage of points being incorrectly classified. The aim of the ML algorithm is to find a new representation of the input data which correctly separate the red points from the blue ones. For example, after many attempts of data transformation, the algorithm chooses the one that leads to the minimum percentage of misclassified points. In particular, the algorithm finds that a coordinate change is the most useful representation of the input data and it learns the following rule (136):

"if x'>0, then the point is blue, otherwise it is red".

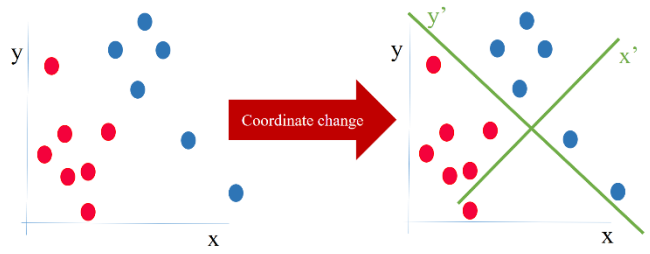


Figure 16: Example of data transformation,

ML has been developed to overcome many weaknesses which occur in classical programming. For example, it requires strong assumptions, it has difficulties to deal with long-term dependencies, it requires the experts' ability to define appropriate features, it is not able to deal with irregularities of the intervals between two events, and so on. Furthermore, when a big amount of data is available, traditional algorithms are not able to increase the performance when the data amount is increasing. Large neural networks (NN) are instead able to take advantage as the amount of data increases (Figure 17).

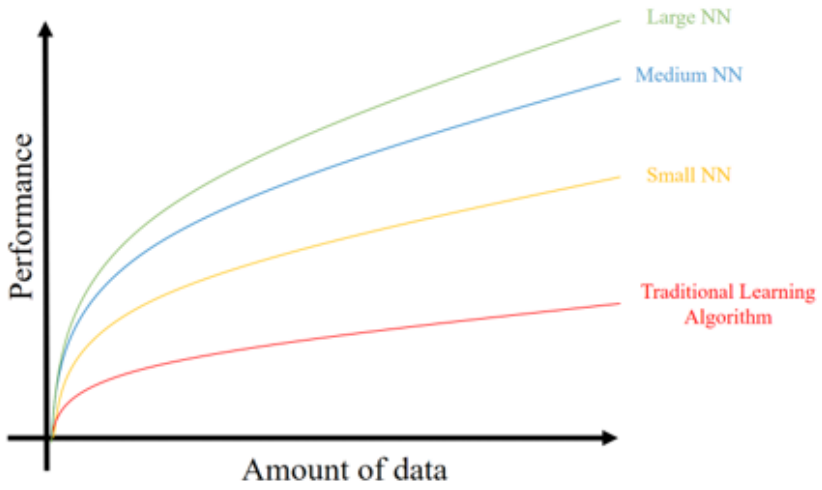


Figure 17: Deep learning and big data.

DL is a subfield of ML, and it learns successive layers of representations of data that have an increasingly meaning. These layers of successive representation of input data, are learned via neural networks (NN), that are models composed by several successive layers. A simple NN has only one hidden layer, which makes only one data transformation to obtain a better representation of the input data. If the hidden layers are instead almost two, we refer to that NN as a DL NN (Figure 18).

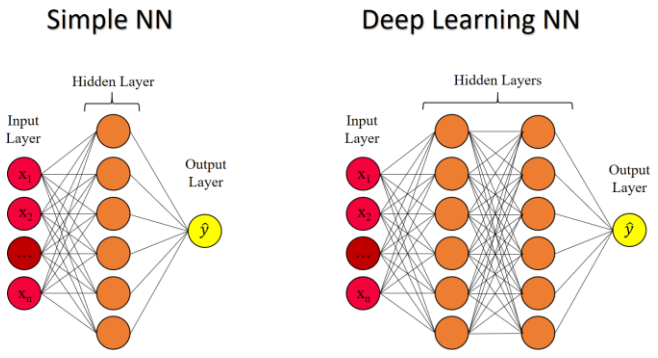


Figure 18: Simple Neural Network vs Deep Learning Neural Network.

Simple NN

The simple architecture behind DL is the artificial neural network (ANN). Each perceptron is composed by input data, a hidden layer where computations are performed, and an output layer (Figure 18).
The first step in a ANN is to compute a weighted sum Z, composed by a bias term b, a weights matrix W, and an input matrix **X** (137).

$$Z = b + \sum_{i=1}^{n} W_i \boldsymbol{X_i}$$

The output of step 1 is then passed to the activation function g, which is a mathematical function that transforms the output to a non-linear format in a desired range, and it is successively passed to the next layer. The most used activation functions are the sigmoid, relu, and hyperbolic tangent functions (137).

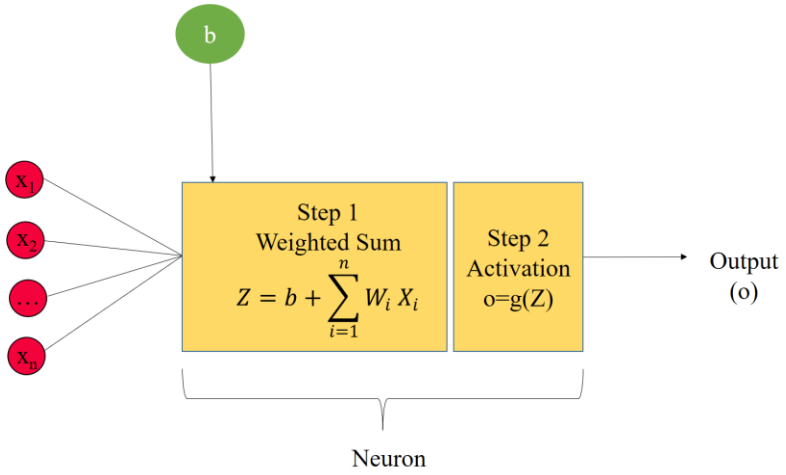


Figure 19: Simple NN structure.

These computations are performed in each neuron in a NN, including the output layer, and one of this passage is called forward propagation.

Understanding how DL works

A supervised DL NN needs to observe several examples of input-output pairs to learn the better input-to-output mapping via a deep sequence of data transformations (layers), which are expressed through numbers called weights or, sometimes, parameters of the layer (136). Weights contain the information learned by the NN from the training data, i.e. from the exposure to input-output pairs. Learning means

find the optimal set of weights for all the NN layers that allows to the NN to correctly map inputs to their outputs.

First, weights are set completely at random. Then, NN tries to associate an output to each input through these values of weights. Obviously, when weights are set completely at random, when the NN tries to associates outputs to inputs it makes errors. So, the NN computes the error that it performs, i.e. it evaluates through a loss function a distance score, which measures how far is the predicted output from the true.

One example of loss function is the Mean Squared Error (MSE) defined as

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$$

where $\widehat{Y}_i$ is the predicted output, Y is the true output, and n is the sample size.

This score is then used as a feedback signal to optimize the weights set, that are changed in order to lower the loss score. This procedure is performed by an optimizer, which takes advantage from the backpropagation algorithm.

For example, one algorithm typically used to minimize the loss function is the stochastic gradient descent (SGD). When the loss value which measures the mismatch between the output predicted and the true one is obtained, the gradient of the loss with regard to the NN weights is computed. Then, weights are moved in the opposite direction from the gradient and in this way the loss in reduced. The amount of change during each step of this updating process, is called learning rate, which controls how quickly a NN model learns from data.

If gradient values of a NN are computed via the chain rule (f(g(x)) = f'(g(x)) * g'(x)), a back-propagation algorithm is applied. Back-propagation starts with the final loss value and going backward from the last layer to the first layer, it applies the chain rule to compute the contribution that each weight had in the computation of the loss value.

Another way of minimizing the loss function is the adaptive moment estimation (Adam) optimizer, which is an extension of the SGD, often used in the Natural Language Processing (NLP) context (138). In the classical SGD, a single learning rate is maintained throughout the training. Contrariwise, in Adam optimizer individual adaptive learning rates are computed for different parameters from estimates of first and second moments of the gradients (138).

The process of learning described above is represented in Figure 20 (136).

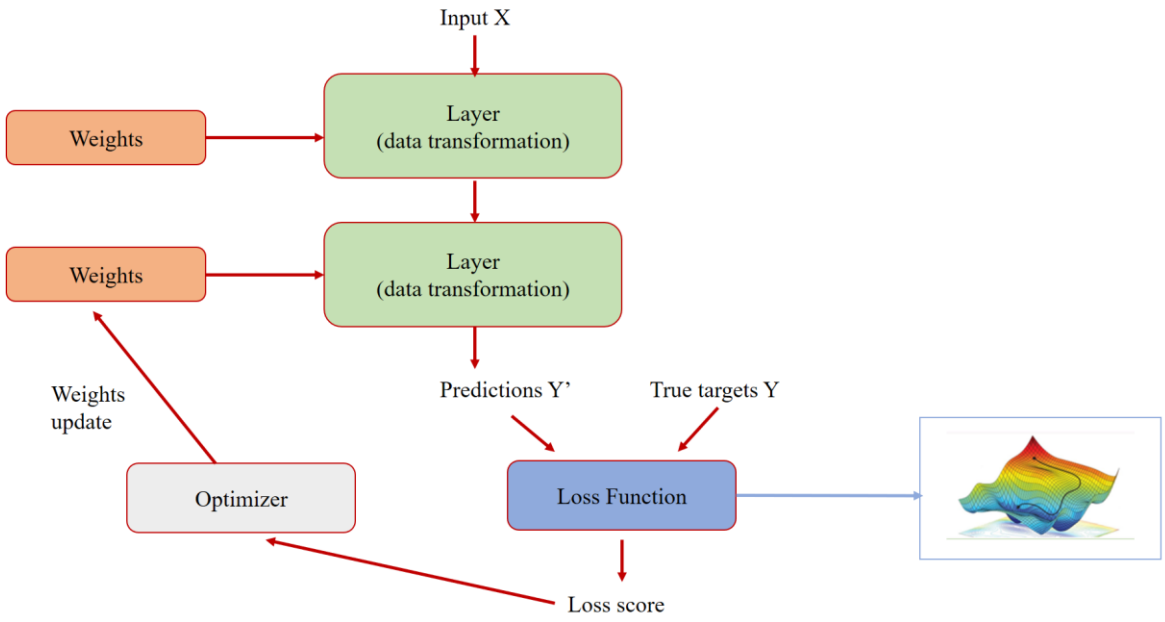


Figure 20: The process of learning of a DL algorithm.

Subsequently, when the model has been trained, it must be evaluated for its ability to generalize what it has learned.

Data are split into three sets: training, validation, and test sets. The model is trained on training set, then it is evaluated on validation set. In this phase, hyper-parameters (i.e. the number of layers, the size of layers, the learning rate, etc..) are tuned.

Finally, the model is tested one final time on the test set, consisting of data that the algorithm has never seen (136).

DL on text sequences

In my application, I will consider healthcare trajectories of elderly people with the aim to learn history of poly-pharmacy and multi-morbidity using hospitalization diagnoses and medication prescriptions, extracted from HADs. More in detail, each ATC code or diagnosis category corresponds to a word in the NLP field, one hospitalization or the set of the medical prescriptions made in the same day corresponds to a sentence, the entire healthcare trajectory of a subject corresponds to a document.

So, let's see more in detail how DL models can process text, i.e. a sequence of words, time-series and, more in general, sequence data. In fact, DL is able to understand sequence data and it can produce a basic form of natural language understanding.

However, when DL is working on text, it is not able to work with input raw text, but it is necessary to perform data pre-processing to transform each word of the input raw text into a numeric tensor, through the so called process of "vectorization". Each word is called token, and this process could also be called tokenization.

There exist multiple ways to associate a vector with a token. The most used are the one-hot encoding and the token embedding (136). In the first one, a unique integer index is associated with every word and then it is turned into a 0-1 vector of size N (the size of the vocabulary). The vector has 1 in the i^{th} position, otherwise it is zero, leading to a sparse vector. Word embeddings are low dimensional dense floating-point vectors, directly learned from data, jointly with the main task of the training. Furthermore, while in one-hot encoding all the words have the same distance between them, in word embedding words with a similar meaning have a lower distance if compared with that one of words with different meanings (Figure 21) (136).

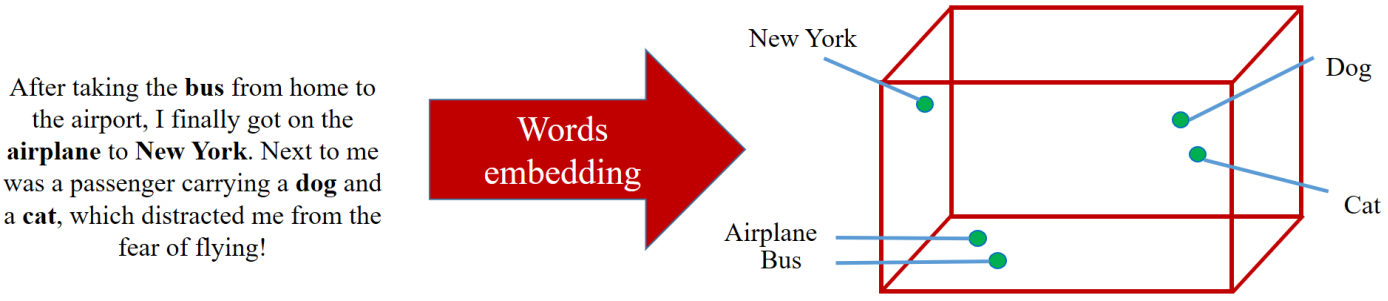


Figure 21: An example of word embedding.

BERT: Bidirectional Encoder Representations from Transformers

In 2019 Jacob Devlin et colleagues (134), members of the Google AI group, developed a new language representation model called BERT, which is the acronyms of Bidirectional Encoder Representations from Transformers.

Transformer is a particular DL model which takes advantage from the attention mechanism, that is a process that weights differently the influence of distinct parts of the input data (139). Furthermore, the attention units produce embeddings for every token, containing information about the token itself and a weighted combination of other tokens that the attention mechanism choose as relevant for the task.

Transformers were developed by Google to handle sequential data, for example for tasks like language translations or text summarizations in the NLP context.

Transformer is an encoder-decoder architecture, which is represented in Figure 22. More in detail, in the encoder part there are a set of encoding layers that iteratively process the input to generate encodings containing information on which part of the input is more relevant, taking advantage from the attention mechanism. Then, the input is passed to the decoder part, composed by a set of decoder layers that process the input which came from the encoder part, in the opposite direction. More in detail, decoder layer extracts from encodings their contextual information and generate the output. Both encoder and decoder layers have a feed-forward NN to further process the outputs encoding individually. The first encoder layer takes as input the embeddings of the input sequence. Finally, a linear transformation and a softmax layer are placed after the last decoder layer, to generate the output probabilities over the vocabulary.

Transformer are semi-supervised learning, i.e. it has a first phase of unsupervised pre-training in which the algorithm is used to learn the data structure, followed by a supervised fine-tuning phase with a specific task.

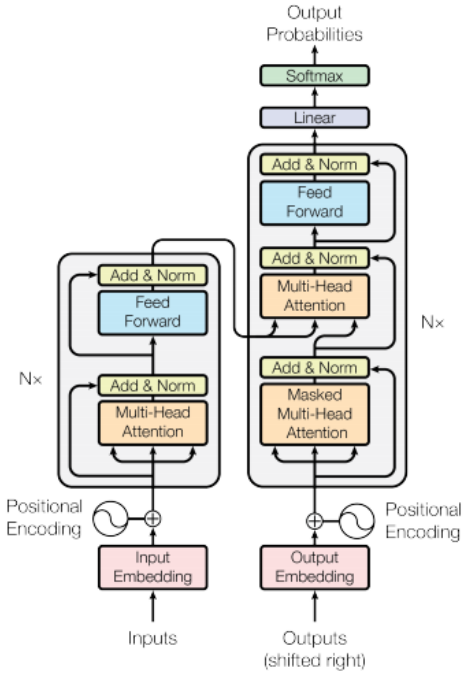


Figure 22: The Transformer-model architecture, taken from (139).

More in detail, the unsupervised pre-training is performed via two different mechanisms (134):

1) Masked Language Modeling (MLM). It randomly selects some words and it masks them. Then, BERT learns to predict the original words, taking advantage from the bidirectional context. More

in detail, it selected 15% of words randomly, and they were modified according to the following probabilities:

a. 80% of the times → (MASK)
b. 10% of the times → a random word was substituted
c. 10% of the times → unchanged

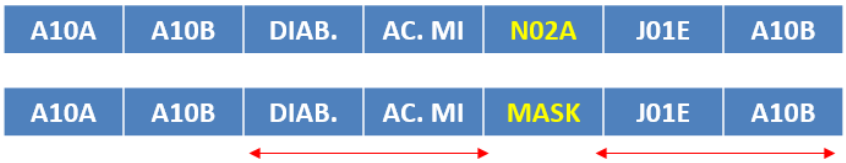


Figure 23: Masked Language Modelling

2) Next Sentence Prediction (NSP). Given two sentences, BERT learns whether one sentence follows the other or not. So, given the sentence A, is sentence B the following? The algorithm answers YES/NO.

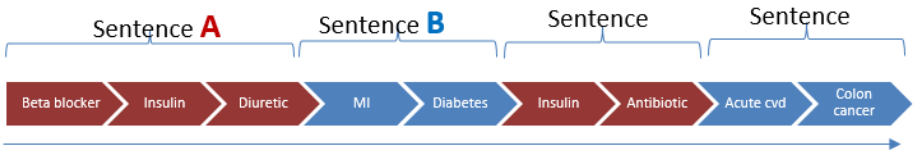


Figure 24: Next Sentence Prediction.

In the supervised fine-tuning phase, BERT is initialized with the parameters learned in the pre-training phase. Then, these parameters are updated using labeled data which are task-specific.
BERT is a deeply bidirectional transformer, that means that the encoder takes into account the context (left and right side) in which a word occurs, being able to give different meanings to the same words.

BERT applied on HADs

Data pre-processing

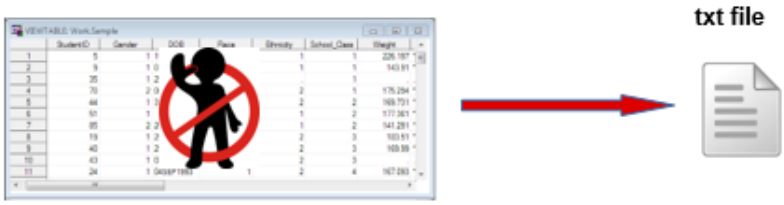
Medication prescriptions were represented through 4-digits of the Anatomical Therapeutic Chemical Classification System (ATC) codes.

Diagnoses were collected from the hospital discharges records, through the International Classification of Diseases, 9th edition (ICD-9-CM). However, to reduce dimensionality, we grouped diagnoses via the Single-level Clinical Classification Software (CCS) for ICD-9-CM (140).

If one subject died or emigrated outside Piedmont, we removed his/her last 3 months before the date of death/emigration.

Input data are not in a structured form, but they are stored into a text-file as shown in Figure 25. More specifically, for each subject it was created one row for each event that occurred in his/her healthcare trajectory (one hospitalization or one medication prescription). Each event is a set of ATC and/or ICD-9-CM codes. In the first row, ATC or diagnoses of the first event are reported. More in detail. for each event a new row, which contains the previous events and the new one, is added. Then, to each row was added a 0-1 label, which indicates if in the next 3 months an urgent hospitalization (i.e. a non-programmed hospitalization) occurred.

-




```
0
88 1 3
G03H C08C [SEP] G03H C08C [SEP] G03H [SEP] G03H C01D [SEP] G03H [SEP] G03H [SEP] C08C [SEP] B01A [SEP] G03H [SEP] C08C [SEP]
Peripathero Coronathero Prostatecan Otcirculdx [SEP] B01A [SEP] B03B G04C [SEP] C08C [SEP] C08C [SEP] C01D C08C [SEP] C08C
[SEP] Otherscreen Coronathero [SEP] Precereoccl [SEP] C09A C10A [SEP] G03H [SEP] G04C G03H B01A [SEP] Fxhip Htncomplicn
Otaftercare Fluidelcdx Acutphanm Complicproc [SEP] Otconntiss Fxhip Coronathero [SEP] G03H [SEP] A02B B01A C08C [SEP] G03H
[SEP] G03H [SEP] A02B G03H C08C B01A [SEP] N06A


0
81 1 3
J01E C03C B01A M01A C09A A02A A02B R03A R03B C07A [SEP] C09A J01M A02B M04A [SEP] H02A C03C A02A B05B [SEP] C03C M01A A02B
G04C [SEP] M04A [SEP] B01A C03C C09A J01M A02B G04C C07A [SEP] C03C J01M A02A M04A [SEP] R03A R03B [SEP] C03C C09A A02B G04C
M04A [SEP] J01M [SEP] B01A J01M [SEP] A02A [SEP] C03C R03B [SEP] C03C J01M B01A C09A A07E A02B M04A C07A [SEP] M04A [SEP] J01M
[SEP] R03B [SEP] C03C A10A [SEP] B01A C03C J01M C09A A02B [SEP] M04A M01A [SEP] J01C [SEP] B01A [SEP] N02B [SEP] A02A J01M
[SEP] B01A J01M A02B C07A [SEP] J01M [SEP] C07A A02B J01M M04A [SEP] A02A C03C [SEP] C03C C09A J01M A02B R03B R03A [SEP] R03B
R03A [SEP] C03C [SEP] M01A A02A C07A [SEP] C09A A02B M01A C03C [SEP] V03A [SEP] A02B M04A C09A C03C B01A [SEP] M01A [SEP] J01M
C03C [SEP] B01A A02B J01M C07A R03B C03C C09A [SEP] M01A [SEP] A02A A02B G04C J01M [SEP] Fxhip DiabMelnoc Chrkidneydisease
COPD Coronathero Acutphanm [SEP] Fxhip Rehab Complicproc Anemia UTI Adltrespfl [SEP] M01A R03A A02B R03B J01C [SEP] J01M [SEP]
J01X [SEP] J01C G04C [SEP] C09A A02B R03A R03B A02A C03C C07A [SEP] J01C [SEP] A07A A02B [SEP] COPD Acrenlfail [SEP] B01A
[SEP] A02B [SEP] R03A J01M R03B [SEP] B01A A02B M01A C09A H02A J01C M04A [SEP] J01C B05B M01A A07A C07A [SEP] A02B C03C [SEP]
M01A [SEP] A02B C09A G04C C03C [SEP] A02B M01A
```

Figure 25: Input data format.

Furthermore, three numbers indicating age, gender and SES were also added, through a wide-and-deep model, i.e. concatenating them to the input of the last BERT layer, whose output is then converted into class probabilities for the prediction task (Figure 25 and Figure 26).

In the training phase we randomly allocated 100 000 samples in the training set, 25 000 in the validation set, and 25 000 in the test set.

Due to computational limits, at this stage of the work we were not able to use all the available data. To handle un-balancement in the data set, 1:1 random oversampling was performed.

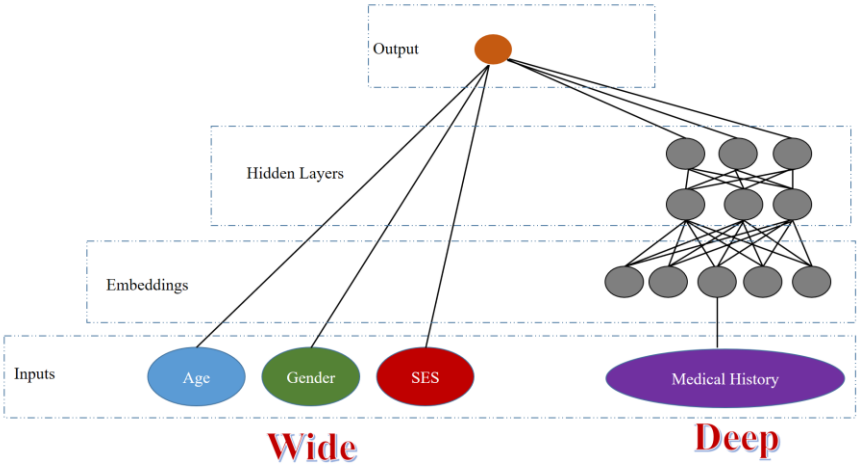


Figure 26: Wide and deep model.

Scenarios

Four different scenarios were analyzed, to asses which information (between drugs prescriptions, hospitalization diagnoses and demographic data) is more useful to predict urgent hospitalizations in elderly subjects.

i. Scenario (1): medications' prescriptions (ATC7 codes), hospitalizations diagnoses (ICD9CM codes) and individual characteristics (IC), i.e. age, gender and SES
ii. Scenario (2): medications' prescriptions (ATC7 codes), and hospitalizations diagnoses (ICD9CM codes)
iii. Scenario (3): hospitalizations diagnoses (ICD9CM codes) and individual characteristics (IC)
iv. Scenario (4): medications' prescriptions (ATC7 codes), and individual characteristics (IC).

In each scenario, the goal was to predict a new urgent hospitalization within 3 months from the last event, i.e. a hospitalization or a medication prescription according to different scenarios.

The application of BERT to HADs

BERT was applied to healthcare trajectories of elderly people to learn history of poly-pharmacy and multi-morbidity using hospitalization diagnoses and medication prescriptions, extracted from HADs described in Chapter 1.

Each ATC code or diagnosis category corresponds to a word in the NLP field, one hospitalization or the set of the medical prescriptions made in the same day corresponds to a sentence, the entire healthcare trajectory of a subject corresponds to a document.

BERT was pre-trained to learn the data structure, through the original pre-training algorithm of the BERT algorithm developed by Google, i.e. MLM and NSP. More in detail, three pre-trainings were conducted on the whole sample and in three scenarios aforementioned. First, only hospitalizations' CCS categories were used. Then, only medications' ATC codes were considered and finally both hospitalizations' CCS categories and medications' ATC codes were taken into account.

Furthermore, to qualitatively evaluate the goodness of the embedding, a t-SNE reduction was performed, to have a dimensional graphical representation more interpretable. When only medications or only diagnoses were taken into account, the top 10 occurring codes and their nearest neighbors in term of cosine similarity were represented. When instead both diagnoses and medications were included, the top 5 medications and the top 5 diagnoses with their nearest neighbors were represented.

Finally, the attention mechanism enables prediction interpretation. Some examples are reported to show how attention weights from transformer layers connect some codes with each other in the pre-trained model.

With respect the original Google's BERT algorithm, we used a smaller one, to avoid over-fitting. More in detail, we used 6 layers, each with 2 attention mechanisms, 512 intermediate layers and 288 hidden size layers, in accordance with the work published by Li et colleagues (133).

The maximum sequence length was 512, and the vocabulary size changed in accordance with the considered scenario, to match the number of different medical codes to be considered (plus the special words needed by BERT for sentence separation, token masking, etc.). The resulting vocabulary sizes are 263 (hospitalization diagnoses only), 199 (medication prescriptions only) and 457 (hospitalization diagnoses and medication prescriptions).

Pre-training was performed with the original code by Google, based on the TensorFlow Python library (141) using the Adam optimizer for 20 epochs, i.e. the whole data was seen 20 times by the algorithm. Furthermore, the learning rate was set to 1e-4, the dropout rate to 0.1 and the batch size to 8, which means that 8 samples from the training set will be used to estimate the error gradient before the updating of weights.

Then, it was performed a fine-tuning in a supervised way to predict urgent hospitalization within 3 months from the last event, i.e. a hospitalization or a medication prescription, according to each scenario considered.

To import the pre-trained model we used the "hugging-face" Python library, based on PyTorch (142). Also in this phase, we used the Adam optimizer but with a smaller learning rate (2e-5), while keeping the same dropout rate, epochs, and batch sizes which were selected for pre-training.

Both pre-training and fine-tuning have been performed on a system equipped with a dual-core Intel Xeon processor with 40 cores, 128GB of RAM, 8TB of SSD drive, and an NVIDIA Titan XP GPU with 12GB of graphics memory.

Results of the pre-training phase were reported in terms of accuracy (ACC), i.e. the percentage of correctly classified terms.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP states for True Positive, TN for True Negative, TP for True Positive, TN for True Negative, FP states for False Positive and FN states for False Negative.

Considering the un-balancement in the outcome, results of the prediction task were instead reported in terms of:

i. Precision (PR)

$$PR = \frac{TP}{TP + FP}$$

ii. Recall (RC)

$$RC = \frac{TP}{TP + FN}$$

iii. F1 score (F_1)

$$F_1 = \frac{2\,TP}{2\,TP + FP + FN}$$

iv. Area Under the Receiver Operating Characteristic (AUROC), that gives information about the ability of a binary classificator to correctly discriminate between the two classes. More the value of AUROC is near 100%, better the classificator is.

Results

Results about MLM and NSP accuracies in the pre-training phase of BERT were reported in Table 19, according to different scenarios. When only hospitalization diagnoses were considered, the lowest accuracies both in MLM and NSP were reached (respectively 92% and 97%). Contrariwise, when both medication prescriptions and hospitalization diagnoses were taken into account, the highest accuracies were reported, both in MLM (97%) and in NSP (99%).

Table 19: Pre-training BERT accuracies, according to different scenarios and pre-training methods. Masked Language Modelling (MLM) and Next Sentence Prediction (NSP) approaches.

Scenarios	**MLM accuracy**	**NSP accuracy**
Medications only	94.57 %	99.63 %
Diagnoses only	92.18 %	97.25 %
Medications and diagnoses	96.58 %	99.75 %

In Figure 27, the embedding results are shown, according to different scenarios. The embedding clusters are potentially reflecting co-occurring conditions and/or the belonging to the same clinical group.

If we analyze more in details, in Figure 27A we can see some clusters of diagnoses. In particular, the leftmost part of the figure presents a cluster composed by diabetes, cancer in urinary organs and hypertension complications; in the central part there is the cluster of respiratory diseases grouping chronic obstructive pulmonary disease (COPD), asthma and respiratory failure. Below it, a group formed by circulatory diseases and dysrhythmia is represent, but it could be overlapped with respiratory failure; on the top of the central part appears a cluster which contains rehabilitation, osteoarthritis, fracture femur and connective tissue disease. In the rightmost part of the figure esophagus cancer and orthodontic aftercare diagnoses are nearby.

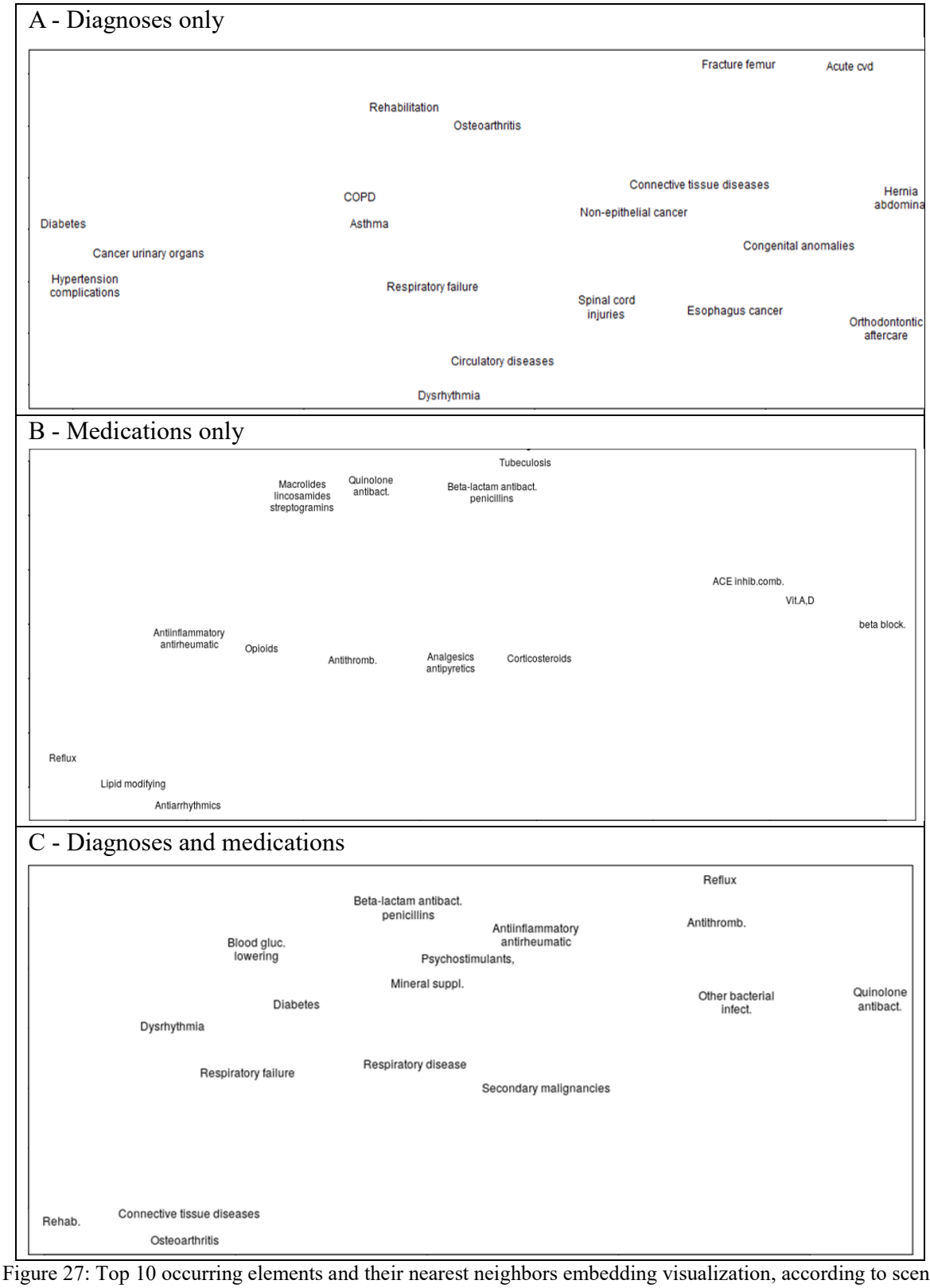


Figure 27: Top 10 occurring elements and their nearest neighbors embedding visualization, according to scenarios.

Furthermore, in Figure 27 B we can see the embedding of the pre-training scenario including medications only. In this situation, four clusters have been identified: one in the lower left part, composed by lipid modifying agents, anti-arrhythmic agents and reflux drugs; in the center, one composed by anti-inflammatory/anti-rheumatic, opioids, anti-thrombotic, analgesics/antipyretics and corticosteroids; on the top, a cluster of antibiotics and on the most right a cluster of ACE inhibitors and beta-blockers agents. Finally, if we consider both medications and diagnoses (Figure 27C), we can see in the bottom left of the figure that rehabilitations, osteoarthritis and connective tissue disease are once again nearby. Then, in the central part, respiratory failure is represented close to respiratory and dysrhythmia diagnoses; meanwhile blood glucose lowering drugs are near diabetes diagnoses. Finally, in the right part of the figure, we find two clusters: one groups anti-inflammatory/anti-rheumatic and anti-thrombotic drugs with reflux drugs and one composed by beta-lactame/penicillin antibacterial, anti-inflammatory/anti-rheumatic, quinolone antibacterial drugs and other bacterial infections diagnoses.

In

Table 20 we can find results about the prediction task (urgent hospitalization within 3 months). If we compare the different scenarios in terms of precision, recall, F1 score and AUROC, the best was the one with all the available information, i.e. medication prescriptions, hospitalizations diagnoses and demographics characteristics. In fact, in this scenario the algorithm reached good results, with a precision of 61%, a recall of 89%, a F1 score of 73% and a AUROC of 97%.

The worst performance was instead observed when only hospitalization diagnoses and demographic characteristics were considered, with a precision of 21%, a recall of 37%, a F1 score of 27% and a AUROC of 62%.

The Scenario 2, i.e. considering medication prescriptions and demographic characteristics showed a lower precision and F1 score if compared with the Scenario 1.

Finally, the Scenario 4 is pretty identical to the Scenario 1, which means that demographic characteristics i.e. age, gender and SES, do not help to learn the prediction task.

Table 20: Results of prediction task according to different scenarios. Med=Medications, Diag=diagnoses, Demo= demographics (age, SES, gender).

	Precision	**Recall**	**F1 score**	**AUROC**
Scenario 1: Med + Diag + Demo	61%	89%	73%	97%
Scenario 2: Med + Demo	51%	89%	65%	97%
Scenario 3: Diag + Demo	21%	37%	27%	62%
Scenario 4: Med + Diag	62%	87%	73%	97%

In conclusion, two randomly selected examples of attention mechanism graphs are represented in Figure 28, where healthcare trajectories are plotted against themselves. Links identify relations between events that are detected by the attention mechanism. Their thickness represents the strength of the relation detected by the attention mechanism. In the selected examples, we can see that the medication prescription identified by the ATC7 code H03A, which identify the class of drugs that are grouped into "thyroid preparations" class, is linked to the "thyroid disorders" hospitalization diagnosis. Then, in Figure 28 B, the calcium channel blockers, with mainly vascular effects (C08C) is linked to "coronary atherosclerosis and other heart disease" CCS class.

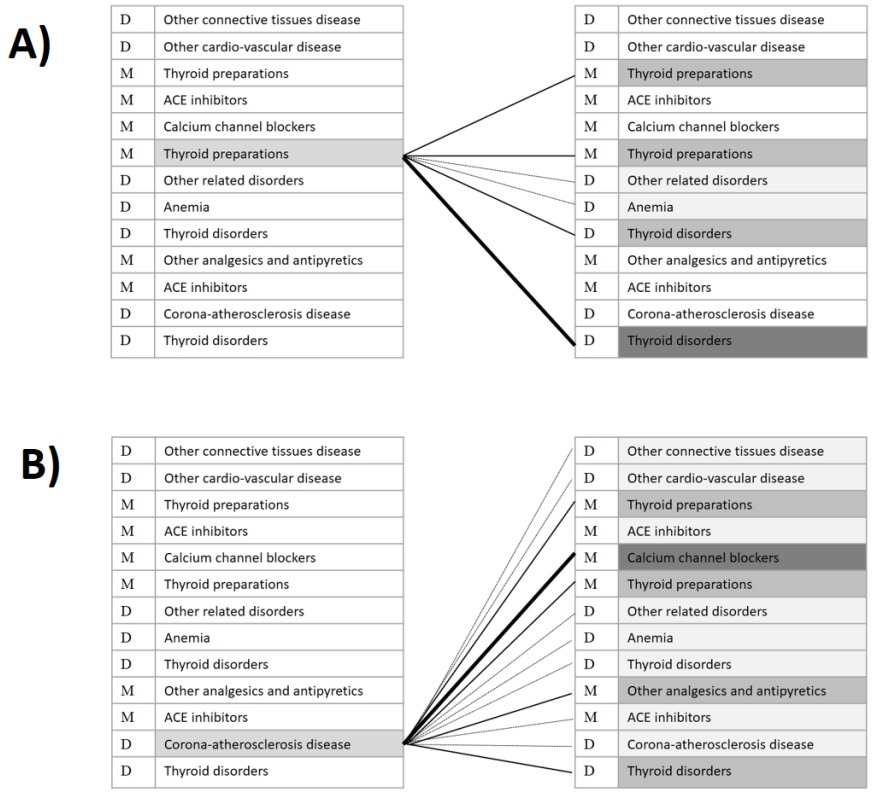


Figure 28:Attention mechanism in the pre-training phase. D= hospitalization diagnoses, M=medications

Discussion

Italy has the biggest proportion in Europe of citizens aged almost 65 years (143). Lots are the consequences of ageing in the population on the health care system. In fact, elderly people are often affected by multi-morbidity, i.e. they have to deal simultaneously with more than two chronic diseases at a time, which make their management more difficult (7).

In a report published in 2011, WHO pointed out the necessity of developing advanced models to predict the healthcare trajectories of elderly populations, to prevent adverse outcomes improving the management of such people, in a context of personalized medicine (144).

In the last years, lots experiments of application of novel methods to prevent adverse outcomes in elderly people have been conducted. One of them concerns the application of BERT to a primary care dataset of diagnoses to predict the next diagnoses occurring in the subsequent months (134) (133). The authors used a sample of subjects without age restrictions. In particular, they showed that BERT could be applied to structured data, and it learns history of the past diseases to predict future diagnoses.

At our knowledge, we tried for the first time to apply BERT to HADs to learn healthcare trajectories of people aged at least 65 years, to predict urgent hospitalizations occurring in the next 3 months. The peculiarity of this work is that BERT was applied to HADs and not to primary care databases, as the other published works did in this setting.

We considered both hospitalization diagnoses and medication prescriptions to reconstruct healthcare trajectories, and with some ablations studies we tried to understand which information is more useful.

We obtained promising results, in fact pre-training MLM, NSP, embeddings and attention mechanisms suggested that BERT is able to learn from structured HADs, because very satisfying performances in terms of ACCs were reached, with values greater than 97% when both medication prescriptions and hospitalization diagnoses were used.

Furthermore, embedding images showed that ATC7 or hospitalization diagnoses codes that often are co-occurring or share a similar medical meaning (for example anti-inflammatory/anti-rheumatic and opioids, or antibiotics) were plotter near in a bi-dimensional space.

In our experiment, we observed that the most informative data are those about medication prescriptions made by a general physician. Hospitalization diagnoses resulted as the less informative data, probably dye to limitations of the hospitalization HADs, in which diagnoses are reported mainly for reimbursement purposes (145,146). However, even if hospitalization diagnoses alone are not informative, we found that adding them to medication prescriptions helps to improve the performance in the prediction of urgent hospitalization within 3 months. However, including demographics information, i.e. age, gender and SES, do not improve the performance. A possible explanation could be that probably this information is incorporated into the healthcare trajectories themselves or it is overpassed by information related to diagnoses and medications.

BERT is able to accurately predict the real occurrence of urgent hospitalizations within 3 months (89% recall), but it tends to produce false positives (61% precision). However, false positives are the least

critical type of misclassification for this kind of application, because it is important to ensure that as many as possible high risk patients are identified by the algorithm.

The novelty of this work is the application of the DL algorithm BERT to healthcare HADs, considering hospitalization diagnoses extracted from the hospital discharge forms and medication prescriptions made by the general physicians, and not data extracted directly from hospitals and clinics as made in previous works (130,133,135).

If used properly HADs can have a key role in epidemiological and medical research (147), and they are not yet used enough for these purposes.

Furthermore, at our knowledge this is the first study in which BERT is applied to an elderly population aged more than 65 years, to predict urgent hospitalizations, modelling the healthcare trajectories extracting information from secondary care information in administrative sources and results seem to be promising.

In conclusion, our results suggest that BERT can be used not only in NLP setting but it is also able to embed medical healthcare trajectories, reconstructed from HADs extracting information about hospitalization diagnoses and medication prescriptions made by general physicians.

This tool could be used for predicting future urgent hospitalizations of elderly population, providing an important tool that could help to plan the allocation of healthcare resources in the future and to manage diseases in a personalized way, helping also to improve the quality of life of aging population with prompt interventions when the probability of urgent hospitalization reaches a critical threshold, suggesting a worsening of the healthcare condition of a subject, which need more attention.

This chapter is in review as

Deep Learning for predicting urgent hospitalizations in elderly population using administrative Electronic Health Records. Sciannameo V, Jahier Pagliari D, Ferracin E, Ricotti A, Ricceri F, Costa G, Berchialla P. Artificial Intelligence in Medicine

Bibliography

1. Bernell S, Howard SW. Use Your Words Carefully: What Is a Chronic Disease? Front Public Health [Internet]. 2016 Aug 2;4:159–159. Available from: https://pubmed.ncbi.nlm.nih.gov/27532034

2. WHO. Noncommunicable Diseases [Internet]. 2016. Available from: http://www.who.int/topics/noncommunicable_diseases/en/

3. Australian Institute of Health and Welfare. Chronic Diseases [Internet]. 2016. Available from: : http://www.aihw.gov.au/chronic-diseases/

4. Uijen AA, van de Lisdonk EH. Multimorbidity in primary care: prevalence and trend over the last 20 years. Eur J Gen Pract. 2008;14(sup1):28–32.

5. Barnett K, Mercer SW, Norbury M, Watt G, Wyke S, Guthrie B. Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. The Lancet [Internet]. 2012 Jul 7;380(9836):37–43. Available from: https://www.sciencedirect.com/science/article/pii/S0140673612602402

6. Calderón-Larrañaga A, Poblador-Plou B, González-Rubio F, Gimeno-Feliu LA, Abad-Díez JM, Prados-Torres A. Multimorbidity, polypharmacy, referrals, and adverse drug events: are we doing things well? Br J Gen Pract [Internet]. 2012 Dec 1;62(605):e821. Available from: http://bjgp.org/content/62/605/e821.abstract

7. Blanda MP. Pharmacologic Issues in Geriatric Emergency Medicine. Geriatr Emerg Med [Internet]. 2006 May 1;24(2):449–65. Available from: http://www.sciencedirect.com/science/article/pii/S0733862706000083

8. Beard JR, Officer A, de Carvalho IA, Sadana R, Pot AM, Michel J-P, et al. The World report on ageing and health: a policy framework for healthy ageing. The Lancet [Internet]. 2016 May 21 [cited 2021 Apr 29];387(10033):2145–54. Available from: https://doi.org/10.1016/S0140-6736(15)00516-4

9. WHO diabetes [Internet]. [cited 2021 Apr 13]. Available from: https://www.who.int/news-room/fact-sheets/detail/diabetes

10. Fadini GP, Zatti G, Consoli A, Bonora E, Sesti G, Avogaro A, et al. Rationale and design of the DARWIN-T2D (DApagliflozin Real World evIdeNce in Type 2 Diabetes): A multicenter retrospective nationwide Italian study and crowdsourcing opportunity. Nutr Metab Cardiovasc Dis [Internet]. 2017 Dec 1;27(12):1089–97. Available from: https://www.sciencedirect.com/science/article/pii/S0939475317301813

11. Emerging Risk Factors Collaboration, Sarwar N, Gao P, Seshasai SRK, Gobin R, Kaptoge S, et al. Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies. Lancet Lond Engl [Internet]. 2010 Jun 26;375(9733):2215–22. Available from: https://pubmed.ncbi.nlm.nih.gov/20609967

12. Bourne RRA, Stevens GA, White RA, Smith JL, Flaxman SR, Price H, et al. Causes of vision loss worldwide, 1990–2010: a systematic analysis. Lancet Glob Health [Internet]. 2013 Dec 1 [cited 2021 Apr 27];1(6):e339–49. Available from: https://doi.org/10.1016/S2214-109X(13)70113-X

13. Saran R, Li Y, Robinson B, Ayanian J, Balkrishnan R, Bragg-Gresham J, et al. US Renal Data System 2014 Annual Data Report: Epidemiology of Kidney Disease in the United States. Am J Kidney Dis [Internet]. 2015 Jul 1 [cited 2021 Apr 27];66(1):A7. Available from: https://doi.org/10.1053/j.ajkd.2015.05.001

14. Nauck MA. Update on developments with SGLT2 inhibitors in the management of type 2 diabetes. Drug Des Devel Ther [Internet]. 2014 Sep 11;8:1335–80. Available from: https://pubmed.ncbi.nlm.nih.gov/25246775

15. Goring S, Hawkins N, Wygant G, Roudaut M, Townsend R, Wood I, et al. Dapagliflozin compared with other oral anti-diabetes treatments when added to metformin monotherapy: a systematic review and network meta-analysis. Diabetes Obes Metab [Internet]. 2014 May 1 [cited 2021 Apr 27];16(5):433–42. Available from: https://doi.org/10.1111/dom.12239

16. Storgaard H, Gluud LL, Bennett C, Grøndahl MF, Christensen MB, Knop FK, et al. Benefits and Harms of Sodium-Glucose Co-Transporter 2 Inhibitors in Patients with Type 2 Diabetes: A Systematic Review and Meta-Analysis. PLOS ONE [Internet]. 2016 Nov 11;11(11):e0166125. Available from: https://doi.org/10.1371/journal.pone.0166125

17. Scheerer MF, Rist R, Proske O, Meng A, Kostev K. Changes in HbA1c, body weight, and systolic blood pressure in type 2 diabetes patients initiating dapagliflozin therapy: a primary care database study. Diabetes Metab Syndr Obes Targets Ther [Internet]. 2016 Oct 31;9:337–45. Available from: https://pubmed.ncbi.nlm.nih.gov/27822077

18. Gorgojo-Martínez JJ, Serrano-Moreno C, Sanz-Velasco A, Feo-Ortega G, Almodóvar-Ruiz F. Real-world effectiveness and safety of dapagliflozin therapy added to a GLP1 receptor agonist in patients with type 2 diabetes. Nutr Metab Cardiovasc Dis [Internet]. 2017 Feb 1 [cited 2021 Apr 27];27(2):129–37. Available from: https://doi.org/10.1016/j.numecd.2016.11.007

19. Sosale B, Sosale A, Bhattacharyya A. Clinical Effectiveness and Impact on Insulin Therapy Cost After Addition of Dapagliflozin to Patients with Uncontrolled Type 2 Diabetes. Diabetes Ther Res Treat Educ Diabetes Relat Disord [Internet]. 2016/10/19 ed. 2016 Dec;7(4):765–76. Available from: https://pubmed.ncbi.nlm.nih.gov/27761881

20. . U.S._Department_of_Health_and_Human_Services_Food_and_Drug _Administration. The drug development process 2018. [Internet]. 2020. Available from: . U.S._Department_of_Health_and_Human_Services_Food_and_Drug _Administration. The drug development process 2018. Available from: https://www.fda.gov/ForPatients/Approvals/Drugs/default.htm.

21. Goetghebeur E, le Cessie S, De Stavola B, Moodie EE, Waernbaum I, "on behalf of" the topic group Causal Inference (TG7) of the STRATOS initiative. Formulating causal questions and principled statistical answers. Stat Med [Internet]. 2020 Dec 30 [cited 2021 May 14];39(30):4922–48. Available from: https://doi.org/10.1002/sim.8741

22. Tashkin D, Amin A, Kervin E. Comparing Randomized Controlled Trials and Real-World Studies in Chronic Obstructive Pulmonary Disease Pharmacotherapy. Comp Randomized Control Trials Real-World Stud Chronic Obstr Pulm Dis Pharmacother. 2020;(15):1225–43.

23. Booth CM, Tannock IF. Randomised controlled trials and population-based observational research: partners in the evolution of medical evidence. Br J Cancer [Internet]. 2014 Feb 1;110(3):551–5. Available from: https://doi.org/10.1038/bjc.2013.725

24. Grootendorst DC, Jager KJ, Zoccali C, Dekker FW. Observational Studies Are Complementary to Randomized Controlled Trials. Nephron Clin Pract [Internet]. 2010;114(3):c173–7. Available from: https://www.karger.com/DOI/10.1159/000262299

25. Murad MH, Asi N, Alsawas M, Alahdab F. New evidence pyramid. Evid Based Med [Internet]. 2016 Aug 1;21(4):125. Available from: http://ebm.bmj.com/content/21/4/125.abstract

26. Sherman RE, Anderson SA, Dal Pan GJ, Gray GW, Gross T, Hunter NL, et al. Real-World Evidence — What Is It and What Can It Tell Us? N Engl J Med [Internet]. 2016 Dec 7 [cited 2021 Apr 29];375(23):2293–7. Available from: https://doi.org/10.1056/NEJMsb1609216

27. corrao G, Mugelli A, Rossi F, Lanati EP. REAL WORLD DATA E REAL WORLD EVIDENCE: considerazioni e proposte da un network di società scientifiche. 2017.

28. Azzolina D, Baldi I, Barbati G, Berchialla P, Bottigliengo D, Bucci A, et al. Machine learning in clinical and epidemiological research: isn't it time for biostatisticians to work on it? 2019;

29. Black N. Why we need observational studies to evaluate the effectiveness of health care. BMJ [Internet]. 1996 May 11;312(7040):1215–8. Available from: https://pubmed.ncbi.nlm.nih.gov/8634569

30. Saturni S, Bellini F, Braido F, Paggiaro P, Sanduzzi A, Scichilone N, et al. Randomized controlled trials and real life studies. Approaches and methodologies: a clinical point of view. Pulm Pharmacol Ther [Internet]. 2014 Apr 1;27(2):129–38. Available from: https://www.sciencedirect.com/science/article/pii/S1094553914000170

31. Patel A, Billot L. Reality and Truth. Circulation [Internet]. 2017;136(3):260–2. Available from: https://www.ahajournals.org/doi/abs/10.1161/CIRCULATIONAHA.117.029233

32. Haut ER, Pronovost PJ, Schneider EB. Limitations of Administrative Databases. JAMA [Internet]. 2012 Jun 27 [cited 2020 Sep 6];307(24):2589–90. Available from: https://doi.org/10.1001/jama.2012.6626

33. Campbell PG, Malone J, Yadla S, Chitale R, Nasser R, Maltenfort MG, et al. Comparison of ICD-9-based, retrospective, and prospective assessments of perioperative complications: assessment of accuracy in reporting. J Neurosurg Spine. 2011 Jan;14(1):16–22.

34. Davies MJ, D'Alessio DA, Fradkin J, Kernan WN, Mathieu C, Mingrone G, et al. Management of Hyperglycemia in Type 2 Diabetes, 2018. A Consensus Report by the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD). Diabetes

Care [Internet]. 2018 Dec 1;41(12):2669. Available from: http://care.diabetesjournals.org/content/41/12/2669.abstract

35. Wiviott SD, Raz I, Bonaca MP, Mosenzon O, Kato ET, Cahn A, et al. Dapagliflozin and Cardiovascular Outcomes in Type 2 Diabetes. N Engl J Med [Internet]. 2018 Nov 10 [cited 2021 May 19];380(4):347–57. Available from: https://doi.org/10.1056/NEJMoa1812389

36. Neal B, Perkovic V, Mahaffey KW, de Zeeuw D, Fulcher G, Erondu N, et al. Canagliflozin and Cardiovascular and Renal Events in Type 2 Diabetes. N Engl J Med [Internet]. 2017 Jun 12 [cited 2021 May 19];377(7):644–57. Available from: https://doi.org/10.1056/NEJMoa1611925

37. Holman RR, Bethel MA, Mentz RJ, Thompson VP, Lokhnygina Y, Buse JB, et al. Effects of Once-Weekly Exenatide on Cardiovascular Outcomes in Type 2 Diabetes. N Engl J Med [Internet]. 2017 Sep 14 [cited 2021 May 19];377(13):1228–39. Available from: https://doi.org/10.1056/NEJMoa1612917

38. Marso SP, Bain SC, Consoli A, Eliaschewitz FG, Jódar E, Leiter LA, et al. Semaglutide and Cardiovascular Outcomes in Patients with Type 2 Diabetes. N Engl J Med [Internet]. 2016 Sep 15 [cited 2021 May 19];375(19):1834–44. Available from: https://doi.org/10.1056/NEJMoa1607141

39. Marso SP, Daniels GH, Brown-Frandsen K, Kristensen P, Mann JFE, Nauck MA, et al. Liraglutide and Cardiovascular Outcomes in Type 2 Diabetes. N Engl J Med [Internet]. 2016 Jun 13 [cited 2021 May 19];375(4):311–22. Available from: https://doi.org/10.1056/NEJMoa1603827

40. Zaccardi F, Dhalwani NN, Dales J, Mani H, Khunti K, Davies MJ, et al. Comparison of glucose-lowering agents after dual therapy failure in type 2 diabetes: A systematic review and network meta-analysis of randomized controlled trials. Diabetes Obes Metab [Internet]. 2018 Apr 1 [cited 2021 May 19];20(4):985–97. Available from: https://doi.org/10.1111/dom.13185

41. Fadini GP, Simioni N, Frison V, Dal Pos M, Bettio M, Rocchini P, et al. Independent glucose and weight-reducing effects of Liraglutide in a real-world population of type 2 diabetic outpatients. Acta Diabetol [Internet]. 2013 Dec 1;50(6):943–9. Available from: https://doi.org/10.1007/s00592-013-0489-3

42. Rubin DB. Causal Inference Using Potential Outcomes. J Am Stat Assoc [Internet]. 2005 Mar 1;100(469):322–31. Available from: https://doi.org/10.1198/016214504000001880

43. Rosenbaum PR. Modern Algorithms for Matching in Observational Studies. Annu Rev Stat Its Appl [Internet]. 2020 Mar 9 [cited 2021 Nov 4];7(1):143–76. Available from: https://doi.org/10.1146/annurev-statistics-031219-041058

44. Robins JM, Hernán MÁ, Brumback B. Marginal Structural Models and Causal Inference in Epidemiology. Epidemiology [Internet]. 2000;11(5). Available from: https://journals.lww.com/epidem/Fulltext/2000/09000/Marginal_Structural_Models_and_Causal_Inference_in.11.aspx

45. McCaffrey DF, Griffin BA, Almirall D, Slaughter ME, Ramchand R, Burgette LF. A tutorial on propensity score estimation for multiple treatments using generalized boosted models. Stat Med

[Internet]. 2013/03/18 ed. 2013 Aug 30;32(19):3388–414. Available from: https://pubmed.ncbi.nlm.nih.gov/23508673

46. Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivar Behav Res [Internet]. 2011/06/08 ed. 2011 May;46(3):399–424. Available from: https://pubmed.ncbi.nlm.nih.gov/21818162

47. Austin PC, Stuart EA. Estimating the effect of treatment on binary outcomes using full matching on the propensity score. Stat Methods Med Res [Internet]. 2017 Dec 1 [cited 2021 Sep 23];26(6):2505–25. Available from: https://doi.org/10.1177/0962280215601134

48. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. J Stat Softw Vol 1 Issue 3 2011 [Internet]. 2011 Dec 12; Available from: https://www.jstatsoft.org/v045/i03

49. Rubin, D.B. Introduction to multiple imputation. Stat. Anal. Missing Data. 2nd ed. NY: Wiley; 2002. 85–93 p.

50. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Stat Med [Internet]. 2011 Feb 20 [cited 2021 Nov 4];30(4):377–99. Available from: https://doi.org/10.1002/sim.4067

51. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. Educ Psychol. 1974;66(5):688–701.

52. Rosenbaum PR, Rubin DB. The Central Role of the Propensity Score in Observational Studies for Causal Effects. Biometrika [Internet]. 1983;70(1):41–55. Available from: www.jstor.org/stable/2335942

53. Ho DE, Imai K, King G, Stuart EA. MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. J Stat Softw [Internet]. 2011;42(8):1–28. Available from: https://www.jstatsoft.org/v42/i08/

54. Bang H, Robins JM. Doubly Robust Estimation in Missing Data and Causal Inference Models. Biometrics [Internet]. 2005 Dec 1 [cited 2021 Nov 4];61(4):962–73. Available from: https://doi.org/10.1111/j.1541-0420.2005.00377.x

55. Fadini GP, Sciannameo V, Franzetti I, Bottigliengo D, D'Angelo P, Vinci C, et al. Similar effectiveness of dapagliflozin and GLP-1 receptor agonists concerning combined endpoints in routine clinical practice: A multicentre retrospective study. Diabetes Obes Metab [Internet]. 2019 Aug 1 [cited 2021 Apr 28];21(8):1886–94. Available from: https://doi.org/10.1111/dom.13747

56. Frías JP, Guja C, Hardy E, Ahmed A, Dong F, Öhman P, et al. Exenatide once weekly plus dapagliflozin once daily versus exenatide or dapagliflozin alone in patients with type 2 diabetes inadequately controlled with metformin monotherapy (DURATION-8): a 28 week, multicentre, double-blind, phase 3, randomised controlled trial. Lancet Diabetes Endocrinol [Internet]. 2016 Dec 1 [cited 2021 Jul 21];4(12):1004–16. Available from: https://doi.org/10.1016/S2213-8587(16)30267-4

57. Xie Y, Brand JE, Jann B. Estimating Heterogeneous Treatment Effects with Observational Data. Sociol Methodol [Internet]. 2012 Aug;42(1):314–47. Available from: https://pubmed.ncbi.nlm.nih.gov/23482633

58. Austin PC, Grootendorst P, Normand S-LT, Anderson GM. Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: a Monte Carlo study. Stat Med [Internet]. 2007 Feb 20 [cited 2021 May 19];26(4):754–68. Available from: https://doi.org/10.1002/sim.2618

59. Gareth James DW Trevor Hastie, Robert Tibshirani. An introduction to statistical learning : with applications in R [Internet]. New York : Springer, [2013] ©2013; 2013. Available from: https://search.library.wisc.edu/catalog/9910207152902121

60. Altman DG, Bland JM. Missing data. BMJ [Internet]. 2007 Feb 22;334(7590):424. Available from: http://www.bmj.com/content/334/7590/424.abstract

61. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ [Internet]. 2009 Jun 29;338:b2393. Available from: http://www.bmj.com/content/338/bmj.b2393.abstract

62. Rombach I, Jenkinson C, Gray AM, Murray DW, Rivero-Arias O. Comparison of statistical approaches for analyzing incomplete longitudinal patient-reported outcome data in randomized controlled trials. Patient Relat Outcome Meas [Internet]. 2018 Jun 21;9:197–209. Available from: https://pubmed.ncbi.nlm.nih.gov/29950913

63. Mallinckrod† CH, Lane PW, Schnell D, Peng Y, Mancuso JP. Recommendations for the Primary Analysis of Continuous Endpoints in Longitudinal Clinical Trials. Drug Inf J [Internet]. 2008 Jul 1 [cited 2021 May 17];42(4):303–19. Available from: https://doi.org/10.1177/009286150804200402

64. Funk MJ, Westreich D, Wiesen C, Stürmer T, Brookhart MA, Davidian M. Doubly robust estimation of causal effects. Am J Epidemiol [Internet]. 2011/03/08 ed. 2011 Apr 1;173(7):761–7. Available from: https://pubmed.ncbi.nlm.nih.gov/21385832

65. van der Laan, Mark J., Rubin D. Targeted Maximum Likelihood Learning. Int J Biostat [Internet]. 2006;2(1). Available from: https://doi.org/10.2202/1557-4679.1043

66. Groenwold RHH, Donders ART, Roes KCB, Harrell FE Jr, Moons KGM. Dealing With Missing Outcome Data in Randomized Trials and Observational Studies. Am J Epidemiol [Internet]. 2012 Feb 1 [cited 2021 Apr 5];175(3):210–7. Available from: https://doi.org/10.1093/aje/kwr302

67. Lewin A, Brondeel R, Benmarhnia T, Thomas F, Chaix B. Attrition Bias Related to Missing Outcome Data: A Longitudinal Simulation Study. Epidemiology [Internet]. 2018;29(1). Available from: https://journals.lww.com/epidem/Fulltext/2018/01000/Attrition_Bias_Related_to_Missing_Outcome_Data__A.12.aspx

68. Schuler MS, Rose S. Targeted Maximum Likelihood Estimation for Causal Inference in Observational Studies. Am J Epidemiol [Internet]. 2017 Jan 1 [cited 2021 Apr 8];185(1):65–73. Available from: https://doi.org/10.1093/aje/kww165

69. Robins J. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. Math Model [Internet]. 1986 Jan 1;7(9):1393–512. Available from: https://www.sciencedirect.com/science/article/pii/0270025586900886

70. Robins JM, Greenland S, Hu F-C. Estimation of the Causal Effect of a Time-Varying Exposure on the Marginal Mean of a Repeated Binary Outcome. J Am Stat Assoc [Internet]. 1999 Sep 1;94(447):687–700. Available from: https://www.tandfonline.com/doi/abs/10.1080/01621459.1999.10474168

71. Naimi AI, Cole SR, Kennedy EH. An introduction to g methods. Int J Epidemiol [Internet]. 2017 Apr 1 [cited 2021 Apr 30];46(2):756–62. Available from: https://doi.org/10.1093/ije/dyw323

72. Fitzmaurice G, Molenberghs G. Advances in longitudinal data analysis: an historical perspective. Longitud Data Anal. 2009;3–30.

73. Keil AP, Edwards JK, Richardson DB, Naimi AI, Cole SR. The parametric g-formula for time-to-event data: intuition and a worked example. Epidemiol Camb Mass [Internet]. 2014 Nov;25(6):889–97. Available from: https://pubmed.ncbi.nlm.nih.gov/25140837

74. Seaman SR, Vansteelandt S. Introduction to Double Robust Methods for Incomplete Data. Stat Sci Rev J Inst Math Stat [Internet]. 2018;33(2):184–97. Available from: https://pubmed.ncbi.nlm.nih.gov/29731541

75. Gruber S, Laan M van der. tmle: An R Package for Targeted Maximum Likelihood Estimation. J Stat Softw Vol 1 Issue 13 2012 [Internet]. 2012 Nov 16; Available from: https://www.jstatsoft.org/v051/i13

76. Super Learner. Stat Appl Genet Mol Biol [Internet]. 2007;6(1). Available from: https://doi.org/10.2202/1544-6115.1309

77. Pang M, Schuster T, Filion KB, Schnitzer ME, Eberg M, Platt RW. Effect Estimation in Point-Exposure Studies with Binary Outcomes and High-Dimensional Covariate Data - A Comparison of Targeted Maximum Likelihood Estimation and Inverse Probability of Treatment Weighting. Int J Biostat [Internet]. 2016 Nov 1;12(2):/j/ijb.2016.12.issue-2/ijb-2015-0034/ijb-2015-0034.xml. Available from: https://pubmed.ncbi.nlm.nih.gov/27889705

78. Laan M, Dudoit S. Unified Cross-Validation Methodology For Selection Among Estimators and a General Cross-Validated Adaptive Epsilon-Net Estimator: Finite Sample Oracle Inequalities and Examples. In 2003.

79. Luque-Fernandez MA, Schomaker M, Rachet B, Schnitzer ME. Targeted maximum likelihood estimation for a binary treatment: A tutorial. Stat Med [Internet]. 2018/04/23 ed. 2018 Jul 20;37(16):2530–46. Available from: https://pubmed.ncbi.nlm.nih.gov/29687470

80. Higdon R. Generalized Additive Models. In: Dubitzky W, Wolkenhauer O, Cho K-H, Yokota H, editors. Encyclopedia of Systems Biology [Internet]. New York, NY: Springer New York; 2013. p. 814–5. Available from: https://doi.org/10.1007/978-1-4419-9863-7_1197

81. Tin Kam Ho. The random subspace method for constructing decision forests. IEEE Trans Pattern Anal Mach Intell. 1998 Aug;20(8):832–44.

82. Strobl C, Malley J, Tutz G. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychol Methods [Internet]. 2009 Dec;14(4):323–48. Available from: https://pubmed.ncbi.nlm.nih.gov/19968396

83. Rubin D. Multiple imputation for nonresponse in surveys. New York: John Wiley and Sons; 1987.

84. Colombo D, Maathuis MH. Order-Independent Constraint-Based Causal Structure Learning. J Mach Learn Res [Internet]. 2014;15(116):3921–62. Available from: http://jmlr.org/papers/v15/colombo14a.html

85. Broom BM, Do K-A, Subramanian D. Model averaging strategies for structure learning in Bayesian networks with limited data. BMC Bioinformatics [Internet]. 2012 Aug 24;13(13):S10. Available from: https://doi.org/10.1186/1471-2105-13-S13-S10

86. Schouten RM, Lugtig P, Vink G. Generating missing values for simulation purposes: a multivariate amputation procedure. J Stat Comput Simul [Internet]. 2018 Oct 13;88(15):2909–30. Available from: https://doi.org/10.1080/00949655.2018.1491577

87. Mallinckrodt CH, Lane PW, Schnell D, Peng Y, Mancuso JP. Recommendations for the Primary Analysis of Continuous Endpoints in Longitudinal Clinical Trials. Drug Inf J DIJ Drug Inf Assoc [Internet]. 2008 Jul 1;42(4):303–19. Available from: https://doi.org/10.1177/009286150804200402

88. Kreif N, Gruber S, Radice R, Grieve R, Sekhon JS. Evaluating treatment effectiveness under model misspecification: A comparison of targeted maximum likelihood estimation with bias-corrected matching. Stat Methods Med Res [Internet]. 2014/02/12 ed. 2016 Oct;25(5):2315–36. Available from: https://pubmed.ncbi.nlm.nih.gov/24525488

89. Lendle SD, Fireman B, van der Laan MJ. Targeted maximum likelihood estimation in safety analysis. Methods Comp Eff Res-Centered Outcomes Res Effic Eff [Internet]. 2013 Aug 1;66(8, Supplement):S91–8. Available from: https://www.sciencedirect.com/science/article/pii/S0895435613002011

90. Gail MH, Wieand S, Piantadosi S. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. Biometrika. 1984 Dec;71(3):431–44.

91. Sander Greenland, Judea Pearl, James M. Robins. Confounding and Collapsibility in Causal Inference. Stat Sci [Internet]. 1999 Feb 1;14(1):29–46. Available from: https://doi.org/10.1214/ss/1009211805

92. Gerstein HC, Colhoun HM, Dagenais GR, Diaz R, Lakshmanan M, Pais P, et al. Dulaglutide and cardiovascular outcomes in type 2 diabetes (REWIND): a double-blind, randomised placebo-controlled trial. The Lancet [Internet]. 2019 Jul 13 [cited 2021 Jul 20];394(10193):121–30. Available from: https://doi.org/10.1016/S0140-6736(19)31149-3

93. Hernandez AF, Green JB, Janmohamed S, D'Agostino RB Sr, Granger CB, Jones NP, et al. Albiglutide and cardiovascular outcomes in patients with type 2 diabetes and cardiovascular

disease (Harmony Outcomes): a double-blind, randomised placebo-controlled trial. The Lancet [Internet]. 2018 Oct 27 [cited 2021 Jul 20];392(10157):1519–29. Available from: https://doi.org/10.1016/S0140-6736(18)32261-X

94. Kristensen SL, Rørth R, Jhund PS, Docherty KF, Sattar N, Preiss D, et al. Cardiovascular, mortality, and kidney outcomes with GLP-1 receptor agonists in patients with type 2 diabetes: a systematic review and meta-analysis of cardiovascular outcome trials. Lancet Diabetes Endocrinol [Internet]. 2019 Oct 1 [cited 2021 Jul 20];7(10):776–85. Available from: https://doi.org/10.1016/S2213-8587(19)30249-9

95. Avogaro A, Fadini GP, Sesti G, Bonora E, Del Prato S. Continued efforts to translate diabetes cardiovascular outcome trials into clinical practice. Cardiovasc Diabetol [Internet]. 2016 Aug 11;15(1):111. Available from: https://doi.org/10.1186/s12933-016-0431-4

96. Nicolucci A, Candido R, Cucinotta D, Graziano G, Rocca A, Rossi MC, et al. Generalizability of Cardiovascular Safety Trials on SGLT2 Inhibitors to the Real World: Implications for Clinical Practice. Adv Ther [Internet]. 2019 Oct 1;36(10):2895–909. Available from: https://doi.org/10.1007/s12325-019-01043-z

97. Birkeland KI, Bodegard J, Norhammar A, Kuiper JG, Georgiado E, Beekman-Hendriks WL, et al. How representative of a general type 2 diabetes population are patients included in cardiovascular outcome trials with SGLT2 inhibitors? A large European observational study. Diabetes Obes Metab [Internet]. 2019 Apr 1 [cited 2021 Jul 20];21(4):968–74. Available from: https://doi.org/10.1111/dom.13612

98. Boye KS, Riddle MC, Gerstein HC, Mody R, Garcia-Perez L-E, Karanikas CA, et al. Generalizability of glucagon-like peptide-1 receptor agonist cardiovascular outcome trials to the overall type 2 diabetes population in the United States. Diabetes Obes Metab [Internet]. 2019 Jun 1 [cited 2021 Jul 20];21(6):1299–304. Available from: https://doi.org/10.1111/dom.13649

99. Wittbrodt E, Chamberlain D, Arnold SV, Tang F, Kosiborod M. Eligibility of patients with type 2 diabetes for sodium-glucose co-transporter-2 inhibitor cardiovascular outcomes trials: An assessment using the Diabetes Collaborative Registry. Diabetes Obes Metab [Internet]. 2019 Aug 1 [cited 2021 Jul 20];21(8):1985–9. Available from: https://doi.org/10.1111/dom.13738

100. Husain M, Birkenfeld AL, Donsmark M, Dungan K, Eliaschewitz FG, Franco DR, et al. Oral Semaglutide and Cardiovascular Outcomes in Patients with Type 2 Diabetes. N Engl J Med [Internet]. 2019 Aug 29 [cited 2021 Jul 20];381(9):841–51. Available from: https://doi.org/10.1056/NEJMoa1901118

101. Yang X-S. 2 - Mathematical foundations. In: Yang X-S, editor. Introduction to Algorithms for Data Mining and Machine Learning [Internet]. Academic Press; 2019. p. 19–43. Available from: https://www.sciencedirect.com/science/article/pii/B9780128172162000090

102. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing [Internet]. Vienna, Austria; 2019. Available from: https://www.R-project.org

103. Schnell O, Standl E, Cos X, Heerspink HJ, Itzhak B, Lalic N, et al. Report from the 5th cardiovascular outcome trial (CVOT) summit. Cardiovasc Diabetol [Internet]. 2020 Apr 17;19(1):47. Available from: https://doi.org/10.1186/s12933-020-01022-7

104. Sharma A, Pagidipati NJ, Califf RM, McGuire DK, Green JB, Demets D, et al. Impact of Regulatory Guidance on Evaluating Cardiovascular Risk of New Glucose-Lowering Therapies to Treat Type 2 Diabetes Mellitus. Circulation [Internet]. 2020 Mar 10 [cited 2021 Jul 21];141(10):843–62. Available from: https://doi.org/10.1161/CIRCULATIONAHA.119.041022

105. Sciannameo V, Berchialla P, Orsi E, Lamacchia O, Morano S, Querci F, et al. Enrolment criteria for diabetes cardiovascular outcome trials do not inform on generalizability to clinical practice: The case of glucagon-like peptide-1 receptor agonists. Diabetes Obes Metab [Internet]. 2020 May 1 [cited 2021 Jul 21];22(5):817–27. Available from: https://doi.org/10.1111/dom.13962

106. Castellana M, Procino F, Sardone R, Trimboli P, Giannelli G. Generalizability of sodium-glucose co-transporter-2 inhibitors cardiovascular outcome trials to the type 2 diabetes population: a systematic review and meta-analysis. Cardiovasc Diabetol [Internet]. 2020 Jun 13;19(1):87–87. Available from: https://pubmed.ncbi.nlm.nih.gov/32534590

107. Chatterjee S, Davies MJ, Khunti K. What have we learnt from "real world" data, observational studies and meta-analyses. Diabetes Obes Metab [Internet]. 2018 Feb 1 [cited 2021 Jul 21];20(S1):47–58. Available from: https://doi.org/10.1111/dom.13178

108. Stuart EA, Cole SR, Bradshaw CP, Leaf PJ. The use of propensity scores to assess the generalizability of results from randomized trials. J R Stat Soc Ser A Stat Soc [Internet]. 2001 Apr 1;174(2):369–86. Available from: https://pubmed.ncbi.nlm.nih.gov/24926156

109. Hong J-L, Webster-Clark M, Jonsson Funk M, Stürmer T, Dempster SE, Cole SR, et al. Comparison of methods to generalize randomized clinical trial results without individual-level data for the target population. Am J Epidemiol. 2019;188(2):426–37.

110. Stuart EA, Bradshaw CP, Leaf PJ. Assessing the Generalizability of Randomized Trial Results to Target Populations. Prev Sci [Internet]. 2015 Apr 1;16(3):475–85. Available from: https://doi.org/10.1007/s11121-014-0513-z

111. Zinman B, Wanner C, Lachin JM, Fitchett D, Bluhmki E, Hantel S, et al. Empagliflozin, Cardiovascular Outcomes, and Mortality in Type 2 Diabetes. N Engl J Med [Internet]. 2015 Nov 26 [cited 2021 Jul 21];373(22):2117–28. Available from: https://doi.org/10.1056/NEJMoa1504720

112. Zinman B, Inzucchi SE, Lachin JM, Wanner C, Ferrari R, Fitchett D, et al. Rationale, design, and baseline characteristics of a randomized, placebo-controlled cardiovascular outcome trial of empagliflozin (EMPA-REG OUTCOME™). Cardiovasc Diabetol [Internet]. 2014 Jun 19;13(1):102. Available from: https://doi.org/10.1186/1475-2840-13-102

113. Green JB, Bethel MA, Armstrong PW, Buse JB, Engel SS, Garg J, et al. Effect of Sitagliptin on Cardiovascular Outcomes in Type 2 Diabetes. N Engl J Med [Internet]. 2015 Jul 16 [cited 2021 Jul 21];373(3):232–42. Available from: https://doi.org/10.1056/NEJMoa1501352

114. Scirica BM, Bhatt DL, Braunwald E, Steg PG, Davidson J, Hirshberg B, et al. Saxagliptin and Cardiovascular Outcomes in Patients with Type 2 Diabetes Mellitus. N Engl J Med [Internet]. 2013 Oct 3 [cited 2021 Jul 21];369(14):1317–26. Available from: https://doi.org/10.1056/NEJMoa1307684

115. Bonora E, Cataudella S, Marchesini G, Miccoli R, Vaccaro O, Fadini GP, et al. Clinical burden of diabetes in Italy in 2018: a look at a systemic disease from the ARNO Diabetes Observatory. BMJ Open Diabetes Res Care [Internet]. 2020 Jul;8(1):e001191. Available from: https://pubmed.ncbi.nlm.nih.gov/32713842

116. Bonora E, Cataudella S, Marchesini G, Miccoli R, Vaccaro O, Fadini GP, et al. A view on the quality of diabetes care in Italy and the role of Diabetes Clinics from the 2018 ARNO Diabetes Observatory. Nutr Metab Cardiovasc Dis [Internet]. 2020 Oct 30 [cited 2021 Jul 21];30(11):1945–53. Available from: https://doi.org/10.1016/j.numecd.2020.08.018

117. Sciannameo V, Berchialla P, Avogaro A, Fadini GP, Consoli A, Formoso G, et al. Transposition of cardiovascular outcome trial effects to the real-world population of patients with type 2 diabetes. Cardiovasc Diabetol [Internet]. 2021 May 10;20(1):103. Available from: https://doi.org/10.1186/s12933-021-01300-y

118. World Health Organization. Global Health and Aging [Internet]. 2011. Available from: https://www.who.int/ageing/publications/global_health.pdf?ua

119. Maher RL, Hanlon J, Hajjar ER. Clinical consequences of polypharmacy in elderly. Expert Opin Drug Saf [Internet]. 2014 Jan 1;13(1):57–65. Available from: https://doi.org/10.1517/14740338.2013.827660

120. Bourgeois FT, Shannon MW, Valim C, Mandl KD. Adverse drug events in the outpatient setting: an 11-year national analysis. Pharmacoepidemiol Drug Saf [Internet]. 2010 Sep 1 [cited 2020 Jun 4];19(9):901–10. Available from: https://doi.org/10.1002/pds.1984

121. Leendertse AJ, Egberts ACG, Stoker LJ, van den Bemt PMLA, HARM Study Group. Frequency of and Risk Factors for Preventable Medication-Related Hospital Admissions in the Netherlands. Arch Intern Med [Internet]. 2008 Sep 22 [cited 2020 Apr 6];168(17):1890–6. Available from: https://doi.org/10.1001/archinternmed.2008.3

122. Kurt I, Ture M, Kurum AT. Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. Expert Syst Appl [Internet]. 2008 Jan 1;34(1):366–74. Available from: http://www.sciencedirect.com/science/article/pii/S0957417406002855

123. Carroll RJ, Eyler AE, Denny JC. Naïve Electronic Health Record phenotype identification for Rheumatoid arthritis. 2011;189–96.

124. Rahimian F, Salimi-Khorshidi G, Payberah AH, Tran J, Ayala Solares R, Raimmondi F. Predicting the risk of emergency admission with machine learning: Development and validation using linked electronic health records. PLoS Med. 2018;15(11).

125. Roberto J, Solares A, Elisa F, Raimondi D, Zhu Y, Rahimian F, et al. Deep learning for electronic health records : A comparative review of multiple deep neural architectures. J Biomed Inform [Internet]. 2020;101(September 2019):103337. Available from: https://doi.org/10.1016/j.jbi.2019.103337

126. Schmidhuber J. Deep learning in neural networks: An overview. Neural Netw [Internet]. 2015 Jan 1;61:85–117. Available from: https://www.sciencedirect.com/science/article/pii/S0893608014002135

127. Leo Breiman. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). Stat Sci [Internet]. 2001 Aug 1;16(3):199–231. Available from: https://doi.org/10.1214/ss/1009213726

128. Nicholas G. Polson, Vadim Sokolov. Deep Learning: A Bayesian Perspective. Bayesian Anal [Internet]. 2017 Dec 1;12(4):1275–304. Available from: https://doi.org/10.1214/17-BA1082

129. Nguyen P, Tran T, Wickramasinghe N, Venkatesh S. Deepr : A Convolutional Net for Medical Records. 2016;(January 2018).

130. Pham T, Tran T, Phung D, Venkatesh S. DeepCare : A Deep Dynamic Memory Model for Predictive Medicine. 2017;(i):1–27.

131. Sherstinsky A. Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network. Phys Nonlinear Phenom [Internet]. 2020 Mar 1;404:132306. Available from: https://www.sciencedirect.com/science/article/pii/S0167278919305974

132. Choi E, Bahadori MT, Kulas JA, Schuetz A, Stewart WF, Sun J. RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism. 2016.

133. Li Y, Rao S, Roberto J, Solares A, Hassaine A, Ramakrishnan R, et al. OPEN BEHRT : Transformer for Electronic Health Records. 2020;1–12.

134. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) [Internet]. Minneapolis, Minnesota: Association for Computational Linguistics; 2019. p. 4171–86. Available from: https://www.aclweb.org/anthology/N19-1423

135. Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D. Med-BERT: pre-trained contextualized embeddings on large-scale structured electronic health records for disease prediction. 2020.

136. Chollet F, Allaire JJ. Deep learning with R / François Chollet with J.J. Allaire. 1st edition. Shelter Island, NY: Manning Publications; 2018.

137. Liou DR, Liou JW, Liou CY. Learning Behaviors of Perceptron. iConcept Press;

138. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. In: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings [Internet]. 2015. Available from: http://arxiv.org/abs/1412.6980

139. Vaswani A. Attention Is All You Need. 2017;(Nips).

140. Elixhauser A, Steiner C, Palmer L. Clinical Classifications Software (CCS) for ICD-9-CM [Internet]. 2015. Available from: https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp

141. Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems [Internet]. 2015. Available from: http://tensorflow.org/

142. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: Advances in Neural Information Processing Systems 32 [Internet]. Curran Associates, Inc.; 2019. p. 8024–35. Available from: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

143. Mazzola P, Rimoldi SML, Rossi P, Noale M, Rea F, Facchini C, et al. Aging in Italy: The Need for New Welfare Strategies in an Old Country. The Gerontologist [Internet]. 2016 Jun 1 [cited 2021 Apr 16];56(3):383–90. Available from: https://doi.org/10.1093/geront/gnv152

144. Birkhead GS, Klompas M, Shah NR. Uses of Electronic Health Records for Public Health Surveillance to Advance Public Health. Annu Rev Public Health [Internet]. 2015 Mar 18 [cited 2020 Jun 4];36(1):345–59. Available from: https://doi.org/10.1146/annurev-publhealth-031914-122747

145. Skrami E, Carle F, Villani S, Borrelli P, Zambon A, Corrao G, et al. Availability of Real-World Data in Italy: A Tool to Navigate Regional Healthcare Utilization Databases. Int J Environ Res Public Health. 2020;17(1).

146. van Walraven C, Bennett C, Forster AJ. Administrative database research infrequently used validated diagnostic or procedural codes. J Clin Epidemiol [Internet]. 2011 Oct 1;64(10):1054–9. Available from: https://www.sciencedirect.com/science/article/pii/S0895435611000138

147. Johnson EK, Nelson CP. Values and pitfalls of the use of administrative databases for outcomes assessment. J Urol [Internet]. 2013/04/20 ed. 2013 Jul;190(1):17–8. Available from: https://pubmed.ncbi.nlm.nih.gov/23608038

Grazie a Paola,

Daniele, mamma e papà.