School of Medicine and Surgery

PhD Program in Public Health

Curriculum in Health Technology Assessment

Cycle XXXVIII

RANDOM FOREST REGRESSION FOR PREDICTING HEALTHCARE COSTS USING ADMINISTRATIVE DATABASES

Isabella Maria SALA

Registration number: 802991

Tutor:

Prof. Lorenzo Giovanni MANTOVANI

Supervisors:

Prof. Vincenzo BAGNARDI

Dr. Sara CONTI

Dott. Massimo GRANDIS

Prof. Giampiero MAZZAGLIA

Coordinator: Prof. Luigi BADANO

ACADEMIC YEAR 2024/2025

Finanziato dall'Unione europea NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA

Summary

Abstract

Background

Longer life expectancies and increasing prevalence of chronic diseases drive up demand for healthcare services and related costs. In Italy, 32% of people aged 65 and over, and 48% of those over 85 have major chronic conditions and multimorbidity. In 2019, individuals aged 65 and over accounted for 46% of hospital admissions and 60% of pharmaceutical expenditures, highlighting the significant burden of aging on the healthcare system. In terms of costs, population's segments with high prevalence of chronic conditions account for a large portion of healthcare spending. Accurate predictions of future costs for the whole population and for subgroups of patients with high clinical and economic impact is crucial for healthcare planning.

Aim

To predict yearly direct healthcare costs based on data of past Italian National Health Service (NHS) resources utilization for the whole population and for high-impact segments.

Methods

Using administrative healthcare databases, we traced NHS resource utilization (i.e., access to inpatient and outpatient services, drug dispensations) and associated costs for each individual aged ≥18 assisted by the Health Protection Agency of Bergamo between 2011 and 2023. We employed a supervised machine learning approach, specifically the random forest algorithm, to address the prediction problem. We used individual data of 5-year resource utilization of the NHS to predict individual's healthcare costs in the following year. We evaluated different outcome measures: total cost (TC), defined as the sum of all inpatient, outpatient, and drug dispensation costs; total scheduled cost, defined as the sum of scheduled inpatient visits, all outpatient visits, and dispensations; scheduled services cost, defined as the sum of scheduled inpatient and all outpatient visits; and services cost, defined as the sum of all inpatient and outpatient services. Subsequently, individual cost predictions were aggregated to derive total and mean cost estimates both for the whole population and for specific high-impact subgroups, namely patients undergoing dialysis, with type 2 diabetes, heart failure, Parkinson's disease or parkinsonisms, and active neoplasms. The ratio of the difference between predicted and actual cost to actual cost was used as measure of the prediction error (PE). Also, we derived a variability interval for the mean predicted cost based on the 2.5 and 97.5 quantiles of the distribution of the mean costs predicted by each tree of the random forest for subjects included in each group.

Results

Overall, PE values tended to be below zero, indicating a clear tendency to predict lower mean costs than those actually observed. Excluding 2020-2021, when predictions were substantially affected by the COVID-19 pandemic, PE for the whole population were close to zero, indicating a slight discrepancy between predicted and actual costs. In 2023, for example, the predicted total TC was €1,103,322,372 compared to an actual of €1,111,657,382, corresponding to a PE of -0.7%. For high-impact groups, dialysis and diabetic patients showed a consistent temporal trend in PE values, generally negative and not exceeding -10%. For instance, in 2023, the predicted mean TC for dialysis patients was €44,128 [variability interval: €40,421-47,605] compared to an actual of €46,254 (PE = -4.6%), while for diabetic patients it was €3,455 [€3,337-3,571] vs. €3,646 (PE = -5.2%). For patients with heart failure, following a substantial underestimation of mean costs in 2021, a gradual reduction in the deviation between predicted and actual costs was observed in 2022 and 2023. Finally, for patients with Parkinson's disease and active neoplasms, predicted mean costs were markedly underestimated compared to actual costs, with PE values reaching up to -30%, though remaining consistent over time. For instance, in 2023, the predicted mean TC for patients with Parkinson's disease was €3,768 [€3,349-4,434] compared to an actual of €4,698 (PE = -19.8%), while for patients with active neoplasia it was €5,537 [€5,252-5,798] vs. €6,773 (PE = -18.2%).

Conclusions

While a systematic underestimation of mean costs was observed - more pronounced in heterogeneous groups, such as patients with Parkinson's disease or active neoplasms, and less so in homogeneous subgroups, such as dialysis patients - the prediction accuracy remained consistent over time within each subgroup, supporting the robustness and validity of the predictive algorithm.

List of abbreviations

ATC: Anatomical Therapeutic Chemical

BDA: Banca Dati Assistito

CART: Classification and Regression Trees

CKD: Chronic Kidney Disease

CP: Complexity Parameter

HPA: Health Protection Agency

ICD-9-CM: International Classification of Diseases, Ninth Revision, Clinical Modification

IQR: Interquantile Range

MAE: Mean Absolute Error

ML: Machine Learning

MSE: Mean Squared Error

NHS: National Health Service

OECD: Organization for Economic Co-operation and Development

OOB: Out-Of-Bag

PE: Prediction Error

RMSE: Root Mean Squared Error

RHS: Regional Health Services

RSS: Residual Sum of Squares

SC: Services Cost

SSC: Scheduled Services Cost

TC: Total Cost

TSC: Total Scheduled Cost

WHO: World Health Organization

1. INTRODUCTION

1.1 Background and context

The global phenomenon of an increasingly aging population may pose the most significant economic, health, and social challenge of our time. Moreover, a major epidemiological trend in recent years has been the rise in chronic and degenerative diseases. Over the past decades, the proportion of individuals aged 65 and older has doubled on average across the Organization for Economic Co-operation and Development's (OECD) member countries, growing from less than 9% in 1960 to 18% in 2021. (1) This demographic shift can be attributed to declining fertility rates and longer life expectancies, resulting in a higher proportion of elderly individuals within OECD's member populations. Projections indicated that this trend would continue, with the share of the population aged 65 and over expected to rise from 18% in 2021 to 27% by 2050. Notably, more than one-third of the population in Korea, Japan, Italy, Greece, and Portugal is projected to be aged 65 and over by 2050. Moreover, individuals aged 80 and over are expected to rise from 4.8% to 9.8% across OECD countries between 2021 and 2050, with the aforementioned countries anticipating more than one in eight people to be in this age group by mid-century. These demographic shifts underscore the critical importance of adapting healthcare systems to meet the evolving needs of an older population while handling the related rising costs. As people age, they are more likely to need acute or ongoing medical care. (1)

Today, the prevalence of chronic diseases is notably higher among the elderly worldwide. (2) For instance, in the United States, six out of ten adults suffer from at least one chronic disease, and four out of ten have multiple chronic conditions. (3) Similarly, in Germany, 46% of the adult population reports at least one chronic condition. (4) These chronic diseases present massive economic challenges for individuals, societies, and healthcare systems, accounting for a substantial portion of medical expenditures. In the United States alone, they account for 86% of total healthcare costs, reflecting a similar trend worldwide. (5,6)

The presence of multiple chronic conditions has profound implications for healthcare costs and utilization. Bälher et al. (7) showed that the mean total healthcare costs were 5.5 times higher in multimorbid compared to non-multimorbid elderly patients. An increase in costs with increasing number of chronic conditions was also reported by Glynn et al. (8), and, according to a systematic review by Lehnert et al. (9) total healthcare expenditures rose almost exponentially with the number of chronic conditions in different studies. In OECD countries, it is estimated that individuals aged over 65 account for 40% to 50% of total healthcare expenditures, with per capita healthcare costs three to five times higher than those of people under 65. (10)

1.2 Population aging in Italy

In Italy, the effects of population aging have become increasingly evident over the past two decades. From 2004 to 2024, the average population age has grown from 42.3 to 46.6 years. Italian residents aged 65 and over have increased by more than 3 million, now exceeding 14 million, i.e., an increase of 5.1 percentage points compared to 2004. Particularly noteworthy is the fact that more than half of this group is at least 75 years old, marking a 3.8 percentage point increase over the same period. (11) At the same time, Italy's mortality rates have substantially dropped, contributing to six-month gain in life expectancy at birth, which now stands at 83.1 years. This places Italy among the countries with the highest life expectancy globally. While this represents a remarkable achievement, it also poses significant challenges. It requires a concerted effort to ensure that the additional years gained are lived in good health and that the healthcare system can manage a growing burden of chronic and complex conditions. The number of individuals living with long-term comorbidities is expected to rise sharply. (11,12) Moreover, these demographic changes are further compounded by marked regional disparities in health outcomes, driven by a complex interplay of demographic dynamics, healthcare service quality, and socioeconomic inequalities. Northern regions consistently outperform southern ones across different outcomes, such as higher life expectancy, healthier life expectancy, and lower years of life lost, which is partially explained by more advanced healthcare infrastructures and greater investment in resources. (12) The demographic transition toward an aging society can be effectively illustrated through a population pyramid, also known as the age-sex pyramid, which visually represents the age distribution of a population (**Figure 1.1**).

In Italy, 32% of people aged 65 and over, and 48% of those over 85 have major chronic conditions and multimorbidity. The most common conditions include osteoarthritis (48%) and hypertension (47%), followed by heart disease (19%) and diabetes (17%). Non-communicable diseases affect around 24 million people across all stages of life. However, the elderlies are the most vulnerable, with chronic diseases affecting more than 85% of those aged 75 and over. Women, especially those over the age of 55, are also particularly affected. (13–15) Among the degenerative diseases affecting the mental health of older adults, dementia represents a major public health priority. In response, the World Health Organization (WHO) launched the "Global Action Plan on the Public Health Response to Dementia - 2017-2025" to encourage countries to improve the quality of life for people with dementia and their caregivers. In Italy, it is estimated that in 2019, approximately 600,000 people over the age of 65 living at home were affected by dementia or Alzheimer's disease, i.e., about 4.2%

of the elderly population. Prevalence is higher among women (5.1%) than men (3%), and it rises sharply among those aged 85 and over, reaching 15.4% in women and 14% in men. (14)

In terms of healthcare service utilization, the Istituto Nazionale di Statistica (ISTAT) data showed that 90% of people aged 65 and over consulted their General Practitioner at least once in 2019. Furthermore, 48% underwent specialist examinations, and around 14% accessed rehabilitation services at least once. Individuals aged 65 and over accounted for 46% of hospital admissions and 60% of pharmaceutical expenditures, highlighting the significant burden of aging on the healthcare system. (13,15,16) Health problems and loss of autonomy lead to increased healthcare consumption among older adults, particularly after the age of 75. At this stage, rates of specialist consultations and hospital admissions are approximately 1.5 times higher than the national average. There is also a significant rise in the demand for diagnostic tests and rehabilitation services, especially among those with severe difficulties in performing personal care activities. While this increase in healthcare use affects both men and women, the impact is more pronounced among men. (14)

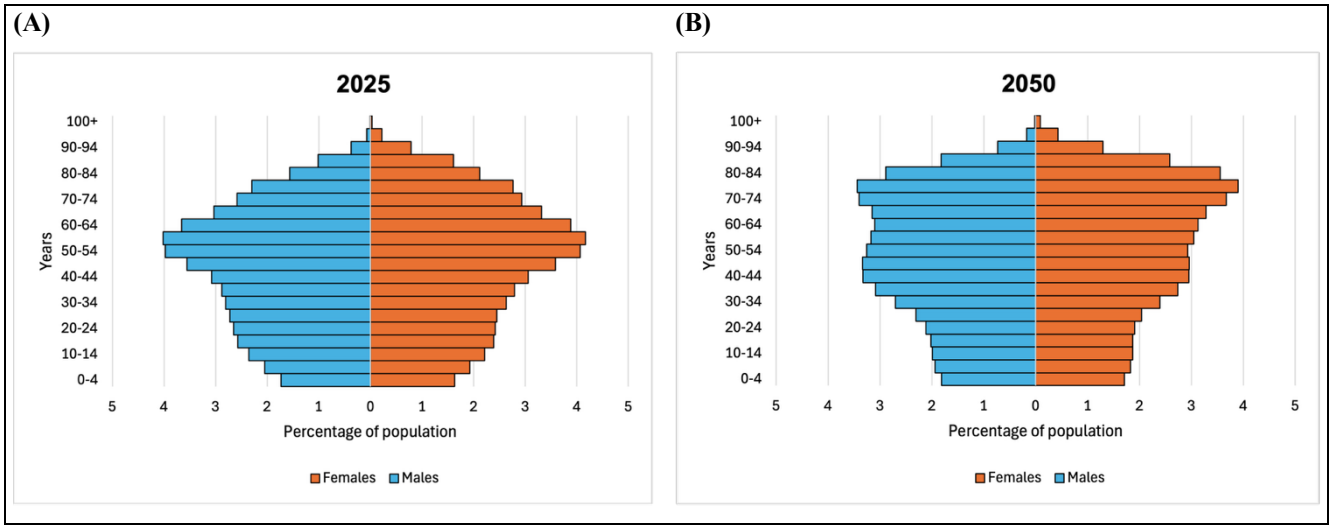


Figure 1.1 Italian resident population pyramids by sex and age. Panel A, population on 1 January 2025; panel B, population projection on 1 January 2050. Data source: Istituto Nazionale di Statistica (ISTAT), accessed at https://demo.istat.it/ on 25 June 2025.

1.3 Analysis of healthcare demand

Given the global trend of aging populations and the increasing prevalence of chronic diseases, the demand for healthcare continues to grow, making the rationalization and sustainability of healthcare costs an increasingly critical challenge. (17) In this context, the development of patient-centered healthcare represents a paradigm shift with significant potential to enhance the quality of care. (18) While efforts are being made to develop personalized care models tailored to individual needs, there remains a pressing need for a comprehensive analysis of healthcare demand of the population. This

includes accurately estimating individual needs to inform strategic decisions aimed at effectively addressing health demands. (19,20) A key approach to achieving this goal is segmenting the population into relatively homogeneous groups based on healthcare needs, allowing for a more efficient analysis and management of healthcare demand and ensuring optimal patient care. (21) Literature offers different models of population segmentation based on healthcare needs. Notable examples include the "Bridges to Health" model by Lynn et al. (21), the segmentation framework applied to the Lombardy population by Madotto et al. (22), and the population-based segmentation model implemented in British Columbia (23).

Italian regulatory frameworks also emphasize the importance of developing statistical models for population stratification, risk factor monitoring, and the integrated management of chronic diseases and complex conditions. Ministerial Decree 70/2015 of the Ministry of Health defines essential levels of care, in conjunction with Law 77/2020 - converting Decree Law 34/2020 on urgent health, labor, economic, and social measures in response to the COVID-19 emergency - and Ministerial Decree 77/2022, establishes models and standards for the development of territorial healthcare within the Italian National Health Service (NHS). Specifically, Section 3 of Annex 1 to Ministerial Decree 77/2022, titled “New Models and Standards for the Development of Territorial Care in the National Health Service”, underscores the necessity of analyzing demographic conditions across different territories to assess healthcare needs and plan appropriate interventions. Population segmentation is identified as the most suitable and comprehensible tool for implementing this regulatory framework.

The accuracy and effectiveness of healthcare demand models depend significantly on the availability and quality of data. (24,25) In this context, administrative healthcare databases of the NHS or Regional Healthcare Services (RHS) represent an invaluable resource, providing vast amounts of high-quality, easily accessible, and relatively inexpensive data on the general population, including demographic, clinical, and economic information. (26–28) These databases enable researchers and policymakers to segment the population into subgroups with similar characteristics, needs, costs, and trends over time. Such segmentation facilitates the development and implementation of targeted interventions tailored to each subpopulation. (29)

Looking ahead, the development of descriptive and predictive statistical models for segmentation will be beneficial in optimizing resource allocation and continuously monitoring the characteristics of different population groups, resulting in a more effective and sustainable healthcare system in the long run.

1.4 Healthcare cost prediction

In the framework of an aging society, such as the Italian population, the growing proportion of elderly people is likely to influence the relationship between health, healthcare, and economic growth. Population ageing is expected to have far-reaching effects across many different areas, including economics and healthcare demand. (30) Predicting future healthcare expenditures is essential for effective planning, enabling organizations to anticipate patient needs, allocate resources efficiently, and ensure financial sustainability.

Accurately predicting healthcare costs is a significant challenge since healthcare costs reflect the evolution of health over time, which is influenced by a variety of factors including social demographics, medical history, environmental conditions, genetics, and unforeseen events such as accidents. Traditionally, healthcare cost prediction has relied on statistical models and rule-based techniques, both of which require substantial domain expertise to be effective. (31) Although statistical models, particularly multiple regression models, are powerful tools for capturing relationships between predictors and an outcome, they face significant challenges. First, the use of multiple independent variables usually introduces multicollinearity, resulting from high correlations among predictors, which can distort estimates and reduce model reliability. Moreover, in high-dimensional contexts, the relationship between predictors and the dependent variable, such as healthcare costs, is seldom purely additive, as variables often interact in complex and nonlinear ways. Traditional models, such as linear regression, are limited in their ability to capture these interactions unless they are explicitly defined. This limitation makes classical models poorly suited for prediction tasks in high-dimensional settings. Also, the performance of these models is constrained by the intrinsic skewness of healthcare data. For instance, cost data often display a spike at zero, heavily right-skewed distributions, and frequent extreme outliers. (32) These characteristics challenge the efficiency of traditional statistical models, and although various advanced statistical techniques have been proposed to address these distributional complexities, they generally fall short of supervised machine learning methods in terms of predictive accuracy and robustness. These methods have the potential to uncover hidden patterns and relationships within vast healthcare datasets, enabling more accurate and dynamic predictions. (33–36)

1.5 Study objective

The aim of this study was to predict the direct healthcare cost based on data of past NHS resource utilization for the whole population and for population segments with high clinical and economic impact. Briefly, we used a supervised machine learning approach, namely random forest regression,

to address the prediction problem. The considered input variables were the available information on individuals, including demographic features (such as age and sex) and their access history to the NHS (such as access to inpatient and outpatient services and dispensed medications), in a given period of time. The outcome of interest (i.e., the object of our prediction) was the individuals' healthcare cost of the subsequent year.

2. MATERIALS AND METHODS

2.1 Data source

Italy has a tax-funded NHS that offers universal coverage, organized across three levels: national, regional (comprising 21 regions and autonomous provinces), and local with an average of 10 Health Protection Agencies (HPA) per region. For this study, data were retrieved from the administrative healthcare databases of the HPA of Bergamo, located in Northern Italy. These databases collect pseudo-anonymized longitudinal data at the patient level, detailing healthcare services reimbursed by the NHS and provided to all residents registered with a general practitioner within the HPA's jurisdiction. For each subject, the main information extracted from each database included:

- Demographic registry (NAR flow): age, sex, deaths and migrations, and registration periods in administrative databases.
- Hospital discharge database (SDO flow): admission and discharge dates, hospital departments (at admission, during hospitalization, and at discharge), one primary diagnosis and five secondary diagnoses, six procedures (International Classification of Diseases, Ninth Revision, Clinical Modification [ICD-9-CM]) at discharge.
- Pharmacy claims: dispensed medications (active ingredient, Anatomical Therapeutic Chemical [ATC] code, dispensed date), sourced from hospital pharmacies (File-F flow) and community pharmacies (pharmaceutical territorial flow).
- Outpatient services database (28-SAN flow): booking and service provision dates, service type (e.g., specialistic visit, laboratory exam).

Additionally, information on chronic conditions was retrieved from the Banca Dati Assistito (BDA). The BDA, implemented in the Lombardy Region since 2004, enables the monitoring of chronic diseases within the assisted population by classifying patients according to their clinical characteristics and healthcare needs. (37,38) This classification is achieved through the integration of data from various health-administrative sources, including disease-specific exemptions, pharmaceutical records, hospital admissions, specialist services, admissions to residential care facilities, and psychiatric care. The classification algorithm has been progressively updated over time, both to include previously excluded conditions, such as those related to psychiatric disorders and mental health issues, and to align with newer chronic disease stratification models, for instance the care model introduced by Regional Decree DGR 6164/2017 in 2016. The BDA includes all subjects assisted by the HPA who satisfy the classification algorithms. For each individual, the specific codes corresponding to their chronic diseases, as determined by the classification algorithm, are recorded. The total number of chronic conditions per individual may exceed ten. Among all identified chronic conditions for each patient, one is designated as the primary condition. As of today, the BDA

identifies 53 specific chronic conditions. The algorithms used to identify each condition are described in detail elsewhere. (38,39)

The study was conducted in accordance with the Declaration of Helsinki. All procedures for record linkage between demographic and healthcare data were carried out ensuring anonymity, following the management rules of regional information systems and fully complying with current privacy legislation (General Data Protection Regulation - GDPR, European Regulation 2016/679). Data were pseudo-anonymized to perform the record linkage procedures between the relevant healthcare data flows. No personal identifiers were provided to the researchers, and personal data were analyzed in an anonymous manner.

2.2. Study population and variables

In this study, we used data of the administrative health databases of the HPA of Bergamo from the years 2011 to 2023. During the study period, there was a total of 1,388,639 individuals covered by the HPA of Bergamo for at least one year. To be included in the cost prediction analysis, subjects had to be aged 18 or older and assisted by the HPA for at least 6 consecutive years (n = 939,320).
For the study population, annual access to inpatient and outpatient services, annual dispensed medications, and total annual costs based on healthcare service tariffs were traced from administrative healthcare databases. For each subject, for each year, we linked the unique anonymous identification code across all databases (i.e., NAR flow, SDO flow, 28-SAN flow, File-F flow, and pharmaceutical territorial flow). Every row in the databases (except for the NAR flow with demographical data) reports the information on a single service (i.e., a single hospital admission in the SDO flow or a single outpatient visit in the 28-SAN flow), and on a single drug dispensation in the File-F and territorial flows. An individual is allowed to have several rows in each database. Since we needed to transform data from service or drug dispensation to a single record reporting the individual history of healthcare service usage, we synthesized the services and drug dispensations for each subject, for each year:

1. We traced the number of hospitalizations (scheduled and urgent), the number of scheduled hospitalizations, and the number of day-hospital admissions, both overall and stratified by class of main diagnosis classified according to the WHO classification based on the ICD-9-CM code range (40): 1) Infectious and parasitic diseases (001-139), 2) Neoplasms (140-239), 3) Endocrine, nutritional and metabolic diseases, and immunity disorders (240-279), 4) Diseases of the blood and blood-forming organs (280-289), 5) Mental, behavioral and

neurodevelopmental disorders (290-319), 6) Diseases of the nervous system and sense organs (320-389), 7) Diseases of the circulatory system (390-459), 8) Diseases of the respiratory system (460-519), 9) Diseases of the digestive system (520-579), 10) Diseases of the genitourinary system (580-629), 11) Complications of pregnancy, childbirth, and the puerperium (630-679), 12) Diseases of the skin and subcutaneous tissue (680-709), 13) Diseases of the musculoskeletal system and connective tissue (710-739), 14) Congenital anomalies (740-759), 15) Certain conditions originating in the perinatal period (760-779), 16) Symptoms, signs, and ill-defined conditions (780-799), 17) Injury and poisoning (800-999), 18) Supplementary classification of factors influencing health status and contact with health services (V01-V91);

2. We traced the number of scheduled outpatient visits (i.e., not conducted following an emergency room visit), both overall and stratified by the medical specialty in which the visit was carried out: 1) Anesthesiology (01), 2) Cardiology (02), 3) General surgery (03), 4) Plastic surgery (04), 5) Vascular surgery - Angiology (05), 6) Dermatology and Venereology (06), 7) Diagnostic imaging: Nuclear Medicine (07), 8) Diagnostic imaging: Diagnostic Radiology (081), 9) Diagnostic imaging: Interventional (082), 10) Diagnostic imaging: Ultrasound (083), 11) Diagnostic imaging: MRI (084), 12) Diagnostic imaging: CT Scan (085), 13) Endocrinology (09), 14) Gastroenterology - Digestive surgery and Endoscopy (010), 15) Physical medicine and Rehabilitation (012), 16) Nephrology (0131), 17) Dialysis and inpatient dialysis (0131 e 0132), 18) Neurosurgery (014), 19) Neurology (015), 20) Ophthalmology (016), 21) Dentistry and Maxillofacial surgery (017), 22) Oncology (018), 23) Orthopedics and Traumatology (019), 24) Obstetrics and Gynecology (020), 25) Otorhinolaryngology (021), 26) Pulmonology (022), 27) Psychiatry (023), 28) Radiotherapy (024), 29) Urology (025), 30) Other (026).
3. We traced the number of prescriptions for laboratory tests, overall and stratified by laboratory test bundles. We defined the laboratory test bundles by identifying the groups of laboratory tests most frequently prescribed together within the same prescriptions. Specifically, we selected tests prescribed with high frequency (i.e., at least 10,000 annual prescriptions), and kept only the prescriptions that included at least one of these tests. To identify tests that are often prescribed together, we calculated the correlation between all possible pairs of tests within the selected prescriptions. Next, we selected prescriptions containing at least one highly correlated test pair (i.e., Pearson's correlation >0.5). Among these, we further selected combinations of tests prescribed together within the same prescription with high frequency (i.e., at least 500 times in a year). This led us to identify the 102 most frequently co-prescribed

laboratory test combinations. Subsequently, two clinicians independently reviewed these 102 prescriptions and assigned a potential diagnostic question to each test (or subgroup of tests) typically prescribed within each prescription. In cases where discrepancies arose between the clinicians' classifications, they reached consensus through discussion. This process resulted in the definition of 25 laboratory test bundles: 1) Lipid profile: HDL cholesterol + total cholesterol + triglycerides, 2) Glycemic profile: glucose + glycated hemoglobin, 3) Metabolic profile: lipid profile + glycemic profile, 4) Protein profile: protein electrophoresis + total proteins, 5) Electrolyte profile: potassium + sodium + total calcium (*optional*) + chloride (*optional*), 6) Bone metabolism: total calcium + alkaline phosphatase + vitamin D + inorganic phosphate (*optional*) + phosphorus (*optional*), 7) Renal function: creatinine + uric acid + urea + creatinine clearance (*optional*) + urine chemical-physical and microscopic exam (*optional*), 8) Renal damage indices: urinary microalbumin + urine chemical-physical and microscopic exam, 9) Liver function: GPT transaminase (ALT) + GOT transaminase (AST) + gamma-GT + cholinesterase (*optional*) + alkaline phosphatase (*optional*), 10) Liver markers: alpha-fetoprotein + ferritin, 11) Cholestasis indices: total bilirubin (with reflex testing) + alkaline phosphatase + gamma-GT, 12) Pancreatic function: amylase + lipase, 13) Thyroid function: TSH + FT4 + FT3 + reflex TSH (*optional*), 14) Thyroid antibody testing: anti-thyroperoxidase antibodies + anti-thyroglobulin antibodies, 15) Autoantibody screening: ANA, 16) Inflammatory markers: C-reactive protein + erythrocyte sedimentation rate, 17) Coagulation profile: prothrombin time + activated partial thromboplastin time + functional fibrinogen (*optional*), 18) Female hormonal profile: FSH + LH + prolactin (*optional*) + estradiol (*optional*), 19) Lymphocyte typing: subpopulation typing of blood cells with single antibody, 20) Infectious agent detection: microscopic bacterial search + nucleic acid extraction + semen culture (*optional*), 21) Infectious disease antibody testing: toxoplasma IgG and IgM antibodies (*optional*) + treponema pallidum IgG antibodies (*optional*) + cytomegalovirus (CMV) IgG antibodies (*optional*) + CMV IgM antibodies (*optional*) + hepatitis B virus: anti-HBs antibodies (*optional*) + hepatitis B virus: HBsAg (*optional*) + hepatitis C virus: antibody detection (*optional*) + HIV 1-2: antibody detection (*optional*), 22) Cell damage markers: lactate dehydrogenase + creatine kinase (*optional*) + myoglobin (*optional*), 23) Cardiac damage markers: myoglobin + CK-MB, 24) Complete blood count: complete blood count with differential, 25) Tumor markers: CA 125 (*optional*) + CA 15.3 (*optional*) + CA 19.9 (*optional*) + CEA (*optional*).

4. We traced the number of pharmaceutical dispensations for each drug class of interest. If multiple drugs from the same class were prescribed within the same prescription, we counted

it as a single dispensation. The following drug classes of interest were identified based on their ATC codes: 1) Antacids or anti-reflux drugs (A02), 2) Antidiarrheals (A07), 3) Antidiabetics: insulins (A10A), 4) Oral antidiabetics (A10B), 5) New generation oral antidiabetics - GLT1/SGLT2 (A10BJ, A10BK), 6) Anticoagulants or antithrombotics (B01), 7) Antihemorrhagics (B02), 8) Antianemics (B03), 9) Cardiac glycosides (C01A), 10) Antiarrhythmics (C01B), 11) Vasodilators (C01D), 12) Antihypertensives (C02), 13) Diuretics (C03), 14) Beta blockers (C07), 15) Calcium channel blockers (C08), 16) Renin-angiotensin system agents (C09), 17) Lipid-lowering agents (C10), 18) Antipsoriatics (D05), 19) Drugs for prostatic hypertrophy (G04C), 20) Systemic corticosteroids (H02), 21) Thyroid medications (H03), 22) Antibiotics (J01), 23) Antifungals (J02), 24) Antivirals (J05), 25) Antineoplastic agents (L01AA), 26) Antineoplastic agents (L01AX), 27) Antineoplastic agents (L01BA), 28) Antineoplastic agents (L01BB), 29) Antineoplastic agents (L01BC), 30) Antineoplastic agents (L01CA), 31) Antineoplastic agents (L01CD), 32) Antineoplastic agents (L01CE), 33) Antineoplastic agents (L01DA), 34) Antineoplastic agents (L01DB), 35) Antineoplastic agents (L01EA), 36) Antineoplastic agents (L01EB), 37) Antineoplastic agents (L01EC), 38) Antineoplastic agents (L01ED), 39) Antineoplastic agents (L01EE), 40) Antineoplastic agents (L01EF), 41) Antineoplastic agents (L01EG), 42) Antineoplastic agents (L01EH), 43) Antineoplastic agents (L01EJ), 44) Antineoplastic agents (L01EK), 45) Antineoplastic agents (L01EL), 46) Antineoplastic agents (L01EM), 47) Antineoplastic agents (L01EX), 48) Antineoplastic agents (L01FA), 49) Antineoplastic agents (L01FB), 50) Antineoplastic agents (L01FC), 51) Antineoplastic agents (L01FD), 52) Antineoplastic agents (L01FE), 53) Antineoplastic agents (L01FF), 54) Antineoplastic agents (L01FG), 55) Antineoplastic agents (L01FX), 56) Antineoplastic agents (L01XA), 57) Antineoplastic agents (L01XB), 58) Antineoplastic agents (L01XD), 59) Antineoplastic agents (L01XF), 60) Antineoplastic agents (L01XG), 61) Antineoplastic agents (L01XJ), 62) Antineoplastic agents (L01XK), 63) Antineoplastic agents (L01XL), 64) Antineoplastic agents (L01XX), 65) Antineoplastic agents (L01XY), 66) Immunomodulatory antineoplastics: endocrine therapies (L02), 67) Immunostimulants (L03), 68) Immunosuppressants (L04), 69) Anti-inflammatory drugs (M01), 70) Anti-gout medications (M04), 71) Drugs for the musculoskeletal system (M05), 72) Anesthetics (N01), 73) Analgesics (N02A, N02B), 74) Anti-migraine medications (N02C), 75) Antiepileptics (N03), 76) Anti-Parkinson drugs (N04), 77) Psycholeptics (N05), 78) Antidepressants (N06A), 79) Anti-dementia drugs (N06D), 80) Respiratory system drugs (R01, R02, R03, R04, R05, R07), 81) Systemic antihistamines (R06), 82) Anti-glaucoma drugs (S01E).

For the drug classes A02, A07, J01, J02, M04, N01, N02C, and R06, we defined subjects as *non-users* if they had no dispensations during the year, and as *occasional or intermittent users* if they had 1 or more dispensations. For all other drug classes, we defined subjects as *non-users* if they had no dispensations during the year, *occasional users* if they had at most 1 dispensation, and *continuous users* if they had 2 or more dispensations during the year.

2.3 Statistical methods

2.3.1 Learning from data

Machine learning (ML) is the study of computer algorithms that automatically improve with experience. It is the link between statistics, which aims to learn associations from data, and computer science, which focuses on efficient computing algorithms, motivated by the computational challenges of creating statistical models from enormous datasets. (41,42)

ML techniques are traditionally divided into two broad categories: supervised learning and unsupervised learning. Supervised learning methods are widely used because they are very similar to the human approach of learning, that involves experiments and in which mistakes lead to knowledge gain. Two fundamental tasks of supervised learning are classification and regression. A supervised learning algorithm aims to infer a function f: $X \rightarrow Y$ from a sample of data, known as *training set*, which contains pairs of inputs (i.e., features or predictors, $x_i \in X$) and output (i.e., response or target variable, $y_i \in Y$) data. Typically, the predictors x_i can be both continuous and discrete, while y_i is quantitative for regression problems and qualitative for classification problems. When the output values are discrete, the function f is called a classifier and the output values are called classes. In cases where the output is continuous, f is called a predictor. The true function $f(x)$, which assigned the output value y to data point x, would be the best classifier or predictor, but it is obviously unknown. Therefore, the goal of supervised learning consists of choosing a single hypothesis function $h(x)$ that best estimates the true function $f(x)$. The quality of fit is usually measured by a loss function, $L(y, h(x))$, that returns a numerical value indicating the difference between y and $h(x)$. (43–45)

In contrast to supervised learning, unsupervised learning is the ML framework in which algorithms only use unlabeled data to identify patterns. Unsupervised learning models are implemented for two main tasks: clustering and dimensionality reduction. (46)

In this work we focused on two methods that pertains to the field of supervised learning, namely decision trees and random forests.

2.3.2 Decision trees

The basic idea behind a decision tree is a recursive partitioning of data into subsets that are increasingly homogeneous with respect to the target variable. **Figure 2.1** illustrates the structure of a decision tree. The root node, at the beginning of a decision tree, represents the entire sample being analyzed and it gets further divided into two or more homogenous sub-nodes. The process of dividing a node into sub-nodes is called splitting. A sub-node is referred to be a decision node when it splits into further sub-nodes. The nodes of the tree that do not split are called leaves or terminal nodes. In each node, the algorithm applies a condition or rule, which is a Boolean question that focuses on a single feature.

More formally, let $\mathbf{x}=(x_1,x_2,\ldots,x_n)$ represent a vector of feature values, and y the target variable. The decision tree starts at a root node, which considers the entire sample of data D. At each internal node, the algorithm selects a feature $x_i \in \mathbf{x}$ and a threshold t for numerical features or a subset C for categorical features that best splits the data. The selection criterion for the best split depends on the decision tree algorithm that is applied. For numerical features, the decision rule at the node is expressed as $x_i \leq t$. The sample D is then split into two subsets, for example D_1 and D_2:

$D_1 = \{(x,y) \in D \mid x_i \leq t\}$, $D_2 = \{(x,y) \in D \mid x_i > t\}$

For categorical features, the rule takes the form $x_i \in C$. After applying the decision rule, the process recurses on the resulting subsets, treating each as a new dataset to be further split. The splitting continues until a stopping criterion is met. At each leaf node, where no further splitting occurs, each observation is assigned a predicted value for the outcome $\hat{y}$. In classification, this outcome is typically the majority class of the observations in that node, while in regression, it is the average or median value of the outcome y for the observations in the node.

Many different decision tree algorithms have been introduced, including C5 (47,48), ID3 (49) and Classification and Regression Trees (CART) (50). We will use the former method that was proposed by Breiman et al. in 1984. The CART algorithm was also implemented in the package *rpart* of the software R. (51)

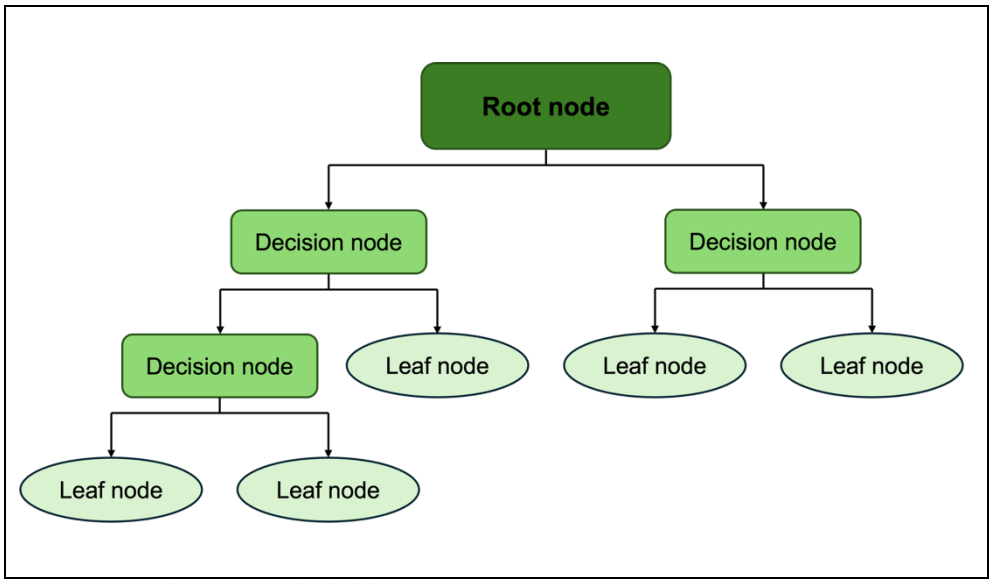


Figure 2.1 Basic structure of a decision tree.

2.3.2 Classification and Regression Tree (CART)

As one can tell from the name, CART is a decision tree algorithm that is used for both classification and regression tasks. The feature space - that is, the set of possible values for the predictors x_1, x_2, …, x_p - is segmented into a number *j* of distinct and non-overlapping segments or regions R_1, R_2, …, R_j. Since it is computationally unfeasible to consider every possible partition of the feature space into *j* regions, the CART algorithm recursively creates a series of binary splits using a top-down, greedy approach that is known as recursive binary splitting. The approach is top-down because it starts at the top of the tree and then progressively divides the feature space. Each split is identified by two new branches further down on the tree. It is greedy because the best split is made at each step of the tree-building process, instead of looking ahead and selecting a split that would result in a better tree at a later step. (52,53)

2.3.2.1 CART for regression

When the target variable *y* is continuous, the algorithm is named *regression tree* and its goal is to predict the value of *y*.

Let D represent the set of data $\{\boldsymbol{x}_i, y_i\}_{i=1}^{N}$, where $\boldsymbol{x} = (x_1, \dots, x_d)$ is the input vector (i.e., vector of features, that can be both numeric or categorical) and $y_i \in Y$ (i.e., numeric target variable). The objective is to identify a function (or predictor) $f: X \to Y$ that minimizes the error $\sum_i |f(\boldsymbol{x}_i) - y_i|^2$.

In the case of the regression tree, such function assigns a constant value for each terminal node. Thus, f is a piecewise constant function.

To create a tree, the CART algorithm at each node selects one feature and splits the node based on the values of that single variable. At the top of the tree, the predictor x_j and the variable cut-off value s are selected such that splitting the predictor space into two regions $R_1(j,s) = \{x|x_j < s\}$ and $R_2(j,s) = \{x|x_j \geq s\}$ minimizes the following cost function: $\sum_{x_i \epsilon R_1(j,s)}(y_i - \hat{y}_{R_1})^2 + \sum_{x_i \epsilon R_2(j,s)}(y_i - \hat{y}_{R_2})^2$, where $\hat{y}_{R_j}$ is the sample mean of y within the j^{th} region, representing the predicted response for the resulting node. This function corresponds to the residual sum of squares (RSS). The process is then repeated, seeking for the best feature and cut-off value to further split the data and minimize the RSS within each of the previously selected regions. The process keeps on until a stopping criterion is met.

Stopping criteria

As stated above, the recursive binary splitting stops when a prespecified criterion is met. Multiple ways for stopping the algorithm exist. One can decide to set *i)* the minimum number of observations in a node that are required in order to make a split, *ii)* the minimum number of observation in a leaf, i.e., if the split leads to a terminal node with too few observations, than the split should not be done, *iii)* the maximum depth of the tree, with the root node counted as depth 0, and *iv)* the threshold complexity parameter, i.e., if the split leads to an increase in the R-squared less than the complexity parameter value, than the split should not be done. (51)

Tree pruning

The procedure outlined above may yield strong predictive performance on the training set. However, it is prone to overfitting - that is, the model may learn patterns that are too specific to the training data, rather than general rules. This often results in poor performance on the test set, which is an independent portion of data used to evaluate how well the model generalizes to new, unseen data. An approach aimed at avoiding overfitting involves growing the smallest tree, stopping the algorithm as soon as the decrease in the RSS due to the current split falls below some threshold. But this strategy is too short-sighted, as an apparently pointless split early in the tree could be followed by a substantial decrease in RSS after subsequent splits. A more effective approach is to first grow a large initial tree T_{max} and then prune it to obtain a subtree T_{pruned}. This is accomplished through *cost-complexity pruning* (or *weakest link pruning*), that is described in detail in the following paragraph.

$T \subset T_{max}$ is defined as any tree that can be derived by pruning T_{max}, specifically by collapsing a subset of its non-terminal nodes. Terminal nodes are indexed by t, with each node t representing a region R_t. Let $\ddot{T}$ represent the number of terminal nodes in T. For each terminal node t, let $N_t = \#\{x_i \in R_t\}$ denote the number of observations in R_t, $\hat{y}_t = \frac{1}{N_t}\sum_{x_i \epsilon R_t} y_i$ the sample mean of the outcome y in the node t, and $Q_t = \frac{1}{N_t}\sum_{x_i \in R_t}(y_i - \hat{y}_t)^2$ the mean squared error. The cost function is then defined as $C(T) = \sum_{t \in \ddot{T}} N_t Q_t$. Finally, the *cost-complexity criterion* is given by $C_\alpha(T) = C(T) + \alpha|\ddot{T}|$, where α is defined as a non-negative tuning (or complexity) parameter. The goal is to identify, for each α, the subtree $T_\alpha \subset T_{max}$ that minimizes $C_\alpha(T)$. The more terminal nodes in the tree, the more complex it is because there is more flexibility in partitioning the space into smaller regions, resulting in more possibilities for fitting the training data. The tuning parameter α balances the trade-off between model fit and complexity. When $\alpha = 0$, the resulting tree is the full tree T_{max} because the complexity penalty is dropped and $C_\alpha(T)$ just measures the training error. Lower values of α result in larger subtrees, while higher values of α yield smaller subtrees. In fact, as α increases, there is a higher penalty to pay for having a tree with many terminal nodes, and so $C_\alpha(T)$ will tend to be minimized for a smaller subtree. As α approaches infinity, the tree T_{root} of size 1 (i.e., a single root node) will be selected. Typically, the best tuning parameter α is selected using K-fold cross-validation. Once the value of α is selected, we get back to the full training set and obtain the subtree corresponding to α. Here is a useful summary of the procedure as reported in "An Introduction to Statistical Learning with Applications in R" (54):

Algorithm 8.1 *Building a Regression Tree*

1. Use recursive binary splitting to grow a large tree on the training data, stopping only when each terminal node has fewer than some minimum number of observations.
2. Apply cost complexity pruning to the large tree in order to obtain a sequence of best subtrees, as a function of α.
3. Use K-fold cross-validation to choose α. That is, divide the training observations into K folds. For each $k = 1, \ldots, K$:
   (a) Repeat Steps 1 and 2 on all but the kth fold of the training data.
   (b) Evaluate the mean squared prediction error on the data in the left-out kth fold, as a function of α.
   Average the results for each value of α, and pick α to minimize the average error.
4. Return the subtree from Step 2 that corresponds to the chosen value of α.

Mathematically, the minimization of $C_\alpha(T)$ through the cost-complexity pruning approach works as follows.

For any internal node $c \in T_{max}$, let $C_\alpha(c) = C(c) + \alpha$. Also, for any branch T_c, let $C_\alpha(T_c) = C(T_c) + \alpha|\ddot{T}_c|$. When $\alpha = 0$, and for sufficiently small α values, the inequality $C_0(T_c) < C_0(c)$ holds. As the value of α gradually increases, the inequality sign will be reversed: $C_\alpha(T_c) > C_\alpha(c)$. The node c that achieves the equality at the smallest α is called the weakest link.

Solving the inequality, we get $\alpha < \frac{C(c)-C(T_c)}{|\ddot{T}|-1} = g_0(c)$. The weakest link c_0 in T_{max} is the node for which $g_0(c_0) = \min_c g_0(c)$. Define $\alpha_1 = g_0(c_0)$, which means that if α increases from zero, c_0 is the first internal node c such that pruning T_c improves the cost-complexity criterion. Then T_1 is define as $T_1 = T_{max} - T_{c_0}$ and the next weakest link c_1 in T_1 is identified. By iterating this process until the tree is reduced to its root node c_{root}, we obtain a decreasing sequence of trees $T_{max} \supset T_1 \supset \ldots \supset T_{root}$, along with an increasing sequence of α values $0=\alpha_0 < \alpha_1 < \alpha_2 < \cdots$. (55)

2.3.2.1 CART for classification

When the target variable y is categorical, the algorithm is named *classification tree* and its goal is to identify the class in which y is most likely to fall into. Growing a classification tree is analogous to growing a regression tree, with the exception that the criterion used to select the variable and cut-off value to determine binary splits can no longer be the RSS. As an alternative, popular metrics that can be used are the classification error rate, the Gini impurity index, and the entropy. The former is simply the fraction of the training observations in the region R_t that do not belong to the most commonly occurring class. In a node t, $\hat{p}_k$ is the proportion of training observations labeled with class k. Then, the classification error is given by $CE = 1 - \max_k(\hat{p}_k)$. However, classification error is not sensitive enough for tree-growing, thus the other two measures are preferred. The Gini impurity index at node t is calculated as: $G(t) = \sum_{k=1}^{K} \hat{p}_k(1 - \hat{p}_k)$. If all the $\hat{p}_k$ in the node are close to 0 or 1, $G(t)$ has a small value. In such cases, most of the observations will be from the same class, and for this reason the Gini index is referred to as a measure of node purity. The entropy at node t is defined as: $E(t) = -\sum_{k=1}^{K} \hat{p}_k \log(\hat{p}_k)$. Like the Gini index, $E(t)$ assumes a small value if the $\hat{p}_k$ are all near 0 or 1, that is if the node is pure.

After a split into two sub-nodes, t_L and t_R, the Gini impurity index becomes: $G(t_L, t_R) = p_L G(t_L) + p_R G(t_R)$, where p_L and p_R are the proportion of observations in the two sub-nodes. Similarly, the entropy becomes: $E(t_L, t_R) = p_L E(t_L) + p_R E(t_R)$. The splitting continues as long as it substantially decreases impurity, which means $\Delta = G(t) - G(t_L, t_R) > \varepsilon$, where the threshold ε is set by the user (analogous for the entropy). The variable and cutoff value for the split are chosen to minimize the impurity (i.e., the pair variable-cutoff that maximizes Δ). (53, 54)

2.3.3 Ensamble methods

Decision trees have some strengths: the decision rule can be easily represented graphically, they are intuitive and understandable even for non-experts, and they can handle both qualitative and quantitative variables. On the other hand, they are highly data-dependent, therefore even small changes in the dataset might result in the estimation of a significantly different tree, and prediction accuracy is quite low. To address these issues, ensamble methods have been proposed. The basic idea is to combine many simple models that perform relatively poorly but better than random guessing. These simple models are referred to as weak learners. (56)

Bagging (57) is an ensemble method that combines the predictions of multiple decision trees to improve model accuracy. It works by generating multiple bootstrap samples from the training data, fitting a separate decision tree to each sample, and then aggregating their predictions (by averaging for regression or majority voting for classification). This process reduces variance and helps prevent overfitting. However, one major drawback of bagging is that the individual trees can be highly correlated, especially when strong predictors dominate the data. This correlation reduces the effectiveness of the ensemble method. (58) To address this issue, random forest was introduced by Breiman in 2001 (59). It is an ensemble technique that builds multiple decision trees for regression or classification tasks, like bagging, but introduces additional randomness. Specifically, random forest not only uses bootstrap samples of the data but also selects a random subset of features at each split, thereby reducing the correlation between trees and improving generalization performance. The random forest algorithm based on Breiman's publication (59) was also implemented in the package *randomForest* of the software R. (60)

2.3.3.1 Random forest

Random forest is an enhanced form of bagging that creates ensembles of independent decision trees. As stated above, one major drawback of bagging is that the individual trees can be highly correlated. This happens because many trees tend to split on the same strong predictors early in the process, which leads them to make similar decisions. To reduce the correlation among the trees, the following strategy is employed. Each tree is trained on a distinct bootstrap sample derived from the full training set, as done in bagging. Then, for each tree at each split, a random subset of q predictors is selected

from all the input variables. The optimal predictor and its corresponding threshold for the split are therefore determined from this subset.

More formally, let D represents the training data $\{\boldsymbol{x}_i, y_i\}_{i=1}^{N}$, where $\boldsymbol{x} = (x_1, \ldots, x_p)$ denotes the p input variables (i.e., vector of features, that can be both numeric or categorical) and y_i denotes the response (i.e., numeric or categorical target variable). For $j = 1$ to J, a bootstrap sample D_j is drawn from D. Let OOB_j, where OOB stands for *out-of-bag*, be the set of data points not in D_j. Thus, D is the disjoint union of the two set, i.e. $D = D_j \cup OOB_j$.

A tree T_j is built using D_j by applying the following steps until a minimum predefined node size or a maximum number of terminal nodes is reached:

(i) At each node, a random subset of $q \leq p$ inputs are selected as splitting variables.

(ii) Among the q selected variables, identify the variable x_j and the corresponding split point s, that either minimizes the root mean squared error (RMSE, for regression tasks) or maximizes the reduction in impurity (for classification tasks).

(iii) The node is split into two child nodes: $\{x_j \leq s\}$ and $\{x_j > s\}$.

Finally, to make a prediction at a new point x_i, the following approach is used. For regression, the prediction is given by $\hat{f}(x_i) = \frac{1}{J}\sum_{j=1}^{J} \hat{h}_j(x_i)$, where $\hat{h}_j(x_i)$ is the prediction of the response variable at x_i based on T_j. In other words, this corresponds to the average of the values of the terminal nodes of the individual trees to which observation x_i is assigned. For classification, the prediction is determined as $\hat{f}(x_i) = \arg\max_{y} \sum_{j=1}^{J} I(\hat{h}_j(x_i) = y)$, where $I(\cdot)$ is the indicator function and $\hat{h}_j(x_i)$ is the predicted class for x_i based on T_j.

The OOB observations are used for estimating the prediction error of the random forest algorithm, namely the OOB error that is calculated by evaluating the performance of the model on these OOB instances. The OOB prediction for an observation is computed by averaging the predicted values from the weak learners (i.e., T_j) in which the observation was not included during training. For regression, the OOB prediction is expressed as $\hat{f}_{OOB}(x) = \frac{1}{J}\sum_{j \in OOB_j} \hat{h}_j(x)$, while for classification is given by $\hat{f}_{OOB}(x) = \arg\max_{y} \sum_{j \in OOB_j} I(\hat{h}_j(x) = y)$. For regression, the prediction error is typically estimated using the OOB mean squared error (MSE), defined as $MSE_{OOB} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{f}_{OOB}(x_i))^2$, where $\hat{f}_{OOB}(x_i)$ is the OOB prediction for observation i. For classification, the prediction error is estimated using the OOB error rate, which is obtained as $E_{OOB} = \frac{1}{N}\sum_{i=1}^{N}(y_i \neq \hat{f}_{OOB}(x_i))$. Since the OOB MSE (or OOB error rate) is computed using unseen data, it provides an effective measure for assessing the goodness of fit of the model. (61,62)

Tuning hyperparameters

In the case of random forests, tuning hyperparameters include the number q of randomly selected splitting variables at each node, the number J of decision trees in the forest, and the tree size defined as the smallest node size for splitting or the maximum number of terminal nodes. (63)

The number q of variables selected at random at each split is a critical hyperparameter. Choosing a low value of q reduces the correlation among the trees in the ensemble and mitigates the dominance of variables with a strong effect on the response, allowing variables with moderate effect to contribute more significantly to the model. However, a smaller q can also lead to less optimal individual trees, as splits may be determined based on noninformative variables from the limited candidate set. If many predictors are relevant, a smaller q is preferable as it allows less influential, yet still useful, variables to contribute to splits. These variables may improve predictions for small groups of observations where stronger features fail to predict accurately. Conversely, when only a few predictors are relevant, a larger q is preferable, ensuring that at least one strong predictor is likely included among the candidates for splitting. As a rule of thumb proposed by Breiman, $q = \frac{p}{3}$ is used for regression tasks, while $q = \sqrt{p}$ is applied for classification trees.

The number J of decision trees in the forest should be set sufficiently high to ensure robust performance. Breiman demonstrated that as J increases the prediction error of a random forest converges almost surely to a limit. This implies that increasing J does not lead to overfitting, making it safe to chose J as large as desired. The only real concern when selecting J is to avoid it being too small. In practice, the OOB error can be used to determine when J is large enough: by plotting the OOB error against the number of trees, it can be observed when the error stabilizes as J increases. As a general rule of thumb, Boehmke et al. suggested to define J as 10 times the number of input variables. (64)

Setting a low value of minimum number of observations in a terminal node results in deeper trees, as additional splits are performed before reaching the terminal nodes. The default value, which is generally considered to yield good results, is usually set to 1 for classification and 5 for regression tasks in many standard statistical software. (65,66) Lastly, restricting the maximum number of terminal nodes results in shallower trees. It is worth recalling that unpruned trees tend to produce more effective weak learners.

In conclusion, iterating over different values of tuning hyperparameters is necessary to determine the ones that provide the most accurate results.

Variable importance

To measure the importance of each predictor, the variable importance is a widely used metric based on the permutation of variables. The basic idea behind this approach is that if x_j is a strong predictor, permuting its values will decrease the prediction accuracy. (67–69) First, for each tree, the goodness of fit is computed using the OOB sample, i.e. the MSE for regression tasks, or the error rate for classification problems. Then, the values of the variable x_j are randomly permuted while keeping all the other predictors fixed. The MSE (or error rate) is recalculated considering the predictions obtained on the permuted OOB data. For each variable x_j the importance is computed in terms of the difference of MSE between the prediction obtained using x_j and its permuted version. This procedure is repeated for all bootstrap samples, and the overall variable importance is computed by averaging all the results. The percentage increase in MSE is also used, defined as MSE after permutation minus MSE before permutation divided by the latter.

More formally, the predictions $\hat{h}_j(x_i)$ are firstly calculated on the original set of OOB data $OOB_j = \{ i \mid (x_i, y_i) \notin D_j \}, i = 1, \dots, N$, where D_j represents the j^{th} bootstrap sample. Then, the values of variable x_j are randomly permuted for the data points $\{x_i : i \in OOB_j\}$ to obtain $P_j = \{x_i^* : i \in OOB_j\}$, and the predictions $\hat{h}_j(x_i^*)$ are computed. Finally, the importance of the variable x_j is given by $Imp_j = \frac{1}{J_i}\sum_j \left(y_i - \hat{h}_j(x_i^*)\right)^2 - \frac{1}{J_i}\sum_j \left(y_i - \hat{h}_j(x_i)\right)^2$ for regression, and by $Imp_j = \frac{1}{J_i}\sum_j \left(y_i \neq \hat{h}_j(x_i^*)\right) - \frac{1}{J_i}\sum_j \left(y_i \neq \hat{h}_j(x_i)\right)$ for classification.

2.4 Statistical analysis

2.4.1 Descriptive analysis

As a first step of the research project, we conducted a descriptive analysis to characterize the target population for each year for which data were available (i.e., from 2011 to 2023).

First, we reported the number and percentage of individuals who used NHS services - specifically inpatient and outpatient services - and we reported the corresponding mean costs with interquartile ranges (IQR).

Then, we identified seven groups of individuals based on their direct annual cost for inpatient and outpatient services: *i)* Group 0 included individuals who did not use any services during the year and were therefore classified as “zero-cost” subjects; *ii)* Group 1 included individuals whose annual cost fell within the 25th percentile of the cost distribution; *iii)* Group 2 included those with an annual cost

between the 25th and 50th percentiles; *iv)* Group 3 included those with an annual cost between the 50th and 75th percentiles; *v)* Group 4 included those with an annual cost between the 75th and 95th percentiles; *vi)* Group 5 included those with an annual cost between the 95th and 99th percentiles; and *vii)* Group 6 included individuals with an annual cost above the 99th percentile. Each group was then analyzed in terms of demographic characteristics (sex and age), inpatients and outpatient healthcare services received, and costs.

2.4.2 Cost prediction analysis

A cost prediction analysis was then conducted to address the main objective of the research project, namely the prediction of healthcare costs based on past NHS resources utilization, for the whole population and for high-impact segments. Briefly, we used individual data of 5-year resource utilization of the NHS to predict individual's healthcare costs in the following year. Subsequently, we aggregated individual cost predictions to obtain predictions for the whole population and for specific segments of interest. We used a supervised machine learning approach, specifically the random forest algorithm, to address the prediction problem.

Given the availability of data from 2011 to 2023, we constructed eight datasets for training and validation of the predictive algorithms. Each dataset included five years of demographic, NHS utilization, and historical costs data (predictors) for individuals assisted by the HPA in that time period, along with their corresponding healthcare costs (outcome) in the subsequent year:

1. Dataset 2011-16: included predictors from 2011 to 2015 and healthcare costs in 2016.
2. Dataset 2012-17: included predictors from 2012 to 2016 and healthcare costs in 2017.
3. Dataset 2013-18: included predictors from 2013 to 2017 and healthcare costs in 2018.
4. Dataset 2014-19: included predictors from 2014 to 2018 and healthcare costs in 2019.
5. Dataset 2015-20: included predictors from 2015 to 2019 and healthcare costs in 2020.
6. Dataset 2016-21: included predictors from 2016 to 2020 and healthcare costs in 2021.
7. Dataset 2017-22: included predictors from 2017 to 2021 and healthcare costs in 2022.
8. Dataset 2018-23: included predictors from 2018 to 2022 and healthcare costs in 2023.

Predictors

Each dataset included a total of 380 predictors, which were categorized into six domains:

1) Demographic data: sex and age group of the individual in the last year for which information was available (e.g., for the 2011-16 dataset, age was considered as of 2015). Age was

categorized as follows: 18-29 years, 30-39 years, 40-49 years, 50-59 years, 60-69 years, 70-79 years, and ≥80 years.

2) Inpatient services data: number and type of hospital admissions (see Section 2.2 for details) in the last year, as well as service utilization in the previous four years. For example, the number of scheduled hospitalizations with a primary diagnosis of neoplasms during the last year, and the frequency of such hospitalizations over the previous four-year period. Individuals were classified as:
    - *non-users* of the service (e.g., scheduled hospitalizations for neoplasms) if they had no more than one admission over the four-year period and zero admissions in the fifth year;
    - *intermittent users* if they had one hospitalization in at least two years of the four-year period, or if they had at least one hospitalization in both the fourth and fifth years;
    - *frequent users* if they had more than one hospitalization in at least two years of the four-year period.
3) Outpatient services data: number and type of outpatient services received (see Section 2.2 for details) in the last year, as well as service utilization in the previous four years. For example, the number of cardiology specialist visits during the last year, and the frequency of such visits over the previous four-year period. Individuals were classified as:
    - *non-users* of the service (e.g., cardiology visits) if they had no more than one visit over the four-year period and zero visits in the fifth year;
    - *intermittent users* if they had one visit in at least two years of the four-year period, or if they had at least one visit in both the fourth and fifth years;
    - *frequent users* if they had more than one visit in at least two years of the four-year period.
4) Laboratory test bundles data: number and type of laboratory test bundles prescriptions (see Section 2.2 for details) in the last year, as well as service utilization in the previous four years. For example, the number of liver function tests prescriptions during the last year, and the frequency of such tests over the previous four-year period. Individuals were classified as:
    - *non-users* of the service (e.g., liver function test) if they had no more than one test over the four-year period and zero tests in the fifth year;
    - *intermittent users* if they had one test in at least two years of the four-year period, or if they had at least one test in both the fourth and fifth years;
    - *frequent users* if they had more than one test in at least two years of the four-year period.

5) Pharmaceutical dispensations data: utilization of drug classes during the last year (see Section 2.2 for details), as well as utilization in the previous four years. For example, use of oral antidiabetics during the last year, and frequency of use over the previous four-year period. Based on utilization patterns over the four-year period, individuals were classified as follows:
    - For drug classes A02, A07, J01, J02, M04, N01, N02C, and R06:
        - *non-users* if they did not use the drug class in any of the four years, or they were classified as occasional or intermittent users only in the fourth year and non-users in the fifth year;
        - *occasional/intermittent users* if they were classified as such for at least two of the four years, or they were occasional/intermittent users in both the fourth and fifth years.
    - For all other drug classes:
        - *non-users* if they did not use the drug class in any of the four years, or they were classified as occasional or frequent users only in the fourth year and non-users in the fifth year;
        - *occasional users* if they were occasional users in the fourth year and occasional or frequent users in the fifth year; or if they were frequent users in one year and occasional users in a subsequent year; or if they were frequent users in two years and occasional users in the other two years (with occasional use in the fourth year); or if they were classified as occasional or frequent users in three or four years, with a greater number of years being occasional users;
        - *frequent users* if they were frequent users in the fourth year and occasional or frequent users in the fifth year; or if they were occasional users in one year and frequent users in a subsequent year; or if they were frequent users in two years and occasional users in the other two years (with frequent use in the fourth year); or if they were occasional or frequent users in three or four years, with a greater number of years being frequent users.

6) Historical costs data: mean total healthcare cost calculated over the five-year period, and a binary variable indicating whether the individual had zero costs in the last year (i.e., no NHS service utilization was recorded during that year).

The complete list of predictors is provided in Supplementary Materials (**Table S1**).

Outcomes

For each year and for each subject, we calculated four different outcomes:

1) Total cost (TC): sum of all inpatient and outpatient services and drug dispensations direct costs.
2) Total scheduled cost (TSC): sum of scheduled inpatient visits, all outpatient visits, and drug dispensations direct costs.
3) Services cost (SC): sum of all inpatient and outpatient services direct costs.
4) Scheduled services cost (SSC): sum of scheduled inpatient and all outpatient services direct costs.

Training and validation of random forest regression algorithms

The aforementioned datasets were used for training and validating predictive algorithms across eight scenarios defined by combinations of predictors and outcomes:

1. Demographic data, inpatient and outpatient service data, laboratory test bundles data, and pharmaceutical dispensation data collected over a five-year period - excluding historical cost information from the set of predictors - were used to predict TCs for the subsequent year.
2. Demographic data, inpatient and outpatient service data, laboratory test bundles data, pharmaceutical dispensation data, and historical costs data collected over a five-year period were used to predict TCs for the subsequent year.
3. Demographic data, inpatient and outpatient service data, laboratory test bundles data, and pharmaceutical dispensation data collected over a five-year period - excluding historical cost information from the set of predictors - were used to predict TSCs for the subsequent year.
4. Demographic data, inpatient and outpatient service data, laboratory test bundles data, pharmaceutical dispensation data, and historical costs data collected over a five-year period were used to predict TSCs for the subsequent year.
5. Demographic data, inpatient and outpatient service data, laboratory test bundles data, and pharmaceutical dispensation data collected over a five-year period - excluding historical cost information from the set of predictors - were used to predict SCs for the subsequent year.
6. Demographic data, inpatient and outpatient service data, laboratory test bundles data, pharmaceutical dispensation data, and historical costs data collected over a five-year period were used to predict SCs for the subsequent year.

7. Demographic data, inpatient and outpatient service data, laboratory test bundles data, and pharmaceutical dispensation data collected over a five-year period - excluding historical cost information from the set of predictors - were used to predict SSCs for the subsequent year.
8. Demographic data, inpatient and outpatient service data, laboratory test bundles data, pharmaceutical dispensation data, and historical costs data collected over a five-year period were used to predict SSCs for the subsequent year.

For each scenario and for each training set, a predictive algorithm was developed using a supervised machine learning approach, specifically random forest regression. Hyperparameter tuning was performed through grid search. The optimal hyperparameters, then applied across all random forest regressions, were as follows: number of trees = 500; number of variables randomly sampled as candidates at each split = 125; minimum size of terminal nodes = 5; and maximum number of terminal nodes per tree = 5,000.

First, we trained the algorithm for each of the eight scenarios on the 70% of individuals' data from 2011 to 2015 with their costs in 2016 as outcome (dataset 2011-16). As validation sets, we used the remaining 30% of the 2011-16 dataset and the subsequent years' datasets, namely the 2012-17, 2013-18, 2014-19, 2015-20, 2016-21, 2017-22, and 2018-23 datasets. The algorithm was validated on 30% of the 2011-2016 dataset to assess the consistency of predictions across different years and to detect potential overfitting; however, the results are not presented here.

Then, we updated the algorithm for each of the eight scenarios using all available datasets. Specifically, the algorithm was trained on data of individuals in the 2012-17 dataset and validated on the 2013-18 dataset; trained on the 2013-18 dataset and validated on the 2014-19 dataset; trained on the 2014-19 dataset and validated on the 2015-20 dataset; trained on the 2015-20 dataset and validated on the 2016-21 dataset; trained on the 2016-21 dataset and validated on the 2017-22 dataset; and finally, trained on the 2017-22 dataset and validated on the 2018-23 dataset.

Individual-level costs' predictions

We considered variable importance, measured as the percent increase in mean squared error (MSE) when a given variable is permuted, as a measure of each predictor's impact on the outcome, and the proportion of explained variance as a measure of how well OOB predictions explain the target variance of the training set. For each test set, we measured the performance of the prediction algorithm using mean absolute error (MAE) and root mean squared error (RMSE). RMSE is defined as the average sum of the squared differences between the actual value and the predicted value, providing a measure of how far predictions deviate from actual outcomes at the individual level. Formally, RMSE

is defined as: $\sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}$. MAE is defined as the average sum of the absolute differences between the actual value and the predicted value, serving as a straightforward measure of model accuracy that is less sensitive to outliers. Formally, MAE is defined as: $\frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|$.

Costs' predictions for the whole population

Then, for each scenario, actual and predicted costs (i.e., TC, TSC, SC, and SSC) for the whole population were calculated as the sum and the mean of all individual's actual and predicted costs, respectively. The ratio of the difference between predicted and actual cost to actual cost was used as measure of the prediction error (PE). PE=0% indicates a perfect prediction, if PE >0% or <0% the prediction overestimates or underestimates the actual cost, respectively. We also derived a variability interval for the mean predicted cost, defined by the 2.5th and 97.5th percentiles of the distribution of mean costs predicted by each weak learner for each subject in the population.

Costs' predictions for high-impact segments

Finally, high-impact population segments were defined based on the classifications provided by the BDA. In the present study, the chronic conditions considered for defining high-impact segments were the following:

- Chronic kidney disease (CKD) - Dialysis (BDA code: K03A)
- Type 2 diabetes mellitus/Type 2 complicated diabetes mellitus (BDA codes: K06B1 and K06B2)
- Heart failure (BDA code: K07C)
- Parkinson's disease/Parkinsonian syndromes (BDA code: K10B)
- Active neoplasia (BDA code: K05A)

BDA classifications were available for subjects assisted by the HPA of Bergamo from 2017. It is important to note that in 2020 significant updates were made to the chronic disease classification algorithm. As a result, relevant differences in the categorization of individuals may exist before and after that year. To identify individuals included in the cost prediction analysis who belonged to the segments of interest, we linked the BDA classifications to our analytical datasets. The record linkage was made possible using pseudo-anonymized identifiers.

As previously mentioned, each individual may be affected by more than one chronic condition. Therefore, we adopted several definitions, ranging from less to more restrictive, to identify the population segments of interest. Specifically:
i) individuals affected by the chronic condition of interest, though it may not be the only one (in the results, referred to as "all");
ii) individuals for whom the chronic condition of interest is classified as the primary condition, though it may not be the only one (in the results, referred to as "primary condition");
iii) individuals affected only by the chronic condition of interest (in the results, referred to as "only");
iv) individuals for whom the chronic condition of interest is classified as the primary condition, and with no more than two additional chronic conditions besides the one of interest (in the results, referred to as "primary condition; ≤2 comorbidities");
iv) individuals for whom the chronic condition of interest is classified as the primary condition, and with three or more additional chronic conditions besides the one of interest (in the results, referred to as "primary condition; >2 comorbidities").

For each scenario, actual and predicted costs (i.e., TC, TSC, SC, and SSC) for each selected high-impact segment of the population were calculated as the sum and the mean of actual and predicted costs, respectively, of individuals classified within each segment. The PE was then calculated for each segment. Also, we derived a variability interval for the mean predicted cost based on the 2.5 and 97.5 quantiles of the distribution of the mean costs predicted by each weak learner for subjects included in the segment.

Statistical analyses were carried out using SAS software v. 9.4 (SAS Institute, Cary, NC) and R v. 4.3.0 (The R Project For Statistical Computing, Vienna, Austria).

3. RESULTS

3.1 Description of the target population

From 2011 to 2023, a total of 1,388,639 individuals were covered by the HPA of Bergamo for at least one year. The mean annual healthcare services cost for the whole population (i.e., hospitalization and outpatient costs) in the 2011-2023 period was €755,226,988 (range: 678,517,959 - 818,184,933).

In 2011, 762,370 out of 1,113,455 individuals (68.5%) underwent at least one hospital admission and/or outpatient service (i.e., outpatient visit or laboratory test). The mean direct services cost to the NHS per subjects, considering the entire population, was €678 (IQR: 0-254). Within the subset of individuals who underwent at least one hospital admission and/or outpatient service, the mean cost was €990 (IQR: 57-470), while a mean cost of €316 (IQR: 51-295) was observed for outpatient services among the 752,801 (67.6%) individuals who received them and €4,407 (IQR: 1,180-4,293) for hospitalizations among the 117,296 (10.5%) individuals who had at least one. **Table 3.1** reports the main demographic and cost characteristics of the 7 cost groups identified based on percentiles of the cost distribution (**Figure 3.1A**). It should be noted that the proportion of individuals aged 60 or older increased significantly with higher cost groups, as also depicted in **Figure 3.2A**. Furthermore, **Figure 3.3A** shows how resource consumption was inversely proportional to the size of the cost group. Group 0 included 351,085 individuals (31.5%) who did not use NHS services during the year and can therefore be defined as "zero-cost" subjects. As the cost group increases, reflecting higher annual healthcare expenditure, the number of individuals decreases, culminating in Group 6, which represented only 0.7% of the population but accounted for 28.3% of the total annual healthcare costs for the entire population.

In 2015, 785,791 out of 1,082,308 individuals (72.6%) underwent at least one hospital admission and/or outpatient service. The mean direct services cost to the NHS per subjects, considering the entire population, was €665 (IQR: 0-248). For subjects who underwent at least one hospital admission and/or outpatient service, the mean cost was €952 (IQR: 57-431). A total of 779,363 (72.0%) individuals received at least one outpatient service, with a mean cost of €350 (IQR: 53-305), while 99,649 (9.2%) subjects had at least one hospitalization with a mean cost of €4,772 (IQR: 1,485-4,997). Main demographic and cost features of the 7 cost groups identified in 2015 are reported in **Table 3.2**, **Figure 3.1B**, **Figure 3.2B** and **Figure 3.3B**. Also for 2015, it can be observed that the highest-cost group (i.e., individuals with an annual cost ≥€14,714) alone accounted for 28.6% of the total healthcare costs of the entire population.

In 2020, 736,487 out of 1,099,868 individuals (67.0%) received at least one hospital admission and/or outpatient service. The mean direct services cost to the NHS per subjects, considering the entire population, was €602 (IQR: 0-192). Within the subset of individuals who underwent at least one

hospital admission and/or outpatient service, the mean cost was €921 (IQR: 61-350), while a mean cost of €338 (IQR: 57-273) was observed for outpatient services among the 730,833 (66.4%) individuals who received them and €5,690 (IQR: 1,996-6,066) for hospitalizations among the 75,794 (6.9%) individuals who had at least one. Main demographic and cost features of the 7 cost groups identified in 2020 are reported in **Table 3.3**, **Figure 3.1C**, **Figure 3.2C** and **Figure 3.3C**. As expected, given the year affected by the COVID-19 pandemic, a higher proportion of zero-cost individuals was observed compared to previous years (33.0% compared to 31.5% in 2011, 32.3% in 2012, 32.2% in 2013, 29.0% in 2014, 27.4% in 2015, 27.7% in 2016, 27.7% in 2017, 27.4% in 2018, and 27.1% in 2019). Additionally, Group 6, consisting of high-cost individuals, absorbed 32.4% of the total annual healthcare costs.

In 2023, 802,826 out of 1,090,875 subjects (73.6%) received at least one hospital admission and/or outpatient service. The mean direct services cost to the NHS per subjects, considering the entire population, was €727 (IQR: 0-285). For subjects who underwent at least one hospital admission and/or outpatient service, the mean cost was €1,019 (IQR: 66-462). A total of 797,907 (73.1%) individuals received at least one outpatient service, with a mean cost of €393 (IQR: 65-349), while 91,079 (8.3%) subjects had at least one hospitalization with a mean cost of €5,536 (IQR: 1,770-6,213). Main demographic and cost features of the 7 cost groups identified in 2023 are reported in **Table 3.4**, **Figure 3.1D**, **Figure 3.2D** and **Figure 3.3D**. After the COVID-19 pandemic years, the proportion of zero-cost individuals decreased to 26.4% in 2023. Group 6 (i.e., individuals with annual healthcare costs ≥€15,594) alone accounted for 29.1% of the total healthcare expenditures of the entire population.

Figure 3.4 shows the transitions between groups of healthcare costs for inpatient and outpatient services (i.e., annual cost equal to €0; <€100; €100-499; €500-999; €1,000-5,000; €5,000-9,999; >€10,000) from 2011 to 2023. This analysis included only individuals who were continuously assisted by the HPA of Bergamo throughout the entire period; as a result, slight differences may be observed in the proportions of zero-cost individuals compared to the descriptive statistics reported above for single years.

The distribution of individuals across the various cost classes, as well as the transitions between these classes over time, were relatively stable between 2011 and 2019. However, starting in 2015, there was a noticeable decrease in the proportion of zero-cost individuals, along with an increase in the proportion of those in the €100-499 cost class. Similarly, the higher-cost classes (€5,000-9,999 and >€10,000) also showed an upward trend beginning in 2015, growing from proportions below 1% to as much as 1% of the annual population. The cost distribution in 2020 reflected the significant disruption to the healthcare system caused by the COVID-19 pandemic. A marked increase in the

proportion of individuals with zero costs was observed, accompanied by a corresponding decrease in the proportion of individuals across other cost classes, except for the highest-cost groups. In 2021, the opposite pattern emerged: there was a substantial decline in the number of individuals who did not utilize inpatient or outpatient services during the year, with the proportion dropping from 31% in 2020 to 23% in 2021. Notably, in the following years, the distribution of individuals across cost classes remained stable compared to 2021, without returning to pre-pandemic levels. Furthermore, in both 2022 and 2023, there was a continued increase in the proportion of individuals in the higher-cost categories (i.e., costs exceeding €5,000), which grow to include up to 2% of the population.

Table 3.1. Main demographic and cost features of the seven cost groups identified based on percentiles of the cost distribution in 2011.

	Group 0 Annual cost: €0	**Group 1** Annual cost: >€0 - 57	**Group 2** Annual cost: €57 - 148	**Group 3** Annual cost: €148 - 470	**Group 4** Annual cost: €470 - 3,965	**Group 5** Annual cost: €3,965 - 15,185	**Group 6** Annual cost: >€15,185
All subjects	**351,085**	**190,810**	**190,869**	**190,100**	**152,473**	**30,494**	**7,624**
Demographic charateristics							
Males, N (%)	201,681 (57.5)	94,056 (49.3)	88,460 (46.3)	82,370 (43.3)	65,143 (42.7)	15,205 (49.9)	4,395 (57.6)
Age class, N (%)							
< 18 years	87,831 (25.0)	36,060 (18.9)	39,845 (20.9)	18,186 (9.6)	19,934 (13.1)	1,137 (3.7)	261 (3.4)
18-29 years	60,158 (17.1)	26,730 (14.0)	21,506 (11.3)	15,949 (8.4)	12,297 (8.1)	1,018 (3.3)	144 (1.9)
30-39 years	68,005 (19.4)	29,312 (15.4)	23,220 (12.2)	21,583 (11.4)	20,220 (13.3)	2,090 (6.9)	247 (3.2)
40-49 years	67,552 (19.2)	31,704 (16.6)	31,973 (16.8)	32,476 (17.1)	20,324 (13.3)	2,710 (8.9)	537 (7.0)
50-59 years	29,807 (8.5)	30,647 (16.1)	27,266 (14.3)	32,342 (17.0)	20,276 (13.3)	3,657 (12.0)	972 (12.7)
60-69 years	17,096 (4.9)	19,901 (10.4)	22,505 (11.8)	31,865 (16.8)	23,132 (15.2)	5,832 (19.1)	1,582 (20.8)
70-79 years	11,797 (3.4)	9,363 (4.9)	14,785 (7.7)	24,912 (13.1)	23,478 (15.4)	7,883 (25.9)	2,367 (31.0)
≥80 years	8,869 (2.5)	7,093 (3.7)	9,768 (5.1)	12,787 (6.7)	12,812 (8.4)	6,167 (20.2)	1,514 (19.9)
Age, median (IQR)	41 (23-55)	43 (30-59)	48 (31-64)	56 (40-70)	59 (40-72)	68 (53-77)	67 (55-76)
Subjects with at least one hospitalization, N (%)	-	1 (0.0)	21 (0.0)	3,765 (2.0)	76,302 (50.0)	29,945 (98.2)	7,262 (95.3)
Subjects with at least one outpatient service. N (%)	-	190,809 (100)	190,854 (100)	189,299 (99.6)	145,387 (95.4)	29,174 (95.7)	7,278 (95.5)
Deaths within the year, N (%)	1,242 (0.4)	643 (0.3)	647 (0.3)	660 (0.3)	1,764 (1.2)	3,061 (10.0)	1,131 (14.8)
Mean costs (€)							
Mean cost of hospitalizations per subject (IQR)	-	47 (47-47)	131 (125-134)	238 (164-316)	1731 (952-2255)	6,886 (4246-8965)	24,472 (15912-26317)
Mean cost of outpatient services per subject (IQR)	-	31 (21-44)	95 (70-116)	260 (187-322)	685 (228-921)	982 (204-1126)	4,961 (276-2439)
Mean cost of all services per subject (IQR)	-	31 (21-44)	95 (70-116)	264 (189-327)	1,520 (726-2274)	7,702 (4951-9849)	28,047 (17944-30325)
Total costs (€)							
Total cost of hospitalizations	-	47	2,754	894,661	132,079,908	206,211,561	177,718,280
Total cost of outpatient services	-	5,889,068	18,061,435	49,241,595	99,615,271	28,650,393	36,109,435
Total cost of all services	-	5,889,115	18,064,189	50,136,255	231,695,179	234,861,953	213,827,715

Table 3.2. Main demographic and cost features of the seven cost groups identified based on percentiles of the cost distribution in 2015.

	Group 0 Annual cost: €0	**Group 1** Annual cost: >€0 - 57	**Group 2** Annual cost: €57 - 141	**Group 3** Annual cost: €141 - 431	**Group 4** Annual cost: €431 - 3,973	**Group 5** Annual cost: €3,973 - 14,714	**Group 6** Annual cost: >€14,714
All subjects	**296,517**	**196,533**	**196,359**	**196,452**	**157,158**	**31,432**	**7,857**
Demographic charateristics							
Males, N (%)	167,643 (56.5)	98,823 (50.3)	91,782 (46.7)	84,811 (43.2)	68,070 (43.3)	15,541 (49.4)	4,374 (55.7)
Age class, N (%)							
< 18 years	62,188 (21.0)	41,950 (21.3)	45,052 (22.9)	22,105 (11.3)	13,729 (8.7)	1,176 (3.7)	188 (2.4)
18-29 years	50,059 (16.9)	26,087 (13.3)	22,801 (11.6)	15,141 (7.7)	11,328 (7.2)	937 (3.0)	129 (1.6)
30-39 years	52,588 (17.7)	25,881 (13.2)	19,944 (10.2)	18,190 (9.3)	17,520 (11.1)	1,928 (6.1)	204 (2.6)
40-49 years	61,995 (20.9)	31,904 (16.2)	30,115 (15.3)	31,237 (15.9)	20,529 (13.1)	2,713 (8.6)	489 (6.2)
50-59 years	33,029 (11.1)	33,711 (17.2)	29,771 (15.2)	35,667 (18.2)	23,910 (15.2)	3,809 (12.1)	992 (12.6)
60-69 years	17,998 (6.1)	20,524 (10.4)	23,275 (11.9)	33,552 (17.1)	26,486 (16.9)	5,901 (18.8)	1,665 (21.2)
70-79 years	10,712 (3.6)	8,879 (4.5)	14,635 (7.5)	25,245 (12.9)	27,362 (17.4)	7,811 (24.9)	2,351 (29.9)
≥80 years	7,948 (2.7)	7,597 (3.9)	10,766 (5.5)	15,315 (7.8)	16,294 (10.4)	7,157 (22.8)	1,839 (23.4)
Age, median (IQR)	37 (20-49)	45 (30-59)	49 (30-65)	57 (41-71)	61 (43-74)	69 (53-78)	68 (55-77)
Subjects with at least one hospitalization, N (%)	-	0	25 (0.0)	2139 (1.1)	60,118 (38.3)	29,973 (95.4)	7,394 (94.1)
Subjects with at least one outpatient service, N (%)	-	196,533 (100)	196,338 (100)	196,044 (99.8)	152,695 (97.2)	30,196 (96.1)	7,557 (96.2)
Deaths within the year, N (%)	1,281 (0.4)	737 (0.4)	752 (0.4)	714 (0.4)	2,004 (1.3)	3,084 (9.8)	1,069 (13.6)
Mean costs (€)							
Mean cost of hospitalizations per subject (IQR)	-	-	131 (129-134)	225 (150-300)	1,853 (1,145-2,357)	6,641 (4,206-8,696)	22,262 (14,728-25,185)
Mean cost of outpatient services per subject (IQR)	-	31 (22-44)	92 (68-111)	246 (178-305)	726 (406-940)	1,342 (218-1,379)	6,486 (312-6,560)
Mean cost of all services per subject (IQR)	-	31 (22-44)	92 (68-111)	248 (179-307)	1,414 (623-2,133)	7,622 (4,969-9,709)	27,189 (17,189-29,718)
Total costs (€)							
Total cost of hospitalizations	-	-	3,266	480,494	111,370,303	199,047,773	164,607,932
Total cost of outpatient services	-	6,123,056	17,974,828	48,243,691	110,799,981	40,536,877	49,017,046
Total cost of all services	-	6,123,056	17,978,094	48,724,185	222,170,284	239,584,650	213,624,978

Table 3.3. Main demographic and cost features of the seven cost groups identified based on percentiles of the cost distribution in 2020.

	Group 0 Annual cost: €0	**Group 1** Annual cost: >€0 - 61	**Group 2** Annual cost: €61 - 127	**Group 3** Annual cost: €127 - 350	**Group 4** Annual cost: €350 - 3,909	**Group 5** Annual cost: €3,909 - 14,741	**Group 6** Annual cost: >€14,741
All subjects	**363,381**	**184,197**	**184,043**	**184,125**	**147,297**	**29,461**	**7,364**
Demographic charateristics							
Males, N (%)	201,822 (55.5)	90,074 (48.9)	87,799 (47.7)	79,865 (43.4)	62,812 (42.6)	15,485 (52.6)	4,343 (59.0)
Age class, N (%)							
< 18 years	78,205 (21.5)	30,298 (16.4)	43,636 (23.7)	20,591 (11.2)	7,129 (4.8)	695 (2.4)	140 (1.9)
18-29 years	58,092 (16.0)	23,443 (12.7)	23,684 (12.9)	16,220 (8.8)	10,187 (6.9)	726 (2.5)	129 (1.8)
30-39 years	51,590 (14.2)	21,665 (11.8)	17,347 (9.4)	15,524 (8.4)	14,691 (10.0)	1,357 (4.6)	172 (2.3)
40-49 years	61,352 (16.9)	30,746 (16.7)	25,147 (13.7)	26,728 (14.5)	17,564 (11.9)	2,240 (7.6)	418 (5.7)
50-59 years	53,445 (14.7)	32,314 (17.5)	27,035 (14.7)	34,521 (18.7)	25,870 (17.6)	3,923 (13.3)	1,020 (13.9)
60-69 years	30,562 (8.4)	21,959 (11.9)	20,637 (11.2)	29,132 (15.8)	25,584 (17.4)	5,626 (19.1)	1,668 (22.7)
70-79 years	16,657 (4.6)	13,014 (7.1)	14,899 (8.1)	24,597 (13.4)	27,983 (19.0)	7,703 (26.1)	2,203 (29.9)
≥80 years	13,478 (3.7)	10,758 (5.8)	11,658 (6.3)	16,812 (9.1)	18,289 (12.4)	7,191 (24.4)	1,614 (21.9)
Age, median (IQR)	38 (20-54)	51 (34-66)	52 (31-68)	58 (42-72)	62 (45-74)	69 (55-78)	69 (57-77)
Subjects with at least one hospitalization, N (%)	-	0 (0.0)	1 (0.0)	758 (0.4)	40,559 (27.5)	27,674 (93.9)	6,802 (92.4)
Subjects with at least one outpatient service, N (%)	-	184,197 (100)	184,042 (100)	183,907 (99.9)	143,973 (97.7)	27,721 (94.1)	6,993 (95.0)
Deaths within the year, N (%)	2,703 (0.7)	1,816 (1.0)	1,201 (0.7)	1,157 (0.6)	3,037 (2.1)	4,271 (14.5)	1,069 (14.5)
Mean costs (€)							
Mean cost of hospitalizations per subject (IQR)	-	-	125 (125-125)	204 (149-229)	1,977 (1,252-2,493)	6,499 (4,085-8,349)	25,150 (14,781-26,985)
Mean cost of outpatient services per subject (IQR)	-	32 (20-45)	86 (66-102)	209 (156-255)	693 (383-843)	1,398 (198-1,307)	6,949 (287-7,844)
Mean cost of all services per subject (IQR)	-	32 (20-45)	86 (66-102)	210 (156-255)	1,222 (481-1,743)	7,420 (4,762-9,466)	29,830 (17,475-32,555)
Total costs (€)							
Total cost of hospitalizations	-	-	125	154,913	80,199,765	179,847,640	171,072,079
Total cost of outpatient services	-	5,831,485	15,764,547	38,517,929	99,773,692	38,760,859	48,595,975
Total cost of all services	-	5,831,485	15,764,672	38,672,843	179,973,457	218,608,499	219,668,054

Table 3.4. Main demographic and cost features of the seven cost groups identified based on percentiles of the cost distribution in 2023.

	Group 0 Annual cost: €0	**Group 1** Annual cost: >€0 - 66	**Group 2** Annual cost: €66 - 162	**Group 3** Annual cost: €162 - 462	**Group 4** Annual cost: €462 - 4,175	**Group 5** Annual cost: €4,175 - 15,594	**Group 6** Annual cost: >€15,594
All subjects	**288,049**	**215,630**	**185,783**	**200,707**	**160,565**	**32,112**	**8,029**
Demographic charateristics							
Males, N (%)	167,315 (58.1)	113,510 (52.6)	81,617 (43.9)	84,973 (42.3)	69,497 (43.3)	16,101 (50.1)	4,528 (56.4)
Age class, N (%)							
< 18 years	64,040 (22.2)	49,760 (23.1)	27,531 (14.8)	15,541 (7.7)	8020 (5.0)	753 (2.3)	159 (2.0)
18-29 years	54,143 (18.8)	32,050 (14.9)	21,308 (11.5)	16,354 (8.1)	10,681 (6.7)	856 (2.7)	106 (1.3)
30-39 years	44,090 (15.3)	23,976 (11.1)	17,793 (9.6)	16,521 (8.2)	14,706 (9.2)	1,608 (5.0)	155 (1.9)
40-49 years	45,356 (15.8)	30,930 (14.3)	27,498 (14.8)	27,750 (13.8)	17,236 (10.7)	2,209 (6.9)	422 (5.3)
50-59 years	37,549 (13.0)	36,972 (17.1)	33,305 (17.9)	40,458 (20.2)	28,417 (17.7)	4,266 (13.3)	986 (12.3)
60-69 years	22,040 (7.7)	21,575 (10.0)	26,510 (14.3)	36,923 (18.4)	30,746 (19.1)	6,267 (19.5)	1,737 (21.6)
70-79 years	11,655 (4.1)	11,170 (5.2)	17,297 (9.3)	28,539 (14.2)	30,872 (19.2)	8,530 (26.6)	2,539 (31.6)
≥80 years	9,176 (3.2)	9,197 (4.3)	14,541 (7.8)	18,621 (9.3)	19,887 (12.4)	7,623 (23.7)	1,925 (24.0)
Age, median (IQR)	35 (19-52)	45 (25-60)	53 (34-67)	58 (44-72)	63 (48-74)	69 (55-78)	70 (59-78)
Subjects with at least one hospitalization, N (%)	-	0	31 (0.0)	1,284 (0.6)	52,228 (32.5)	30,047 (93.6)	7,489 (93.3)
Subjects with at least one outpatient service, N (%)	-	215,630 (100)	185,764 (100)	200,469 (99.9)	157,039 (97.8)	31,201 (97.2)	7,804 (97.2)
Deaths within the year, N (%)	1,390 (0.8)	1,030 (0.5)	963 (0.5)	781 (0.4)	2,241 (1.4)	2,766 (8.6)	883 (11.0)
Mean costs (€)							
Mean cost of hospitalizations per subject (IQR)	-	-	141 (136-149)	237 (162-302)	1,987 (1,270-2,493)	7,094 (4,476-9,381)	24,961 (15,431-28,467)
Mean cost of outpatient services per subject (IQR)	-	38 (23-54)	109 (85-131)	274 (202-337)	819 (479-1,030)	1,625 (252-1,679)	6,591 (350-6,803)
Mean cost of all services per subject (IQR)	-	38 (23-54)	109 (85-131)	275 (203-338)	1,448 (648-2,114)	8,217 (5,358-10,519)	29,688 (18,382-32,530)
Total costs (€)							
Total cost of hospitalizations	-	-	4,367	304,772	103,799,490	213,162,184	186,929,924
Total cost of outpatient services	-	8,175,187	20,158,759	54,885,845	128,646,072	50,712,022	51,432,384
Total cost of all services	-	8,175,187	20,163,126	55,190,617	232,445,562	263,874,206	238,362,308

Figure 3.1. Cost distribution of patients with at least one hospitalization or outpatient service in 2011 (panel A), in 2015 (panel B), in 2020 (panel C), and in 2023 (panel D).

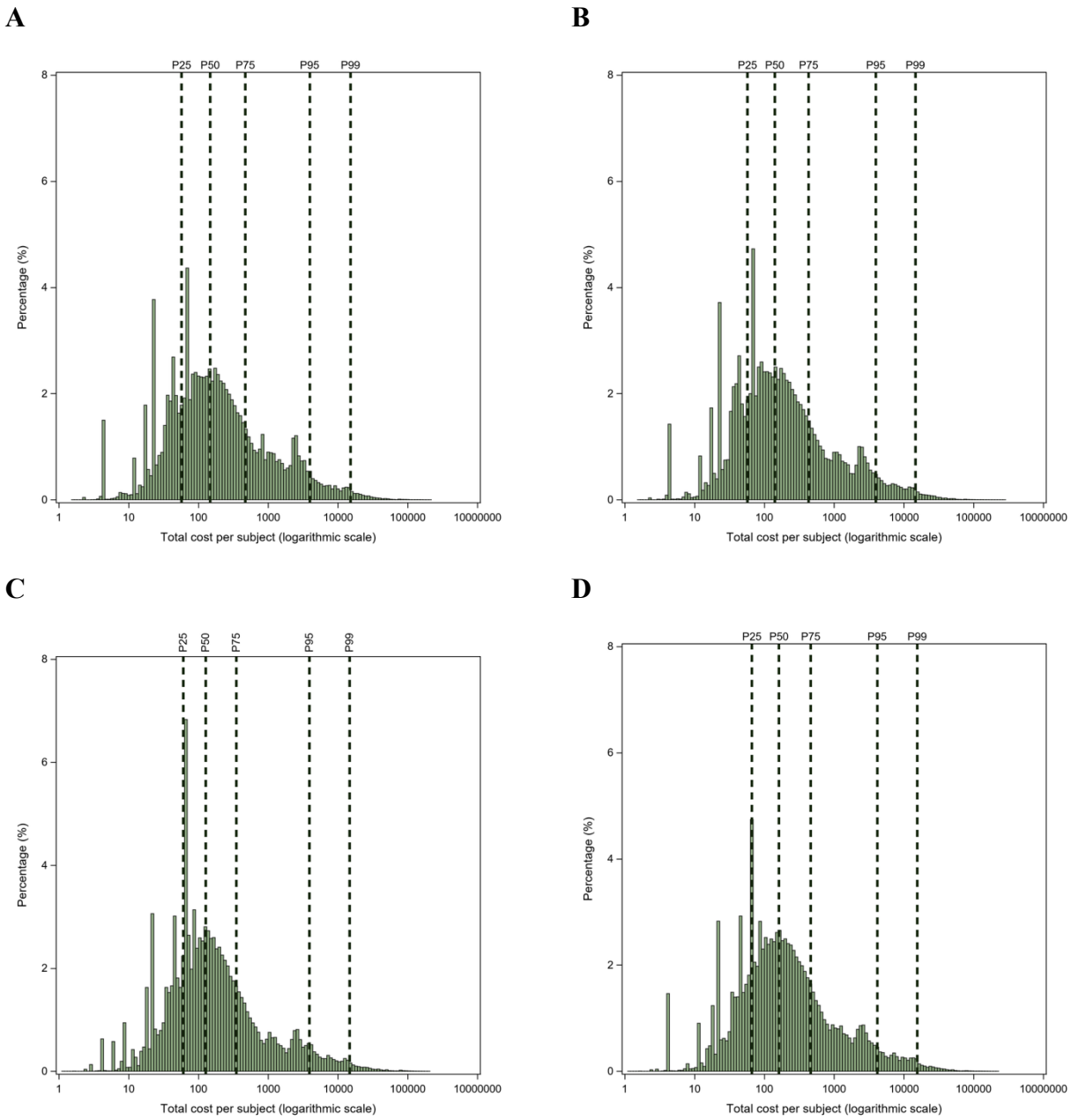


Figure 3.2. Age distribution across the seven cost groups in 2011 (panel A), in 2015 (panel B), in 2020 (panel C), and in 2023 (panel D).

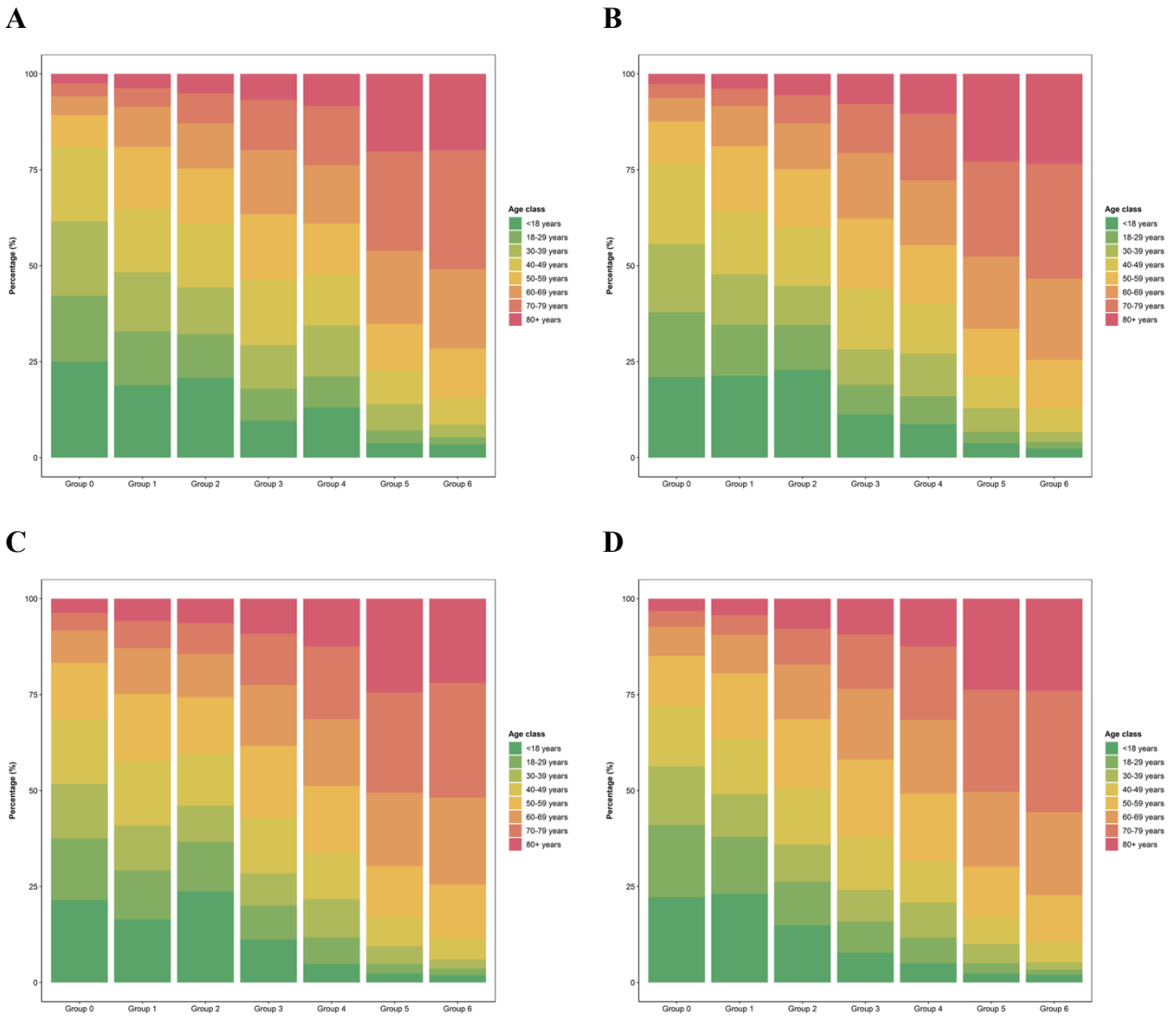


Figure 3.3. Proportion of individuals in each of the seven cost groups (blue dots) and proportion of total costs absorbed by each group (red dots) in 2011 (panel A), in 2015 (panel B), in 2020 (panel C), and in 2023 (panel D).

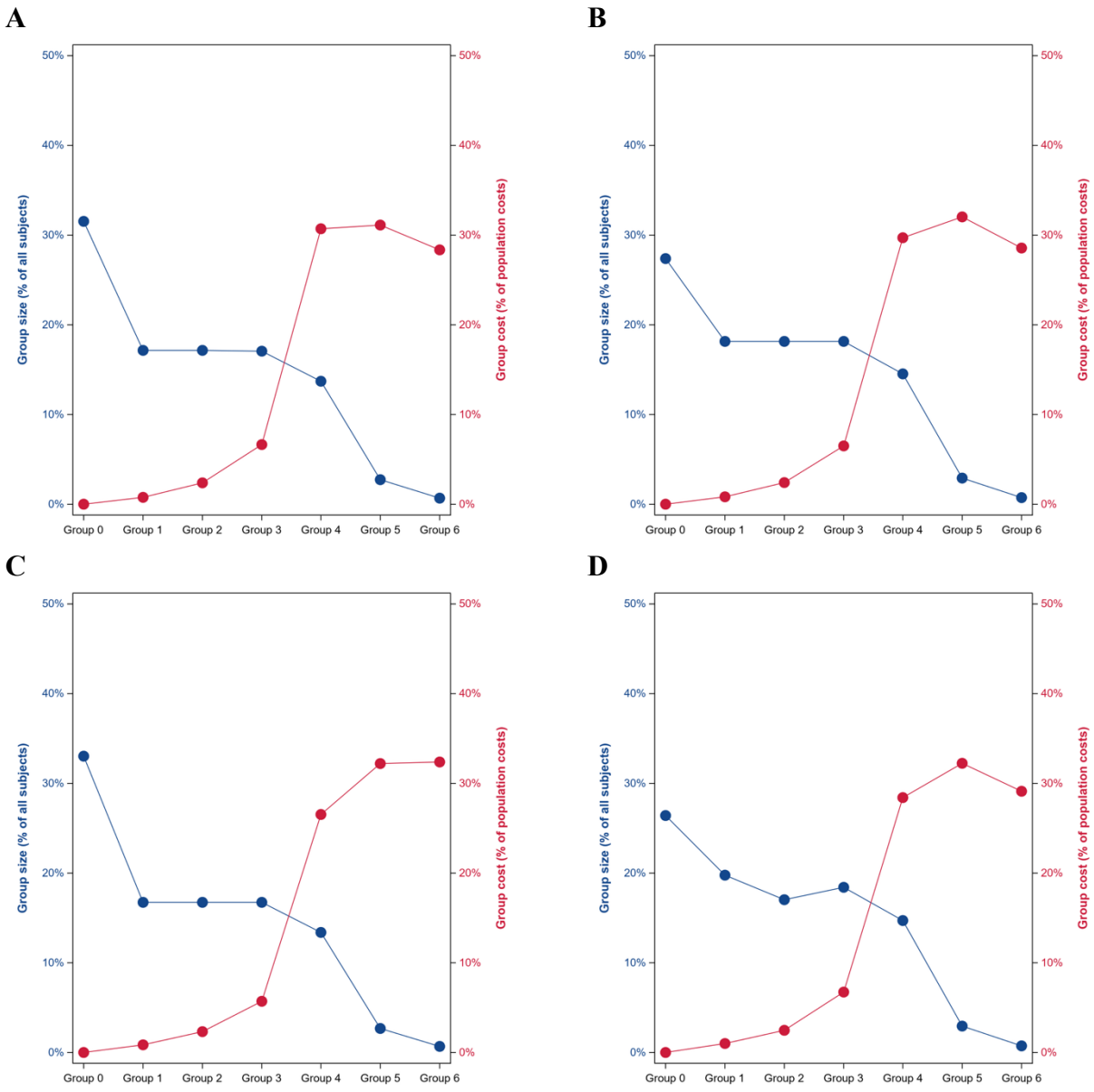
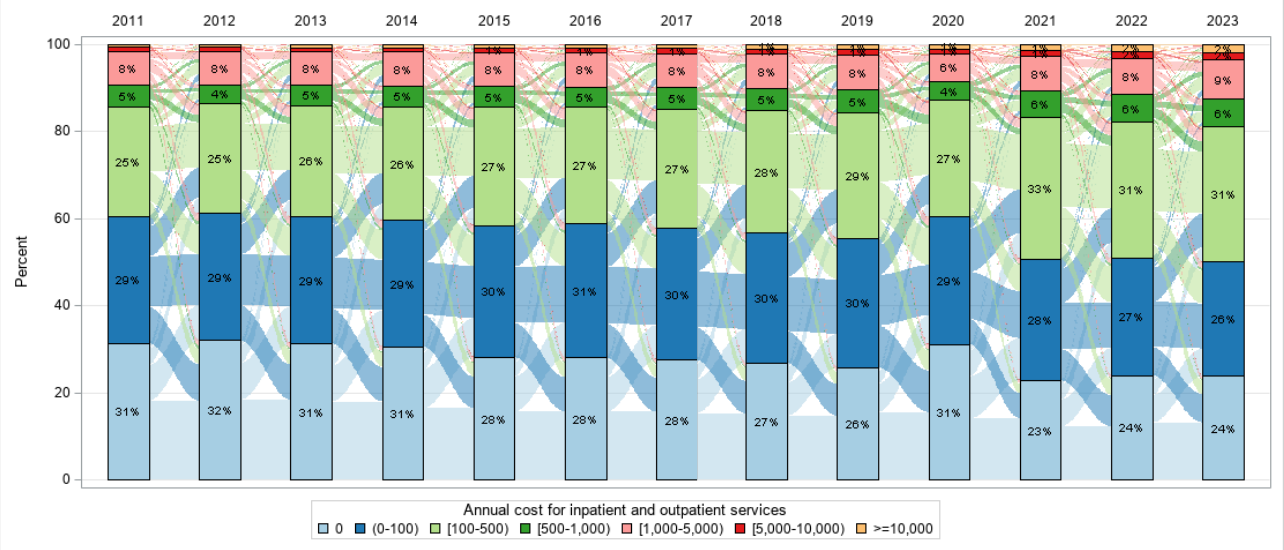


Figure 3.4. Costs for inpatient and outpatient services from 2011 to 2023 for individuals assisted by the HPA of Bergamo throughout the entire period, and transitions between different timepoints.



3.2 Cost prediction results

In this section, we present the results of the random forest algorithms developed to predict direct healthcare costs using administrative database records from the HPA of Bergamo.

First, we report the performance metrics of the predictive algorithms, providing an assessment of their accuracy in predicting costs at the individual level. We then focus on the cost predictions for the whole population as well as for subgroups of high-impact chronic patients, which represent the main objective of this work. Total and mean estimates of predicted costs for these groups were obtained by aggregating individual cost predictions.

Four different cost outcomes were considered - namely, total cost (TC), total scheduled cost (TSC), scheduled services cost (SSC), and services cost (SC). For each outcome, costs were predicted for each available year (from 2017 to 2023 for the whole population, and from 2018 to 2023 for high-impact subgroups) using random forest regression algorithms trained:

i) on the 2011-16 period (hereafter referred to as "non-updated algorithms");

ii) on the immediately preceding period, i.e., predicted costs for 2017 were derived from the algorithm trained over 2011-16, predicted costs for 2018 from the algorithm trained over 2012-17, and so forth (hereafter referred to as "updated algorithms").

Greater emphasis is given to the results obtained with the updated algorithms.

Moreover, each predictive model was trained on two alternative sets of predictors: one excluding subjects' historical cost information and one including such information.

3.2.1 Random forest algorithms' performance

For each considered scenario, we assessed the performance of the prediction algorithms using MAE and RMSE, with lower values indicating better performance (**Table 3.5** and **Table 3.6**). Overall, for all outcomes, slightly lower error values were observed when test set predictions were obtained using the updated algorithm - that is, the algorithm trained on the immediately preceding time period. We also observed a modest further reduction in both MAE and RMSE when historical cost data were included among the predictors. In general, the error values remained relatively high, suggesting that the algorithm was not particularly accurate in predicting costs at the individual level. Overall, the error metrics remained relatively stable over time, albeit at comparatively high levels. However, the extreme heterogeneity of the population complicates the interpretation of these values. For instance, there are individuals whose actual costs are zero or very low, for whom a prediction error of €1,000 may be highly significant, whereas for individuals with exceptionally high costs a comparable error would be virtually negligible.

Regarding the proportion of variance explained, which is a metric indicating how well OOB predictions account for the target variance in the training set, when historical cost information were excluded from the predictor set, values were generally slightly above 20%. Specifically, for TC, the proportion of variance explained ranged from 22.5% for the 2015-2020 training set to 24.9% for 2016-2021; for TSC, from 23.8% for 2012-2017 to 27.9% for 2016-2021; for SSC, from 22.3% for 2017-2021 to 26.4% for 2011-2016; and for SC, from 14.6% for 2015-2020 to 21.4% for 2011-2016. When historical cost data were included among the predictors, the proportion of variance explained increased substantially for all training sets, particularly for TC and TSC. For TC, the proportion of variance explained ranged from 32.1% for the 2014-2019 training set to 39.1% for 2016-2021; for TSC, from 36.4% for 2014-2019 to 46.7% for 2016-2021; for SSC, from 23.8% for 2015-2020 to 27.3% for 2011-2016; and for SC, from 15.1% for 2015-2020 to 22.2% for 2011-2016.

The left-hand panels of **Figures 3.5 and 3.6** display actual individual TC values plotted against individual predictions generated by the updated algorithms, respectively excluding or including historical cost data in the predictors' set, over the period 2017-2023. A clear tendency emerges for the algorithms to fail to capture the extreme observed values (either very low or very high). In 2023, the minimum and maximum predicted costs were €146 and €238,336, respectively, when historical cost data were not included among the predictors, and €69 and €316,240, respectively, when they were.

The right-hand panels of **Figures 3.5 and 3.6** show the distribution of absolute prediction errors, computed as the difference between predicted and actual values. For the algorithms trained without historical cost information (**Figure 3.5**), an overestimation of predicted costs relative to actual costs was observed for the majority of individuals, with a deviation between €100 and €500 recorded in 47.3% of subjects in 2017, 45.6% in 2018, 46.1% in 2019, 44.7% in 2020, 46.9% in 2021, 41.8% in 2022, and 42.9% in 2023. A further positive deviation between €500 and €5,000 involved approximately 30% of subjects, ranging from 27.6% in 2021 to 38.2% in 2022. Notably, the proportion of subjects exhibiting a deviation below -€5,000 (i.e., strong underestimation of observed costs) was consistently higher than the proportion of subjects with deviations exceeding €5,000 (i.e., strong overestimation of observed costs). Specifically, substantial underestimation (i.e., error ≤-€5,000) was observed in 8.7% of subjects compared with substantial overestimation (i.e., error ≥€5,000) in 0.7% in 2017, 8.6% vs. 0.8% in 2018, 8.7% vs. 0.9% in 2019, 7.4% vs. 0.9% in 2020, 9.1% vs. 0.5% in 2021, 8.8% vs. 1.0% in 2022, and 9.0% vs. 0.9% in 2023. It is worth highlighting the marked increase in the proportion of positive errors compared to negative ones in 2020, and the opposite trend in 2021, corresponding to the years in which predictions were heavily affected by the

COVID-19 pandemic and the associated unforeseen disruptions to the organization of the healthcare system.

Comparable proportions were observed when historical cost data were included among the predictors, suggesting that the addition of this information does not materially improve prediction accuracy (**Figure 3.6**). However, it is interesting to note that including such information among the predictors leads to an improvement in the prediction of extremely high costs, as evidenced by the clusters of observations in the upper-right area of the plots in the left-hand panel. Similar patterns were also observed for the other outcomes TSC, SSC, and SC. SC results are presented in the Supplementary materials (**Figures S1 and S2**). In subsequent paragraphs, when we turn to considerations of group-level cost predictions generated as the mean of individual predictions within each group, these strong negative deviations between predicted and actual values will inevitably have an impact, systematically pulling group-level cost predictions downward and resulting in underestimation of the true observed group-level costs.

3.2.2 Variable importance

Figures S3-S10 display the 25 variables with the greatest influence on cost prediction across all training sets, ranked according to the average increase in MSE resulting from permutation of the variable within the dataset. For the TC outcome, when historical cost information were excluded from the predictors' set, age class emerged as the variable associated with the largest average percentage increase in MSE, followed by the number of outpatient dialysis visits during the previous year, the frequency of outpatient visits over the preceding four years, the frequency of antihaemorrhagic drug use in the preceding four years, and the number of urgent hospital admissions for neoplasm in the previous year. Among the variables ranked as most important across the majority of training sets, we also observed the use of antineoplastic agents (L01CD), immunosuppressants and immunostimulants in the previous year, usually prescribed for autoimmune diseases, agents acting on the renin-angiotensin system both in the previous year and over the preceding four years, possible proxy for cardiometabolic diseases and more specifically hypertension, as well as anticoagulants and antithrombotics over the preceding four years, usually prescribed for cardiovascular diseases, highlighting the substantial role of pharmaceutical variables in predicting total costs (**Figure S3**).

When historical cost variables were included among the predictors, the most influential variable for the TC outcome, on average, was the mean total healthcare cost over the five-year period. Other variables consistently ranked among the top five in terms of average percentage increase in MSE included cost class, the number of outpatient dialysis visits during the previous year, the number of

outpatient visits in the previous year, and the frequency of laboratory test prescriptions over the preceding four years (**Figure S4**). Notably, the most influential predictors encompassed not only services utilised during the most recent year, but also variables describing the patient's historical healthcare utilisation patterns, underscoring their significant predictive impact.

For the TSC outcome, the ranking of the most influential variables across all training sets remained largely similar to that observed for TC, both when historical cost information was excluded from the predictor set and when it was included. Notably, urgent events - such as the number of emergency admissions for "Neoplasm" or "Diseases of the circulatory system" in the preceding year - continued to be strong predictors of total cost, even after costs associated with urgent services were removed from the outcome calculation (**Figure S5 and S6**).

The total number of outpatient visits in the previous year, age group, number of dialysis-related outpatient visits in the previous year, frequency of laboratory tests over the preceding four years, and frequency of outpatient visits over the preceding four years were, on average, the variables associated with the largest increase in MSE when considering costs associated to all inpatient and outpatient services (SC) as the outcome (**Figure S7**). Compared with predictions that also included drug costs, the variable indicating the patient's sex emerged as more influential. When historical cost variables were also included among the predictors, the subject's mean total healthcare cost over the five-year period appeared among the top five most influential variables (**Figure S8**). Similar rankings were also observed for the SSC outcome (**Figure S9 and S10**).

3.2.3 Costs' predictions results for the whole population

In a second step, for each outcome and for each test set, the predicted total cost for the entire population was obtained by summing the individual predicted costs, and the relative deviation between predicted and actual costs was measured through the Prediction Error (PE).

Figures 3.7 and 3.9 show, for each outcome, the PEs calculated for cost predictions obtained using the updated algorithms, without and with historical cost information among the predictors, respectively**. Figures 3.8 and 3.10** display the predicted and actual total costs for the whole population for each outcome from 2017 to 2023, again without and with historical cost information among the predictors, respectively. Results obtained without historical cost information among the predictors are also reported in **Table S2**.

When historical cost data were not included among the predictors, the algorithm showed, for all outcomes, a slight tendency toward underestimation (**Figures 3.7 and 3.8**). Specifically, predictions were accurate up to 2019, with PE values close to zero, ranging from -2.7% (PE for TSC in 2017) to 0.4% (PE for TSC in 2019). In 2020 and 2021, prediction accuracy was strongly affected by the

unpredictability of the disruptions to the NHS during the pandemic period, leading to substantial overestimation in 2020 (PE = 9.7% for TC, PE = 16.3% for TSC, PE = 24.4% for SSC, and PE = 11.8% for SC) and marked underestimation in 2021 (PE = -13.6% for TC, PE = -19.3% for TSC, PE = -26.3% for SSC, and PE = -16.2% for SC). From 2022 onward, PE values returned to pre-pandemic levels. In 2022, a slight overestimation was observed (PE = 0.4% for TC, PE = 0.8% for TSC, PE = 1.2% for SSC, and PE = 0.6% for SC), whereas in 2023 the algorithm reverted to its tendency toward underestimation, with a predicted TC of €1,103,322,372 versus an actual value of €1,111,657,382 (PE = -0.7%), a predicted TSC of €929,510,281 versus an actual value of €939,009,798 (PE = -1.0%), a predicted SSC of €556,551,420 versus an actual value of €563,163,596 (PE = -1.2%), and a predicted SC of €733,255,959 versus an actual value of €735,811,181 (PE = -0.3%).

The results were similar when historical cost information were included among the predictors (**Figures 3.9 and 3.10**). PE values were slightly lower from 2017 to 2019 compared to those derived by the algorithm without cost information, while for 2022 and 2023 they were slightly higher - though still close to zero. This suggests that including cost data among the predictors may increase the model's sensitivity to the instability introduced during the 2020-2021 biennium.

Supplementary Figures S11 and S12 show the PE values based on predicted costs derived from the non-updated algorithms trained over the period 2011-16. Without updating the random forest algorithms, prediction accuracy remained comparable to that of the updated algorithms up to 2019, while PE values in 2021 were closer to zero, as they were not affected by training on data influenced by the pandemic year. However, the prediction accuracy for 2022 and 2023 remained comparable to that in 2021, without a return of PE values toward zero, as observed when using the updated algorithms. This suggests that the impact of the COVID-19 pandemic did not only affect the single year of 2020 but persisted in subsequent years, also leading to changes in the cost drivers.

3.2.4 Costs' predictions results for high-impact segments

CKD - Dialysis

Figure S13 presents the actual individual TC values plotted against the corresponding predictions (left-hand panels) generated by the updated algorithms excluding historical cost information from the predictor set, as well as the distribution of absolute prediction errors (right-hand panels), specifically for the subgroup of patients on dialysis. Predicted costs fall within a remarkably narrow range centred around the mean of the actual cost distribution. Consequently, the error distribution appears markedly different from that observed for the general population (**Figure 3.5**): there are high proportions of individuals exhibiting very large deviations between predicted and actual values, both positive (i.e., substantial overestimation of actual costs) and negative (i.e., substantial underestimation).

Figures 3.11, 3.13, 3.15, and 3.17 show the PEs calculated for dialysis patients obtained using the updated algorithms for TC, TSC, SSC, and SC, respectively, with no historical cost information in the predictor set. As mentioned in the Methods section, different definitions were applied to identify the group of individuals affected by the chronic condition of interest. **Figures 3.12, 3.14, 3.16, and 3.18** show the mean annual predicted and mean annual actual costs for this subgroup of patients for TC, TSC, SSC, and SC, respectively. Results are also reported in **Table S3**.

We observe that for all outcomes, PE values tended to be below zero, indicating a clear tendency to predict lower mean costs than those actually observed. However, the variation range of PE values was relatively narrow, with most estimates close to zero and negative peaks recorded for cost predictions in 2021. Considering the different definitions used to identify these patients - namely, "all": individuals with CKD - Dialysis, with or without other conditions; "primary condition": individuals for whom CKD - Dialysis was the main condition; "only": individuals with CKD - Dialysis as their sole chronic condition; "primary condition; ≤2 comorbidities": individuals with CKD - Dialysis as the main condition and at most two other chronic conditions; and "primary condition; >2 comorbidities": individuals with CKD-Dialysis as the main condition and at least three other chronic conditions - the temporal trend of PE values appears very similar across groups, except for the "only" group. It should be noted that the number of individuals affected solely by CKD - Dialysis was very small (the highest count was recorded in 2019, n = 20). Moreover, a major update to the BDA classification algorithm for identifying patients with this chronic condition occurred between 2019 and 2020. This change led to a substantial reduction in the number of individuals classified as having CKD - Dialysis from 2021 onwards. This is because costs were predicted for the year following the five-year observation period (thus, for example, in 2021, we predicted costs for individuals classified as having CKD - Dialysis in 2020). As a result, for instance, the number of individuals classified as "CKD - Dialysis: all" decreased from 735 in 2020 to 197 in 2021. Therefore, caution should be exercised when comparing observed PE values before and after 2020.

Across all outcomes and patients' groups, the deviation between predicted and actual costs decreased in 2022 compared to 2021, before increasing again in 2023. In 2023, PE values remained closer to zero for the "all" and "primary condition; >2 comorbidities" groups (group "all", 2022: mean predicted TC [variability interval] = €45,690 [41,663 - 50,143] vs. mean actual TC = €46,759, PE = -2.3%; group "all", 2023: €44,128 [40,421 - 47,605] vs. €46,254, -4.6%; group "primary condition; >2 comorbidities", 2022: €46,393 [41,708 - 51,127] vs. €47,053, -1.4%; group "primary condition; >2 comorbidities", 2023: €44,467 [39,442 - 49,750] vs.€45,494, -2.3%; **Figure 3.11** and **Table S3**).

Finally, comparable results were obtained when including historical cost information among the predictors. Overall, for the majority of group definitions and all outcomes, PE values tended to be

slightly closer to zero; however, the inclusion of cost predictors did not result in a substantial improvement in cost predictions (**Figures S14-S21**).

It is worth noting that for individuals affected by this chronic condition, no substantial difference was observed in the prediction errors in 2020 and 2021 compared with other years. This suggests that, even during the period impacted by the COVID-19 pandemic, this patient group was particularly safeguarded, and their management did not undergo substantial changes. This can also be confirmed by the results obtained with the non-updated algorithms: the deviation between predicted and actual costs remained comparable before and after 2020 (PE values for the "all" group are shown in **Figures S22 and S23**).

Type 2 diabetes

Figure S24 presents the actual individual TC values plotted against the corresponding predictions (left-hand panels) generated by the updated algorithms excluding historical cost information from the predictor set, as well as the distribution of absolute prediction errors (right-hand panels), for the subgroup of patients with type 2 diabetes. In contrast with individuals on dialysis, for these patients the distribution of prediction errors more closely resembles that observed in the general population, displaying a peak in positive deviations between €1,000 and €5,000, and a higher proportion of subjects with errors <-€5,000 than those with errors >€5,000.

Figures 3.19, 3.21, 3.23, and 3.25 show the PEs calculated for type 2 diabetes patients obtained using the updated algorithms for TC, TSC, SSC, and SC, respectively, with no historical cost information in the predictor set. **Figures 3.20, 3.22, 3.24, and 3.26** show the mean annual predicted and mean annual actual costs for this subgroup of patients for TC, TSC, SSC, and SC, respectively. Results are also reported in **Table S4**.

For the TC and TSC outcomes, temporal trends in PE values were broadly consistent across all groups of patients with diabetes, with the exception of the "primary condition; >2 comorbidities" group. Predicted costs were generally underestimated compared to the actual costs for all years, except in 2020 (and, for TSC, also in 2023 for the "only" group). The underestimation was particularly pronounced in 2021, reflecting the impact of algorithms training on costs affected by the COVID-19 pandemic in 2020, and reached a peak of -20.4% for TSC in the "only" group. In the "all" group, overestimation in 2020 was less marked than in the other groups (group "all": PE for TC = 1.5%, PE for TSC = 2.8%; group "primary condition": PE for TC = 7.8%, PE for TSC = 11.5%; group "only": PE for TC = 7.0%, PE for TSC = 15.1%; group "primary condition; ≤2 comorbidities": PE for TC = 11.7%, PE for TSC = 14.2%; **Figures 3.19 and 3.21**, **Table S4**).

In 2023, PE values returned closer to zero for all four groups. For TC (**Figures 3.19 and 3.20**, **Table S4**): i) group "all": mean predicted cost (variability interval) = €3,455 (3,337 - 3,571) and mean actual cost = €3,646, PE = -5.2%; ii) group "primary condition": €2,321 (2,217 - 2,431) and €2,407, PE = -5.2%; iii) group "only": €1,506 (1,333 - 1,665) and €1,520, PE = -1.0%; iv) group "primary condition; ≤2 comorbidities": €2,547 (2,434 - 2,651) and €2,657, PE = -4.1%. For TSC (**Figures 3.21 and 3.22**, **Table S4**): i) group "all": €2,886 (2,790 - 2,985) and €2,983, PE = -3.2%; ii) group "primary condition": €1,961 (1,871 - 2,056) and €1,987, PE = -1.3%; iii) group "only": €1,282 (1,142 - 1,411) and €1,273, PE = 0.7%; iv) group "primary condition; ≤2 comorbidities": €2,154 (2,053 - 2,258) and €2,191, PE = -1.7%.

In contrast, for the "primary condition; >2 comorbidities" group, PE values for TC were consistently negative across all years, whereas for TSC an overestimation of predicted costs was recorded in 2022. This group also experienced a sharp reduction in sample size between 2020 and 2021 (from 2,622 to 669), attributable to changes in the BDA classification algorithm, warranting caution in interpreting the results. Further evidence of a change in the identification of individuals with more than three chronic conditions and type 2 diabetes mellitus as the primary condition is provided by the marked increase in mean cost per patient observed from 2021 onwards (mean actual TC: 2018, €4,204; 2019, €4,262; 2020, €4,031; 2021, €6,320; 2022, €6,377; 2023, €6,516; and mean actual TSC: 2018, €3,307; 2019, €3,287; 2020, €3,033; 2021, €4,983; 2022, €5,088; 2023, €5,195; **Figure 3.20 and Table S4**).

For the SC outcome (**Figures 3.25 and 3.26**, **Table S4**), the pattern of PE values over time mirrored that observed for TC and TSC, although the deviation between mean predicted and observed values was greater in 2020. In the "all" group, the mean predicted cost (variability interval) was €1,936 (1,869 - 2,012) compared with a mean actual cost of €1,693 (PE = 14.4%), whereas for TC in the same year the corresponding figures were €2,946 (2,856 - 3,048) and €2,903 (PE = 1.5%).

For SSC (**Figures 3.23 and 3.24**, **Table S4**), deviations of PE values from zero were even more pronounced in 2020 and 2021, with a peak of 44.6% in 2020 and -22.0% in 2021 for the "primary condition; ≤2 comorbidities" group. Compared with other outcomes, the algorithm showed a greater tendency to overestimate costs: in 2019, 2022, and 2023, PE values were slightly above zero for most groups.

Finally, the inclusion of historical cost information yielded results broadly consistent with those obtained when these data were excluded. The deviation between predicted and observed values decreased slightly across the most of groups and outcomes, but without materially altering the overall conclusions (**Figures S25-S32**).

Figures S33 and S34 present the PE values for the "all" group derived by the non-updated algorithms.

Heart failure

Figure S35 presents the actual individual TC values plotted against the corresponding predictions (left-hand panels) generated by the updated algorithms excluding historical cost information from the predictor set, as well as the distribution of absolute prediction errors (right-hand panels), for the subgroup of patients with heart failure. The error distribution is consistent with that observed in the general population, but the imbalance between the proportions of subjects with large negative versus large positive errors is even more pronounced: annually, some individuals exhibit prediction errors below -€100,000, whereas no subjects display correspondingly large positive deviations.

Figures 3.27, 3.29, 3.31, and 3.33 show the PEs calculated for patients with heart failure obtained using the updated algorithms for TC, TSC, SSC, and SC, respectively, with no historical cost information in the predictor set. **Figures 3.28, 3.30, 3.32, and 3.34** show the mean annual predicted and mean annual actual costs for this subgroup of patients for TC, TSC, SSC, and SC, respectively. Results are also reported in **Table S5**.

It should be noted that, for three groups, there was a marked change in sample size between 2020 and 2021, again reflecting a modification of the BDA algorithm for identifying chronic patients: the group for which heart failure is the only chronic condition, and the two groups in which heart failure is the primary condition and patients have ≤2 or >2 additional chronic conditions. For these groups, an increase in the mean cost per subject (across all outcomes) was also observed from 2021 onwards. Focusing on PE values for these three groups from 2021 onwards, within-group trends were stable across the four outcomes. In the "only" group, the deviation between mean predicted and actual values was pronounced in 2021 and 2022, peaking at -31.0% for SSC, before returning toward zero in 2023, with a value of -5.4% for TSC. A similar pattern was observed in the "primary condition; ≤2 comorbidities" group, where the closest value to zero in 2023 was for SSC (-11.0%). Conversely, in the "primary condition; >2 comorbidities" group, predicted costs in 2023 exceeded those observed, with PE values ranging from 9.1% for TC to 21.5% for SSC.

For the "all" group, all outcomes exhibited cost underestimation in 2018-2019, followed by an increase in PE values in 2020, exceeding zero for SSC and SC (10.5% and 5.1%, respectively) but remaining below zero for TC and TSC (-2.8% for both). In 2021, there was a substantial underestimation of costs, with PE values ranging from -23.7% for SC to -17.9% for TSC. This was followed by a gradual reduction in the deviation between predicted and actual costs in 2022 and 2023. In 2023, the mean predicted cost (variability interval) was €5,116 (4,895 - 5,367) versus a mean actual cost of €5,558 (PE = -8.0%) for TC (**Figure 3.27 and 3.28, Table S5**); €3,952 (3,765 - 4,164) versus €4,184 (PE = -5.5%) for TSC (**Figure 3.29 and 3.30, Table S5**); €2,330 (2,233 - 2,440) versus €2,414

(PE = -3.4%) for SSC (**Figure 3.31 and 3.32, Table S5**); and €3,485 (3,328 - 3,649) versus €3,788 (PE = -8.0%) for SC (**Figure 3.33 and 3.34, Table S5**).

Finally, for the "primary condition" group, the temporal pattern of PE values closely mirrored that observed for the "all" group, although PE values in 2020 were higher and above zero for all outcomes (2.4% for both TC and TSC, 20.9% for SSC, and 11.8% for SC).

Also for heart failure, the inclusion of historical cost data among the predictors did not result in any substantial changes in the results (**Figures S36-S43**).

Figures S44 and S45 present the PE values for the "all" group derived by the non-updated algorithms.

Parkinson's disease/Parkinsonian syndromes

Figure S46 presents the actual individual TC values plotted against the corresponding predictions (left-hand panels) generated by the updated algorithms excluding historical cost information from the predictor set, as well as the distribution of absolute prediction errors (right-hand panels), for the subgroup of patients with Parkinson's disease/Parkinsonian syndromes. As observed among patients on dialysis, the predicted costs for this group fall within a relatively narrow range, albeit less sharply defined than that seen in CKD patients, partly due to the larger sample size. Indeed, the absolute prediction error distributions reveals that although most individuals exhibit a positive error between €1,000 and €5,000, there remains a non-negligible proportion of subjects for whom costs are severely underestimated (for example, in 2023, 7.8% of individuals had prediction errors below -€10,000, compared with 0.7% with errors above €10,000).

Figures 3.35, 3.37, 3.39, and 3.41 show the PEs calculated for patients with Parkinson's disease/Parkinsonian syndromes obtained using the updated algorithms for TC, TSC, SSC, and SC, respectively, with no historical cost information in the predictor set. **Figures 3.36, 3.38, 3.40, and 3.42** show the mean annual predicted and mean annual actual costs for this subgroup of patients for TC, TSC, SSC, and SC, respectively. Results are also reported in **Table S6**.

For TC and TSC, mean predicted costs substantially underestimated mean actual costs across all subgroups. For all groups - except "primary condition; >2 comorbidities" - the characteristic step change was observed between 2020 (higher PE value) and 2021 (lower PE value), followed by a gradual convergence of values toward zero; however, all PE values remained below -10%. For both TC and TSC, the group with the smallest overall deviation between predicted and actual values was the "all" group. In 2023, for this group, the mean predicted cost (variability interval) was €3,768 (3,349 - 4,434) versus a mean actual cost of €4,698 (PE = -19.8%) for TC, and €3,181 (2,762 - 3,844) versus €3,878 (PE = -18.0%) for TSC (**Figures 3.35-3.38, Table S6**).

The group of patients with Parkinson's disease as the primary condition and at least three additional chronic conditions showed a temporal PE pattern distinct from all other groups. For TC, the PE was -9.6% in 2020, increased to -4.0% in 2021, and then dropped sharply to -24.4% in 2022 (**Figure 3.35**). For TSC, PE values were very similar in 2020 and 2021 (-16.1% and -17.5%, respectively), but in this case as well, the deviation between predicted and actual costs further increased in 2022 (PE = -22.5%; **Figure 3.37**). These findings suggest that, despite Parkinson's disease being the primary condition, the presence of three or more additional chronic diseases makes this group too heterogeneous and insufficiently well-characterized to enable reliable and consistent cost predictions over time.

With regard to cost outcomes considering only inpatient and outpatient services, the temporal patterns were similar to those observed for TC and TSC; however, in 2020 the deviation in PE values was more pronounced, exceeding zero, compared with other years. For both SSC and SC, the PE values were consistently below -10%, and the "all" group still showed the smallest prediction errors overall. In 2023, for this group, the mean predicted cost (variability interval) was €1,912 (1,693 - 2,144) versus a mean actual cost of €2,142 (PE = -10.7%) for SSC (**Figures 3.39 and 3.40, Table S6**), and €2,543 (2,257 - 2,854) versus €2,961 (PE = -14.1%) for SC (**Figures 3.41 and 3.42, Table S6**).

For Parkinson's disease as well, the inclusion of historical cost information did not result in any substantial change in the results (**Figures S47-S54**).

Figures S55 and S56 present the PE values for the "all" group derived by the non-updated algorithms.

Active neoplasia

Finally, **Figure S57** presents the actual individual TC values plotted against the corresponding predictions (left-hand panels) generated by the updated algorithms excluding historical cost information from the predictor set, as well as the distribution of absolute prediction errors (right-hand panels), for the subgroup of patients with active neoplasia. Again, a peak is observed among subjects with errors in the €1,000 - 5,000 range; however, a substantial proportion of individuals exhibit severely underestimated actual costs (in 2023, 10.7% of subjects had error below -€10,000, compared with only 3.8% with error exceeding €10,000).

Figures 3.43, 3.45, 3.47, and 3.49 show the PEs calculated for patients with active neoplasia obtained using the updated algorithms for TC, TSC, SSC, and SC, respectively, with no historical cost information in the predictor set. **Figures 3.44, 3.46, 3.48, 3.50** show the mean annual predicted and mean annual actual costs for this subgroup of patients for TC, TSC, SSC, and SC, respectively. Results are also reported in **Table S7**.

As observed for Parkinson's disease/Parkinsonian syndromes, mean predicted costs for patients with neoplasms tended to underestimate mean actual costs consistently across all years, groups, and outcomes, with the exception of the "primary condition; >2 comorbidities" group for SSC and SC. For TC and TSC (**Figures 3.43-3.46**, **Table S7**), PE values were below -12% for all groups, again with the exception of patients with an active neoplasm as the primary condition and more than two other chronic diseases, who, on average, exhibited smaller deviations between predicted and actual values than the other groups, although their temporal PE trends were similar.

It is also noteworthy that in 2020 and 2021, PE values remained broadly in line with those from other years, suggesting continuity in the care and protection of these patients even during the pandemic period. For example, considering TSC as the outcome for the "primary condition" group, the PE decreased from -24.3% in 2019 (mean predicted cost [variability interval] = €4,990 [4,726-5,252] vs. mean actual cost = €6,590), to -19.5% in 2020 (€5,037 [4,794-5,296] vs. €6,258), then dropped to -31.6% in 2021 (€3,948 [3,737 - 4,167] vs. €5,771), before improving to -13.4% in 2022 (€5,181 [4,904 - 5,495] vs. €5,981), and finally reaching -18.7% in 2023 (€4,634 [4,393 - 4,862] vs. €5,698; **Figure 3.45 and 3.46**, **Table S7**).

The temporal trend in PE values from 2018 to 2023 was similar for SC and SSC (**Figures 3.47 and 3.49**); however, for these outcomes the deviation between mean predicted and actual costs was generally smaller than for TC and TSC. For example, considering SSC as the outcome for the "primary condition" group, the PE was -16.8% in 2019 (mean predicted cost [variability interval] = €2,511 [2,366 - 2,663] vs. mean actual cost = €3,020), -11.1% in 2020 (€2,501 [2,356 - 2,658] vs. €2,815), 29.9% in 2021 (€1,900 [1,795 - 2,004] vs. €2,710), -7.3% in 2022 (€2,664 [2,519 - 2,831] vs. €2,873), and -11.0% in 2023 (€2,390 [2,277 - 2,502] vs. €2,686; **Figure 3.47 and 3.48**, **Table S7**).

The deviation between predicted and observed values decreased for all outcomes and all groups - except for some estimates in the "primary condition; >2 comorbidities" group - when historical cost data were included among the predictors. However, even in this case, no substantial advantage was observed from including these variables in the predictor set (**Figures S58-S65**).

Figures S66 and S67 present the PE values for the "all" group derived by the non-updated algorithms.

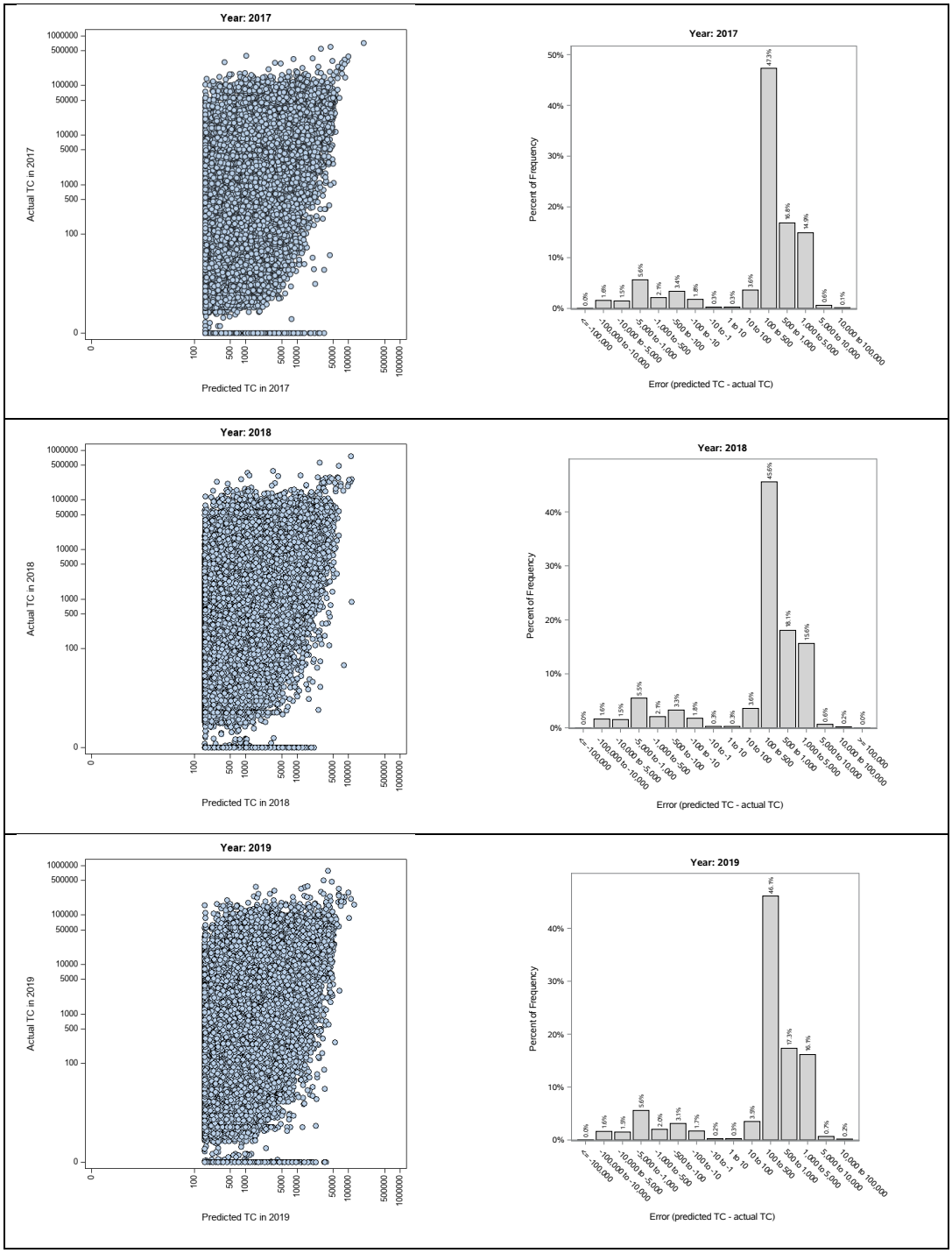
Table 3.5 Mean absolute error (MAE) and root mean squared error (RMSE) for all outcomes and for all test set. Cost predictions were derived from non-updated (2011-16) and updated algorithms, excluding historical cost information from the predictors' set.

TOTAL COST (TC)				TOTAL SCHEDULED COST (TSC)			
		Training set				Training set	
Test set		Non-updated	Updated	Test set		Non-updated	Updated
2012-17	MAE (€)	1191	1191	**2012-17**	MAE (€)	926	926
	RMSE (€)	4115	4115		RMSE (€)	3553	3553
2013-18	MAE (€)	1227	1231	**2013-18**	MAE (€)	965	971
	RMSE (€)	4248	3058		RMSE (€)	3751	3705
2014-19	MAE (€)	1240	1241	**2014-19**	MAE (€)	975	980
	RMSE (€)	4313	4255		RMSE (€)	3737	3674
2015-20	MAE (€)	1239	1229	**2015-20**	MAE (€)	934	929
	RMSE (€)	4379	4286		RMSE (€)	3583	3472
2016-21	MAE (€)	1181	1121	**2016-21**	MAE (€)	904	826
	RMSE (€)	4392	4293		RMSE (€)	3770	3658
2017-22	MAE (€)	1278	1310	**2017-22**	MAE (€)	994	1024
	RMSE (€)	4755	4608		RMSE (€)	4104	3941
2018-23	MAE (€)	1297	1320	**2018-23**	MAE (€)	1024	1043
	RMSE (€)	4787	4615		RMSE (€)	4218	4022
SCHEDULED SERVICES COST (SSC)				**SERVICES COST (SC)**			
		Training set				Training set	
Test set		Non-updated	Updated	Test set		Non-updated	Updated
2012-17	MAE (€)	685	685	**2012-17**	MAE (€)	962	962
	RMSE (€)	2283	2283		RMSE (€)	3068	3068
2013-18	MAE (€)	697	698	**2013-18**	MAE (€)	971	970
	RMSE (€)	2345	2343		RMSE (€)	3054	3058
2014-19	MAE (€)	704	707	**2014-19**	MAE (€)	982	980
	RMSE (€)	2334	2331		RMSE (€)	3143	3143
2015-20	MAE (€)	671	675	**2015-20**	MAE (€)	994	991
	RMSE (€)	2185	2174		RMSE (€)	3320	3310
2016-21	MAE (€)	654	593	**2016-21**	MAE (€)	943	903
	RMSE (€)	2268	2271		RMSE (€)	3169	3175
2017-22	MAE (€)	720	755	**2017-22**	MAE (€)	1019	1054
	RMSE (€)	2475	2467		RMSE (€)	3422	3415
2018-23	MAE (€)	735	765	**2018-23**	MAE (€)	902	1054
	RMSE (€)	2501	2488		RMSE (€)	3367	3345

Table 3.6 Mean absolute error (MAE) and root mean squared error (RMSE) for all outcomes and for all test set. Cost predictions were derived from non-updated (2011-16) and updated algorithms, including historical cost information in the predictors' set.

TOTAL COST (TC)				TOTAL SCHEDULED COST (TSC)			
		Training set				Training set	
Test set		Non-updated	Updated	Test set		Non-updated	Updated
2012-17	MAE (€)	1157	1157	**2012-17**	MAE (€)	895	895
	RMSE (€)	3879	3879		RMSE (€)	3270	3270
2013-18	MAE (€)	1192	1196	**2013-18**	MAE (€)	933	937
	RMSE (€)	3990	3963		RMSE (€)	3449	3423
2014-19	MAE (€)	1204	1203	**2014-19**	MAE (€)	941	944
	RMSE (€)	4024	3977		RMSE (€)	3394	3347
2015-20	MAE (€)	1201	1187	**2015-20**	MAE (€)	898	889
	RMSE (€)	4029	3975		RMSE (€)	3142	3077
2016-21	MAE (€)	1151	1083	**2016-21**	MAE (€)	875	791
	RMSE (€)	3951	3888		RMSE (€)	3251	3172
2017-22	MAE (€)	1229	1240	**2017-22**	MAE (€)	949	959
	RMSE (€)	4323	4242		RMSE (€)	3594	3504
2018-23	MAE (€)	1242	1256	**2018-23**	MAE (€)	973	984
	RMSE (€)	4344	4250		RMSE (€)	3708	3599
SCHEDULED SERVICES COST (SSC)				**SERVICES COST (SC)**			
		Training set				Training set	
Test set		Non-updated	Updated	Test set		Non-updated	Updated
2012-17	MAE (€)	681	681	**2012-17**	MAE (€)	955	955
	RMSE (€)	2269	2269		RMSE (€)	3053	3053
2013-18	MAE (€)	693	694	**2013-18**	MAE (€)	965	965
	RMSE (€)	2332	2333		RMSE (€)	3042	3050
2014-19	MAE (€)	700	703	**2014-19**	MAE (€)	976	975
	RMSE (€)	2323	2321		RMSE (€)	3133	3134
2015-20	MAE (€)	668	671	**2015-20**	MAE (€)	989	987
	RMSE (€)	2172	2163		RMSE (€)	3310	3301
2016-21	MAE (€)	653	590	**2016-21**	MAE (€)	943	900
	RMSE (€)	2255	2261		RMSE (€)	3159	3164
2017-22	MAE (€)	717	749	**2017-22**	MAE (€)	1013	1046
	RMSE (€)	2464	2461		RMSE (€)	3412	3412
2018-23	MAE (€)	732	760	**2018-23**	MAE (€)	1019	1047
	RMSE (€)	2492	2481		RMSE (€)	3343	3336

Figure 3.5 Scatter plots of actual TC vs. predicted TC (left-hand panels) and distributions of absolute errors (right-hand panels), in the whole population. Cost predictions derived from updated algorithms, excluding historical cost information from the predictors' set.



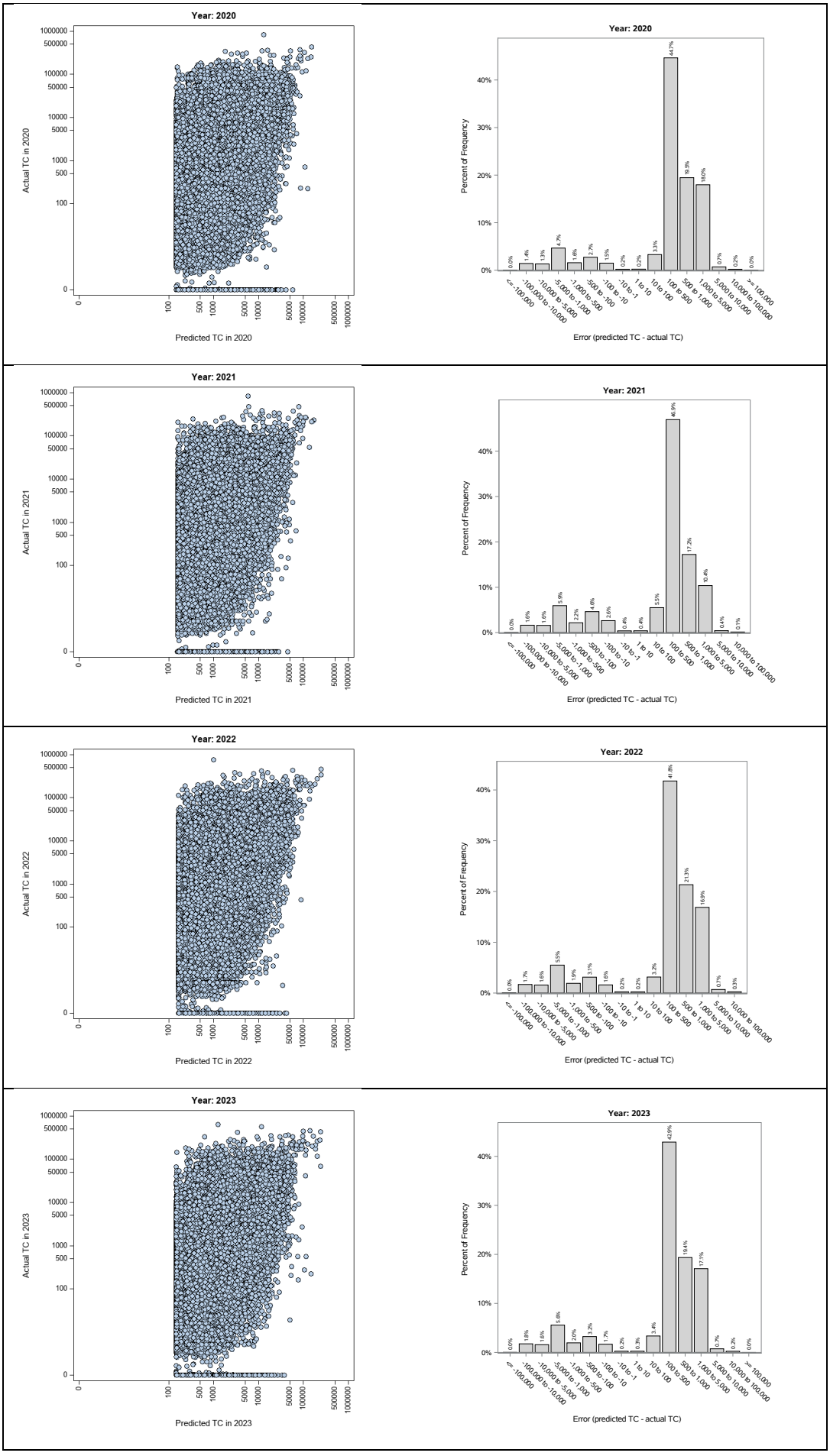
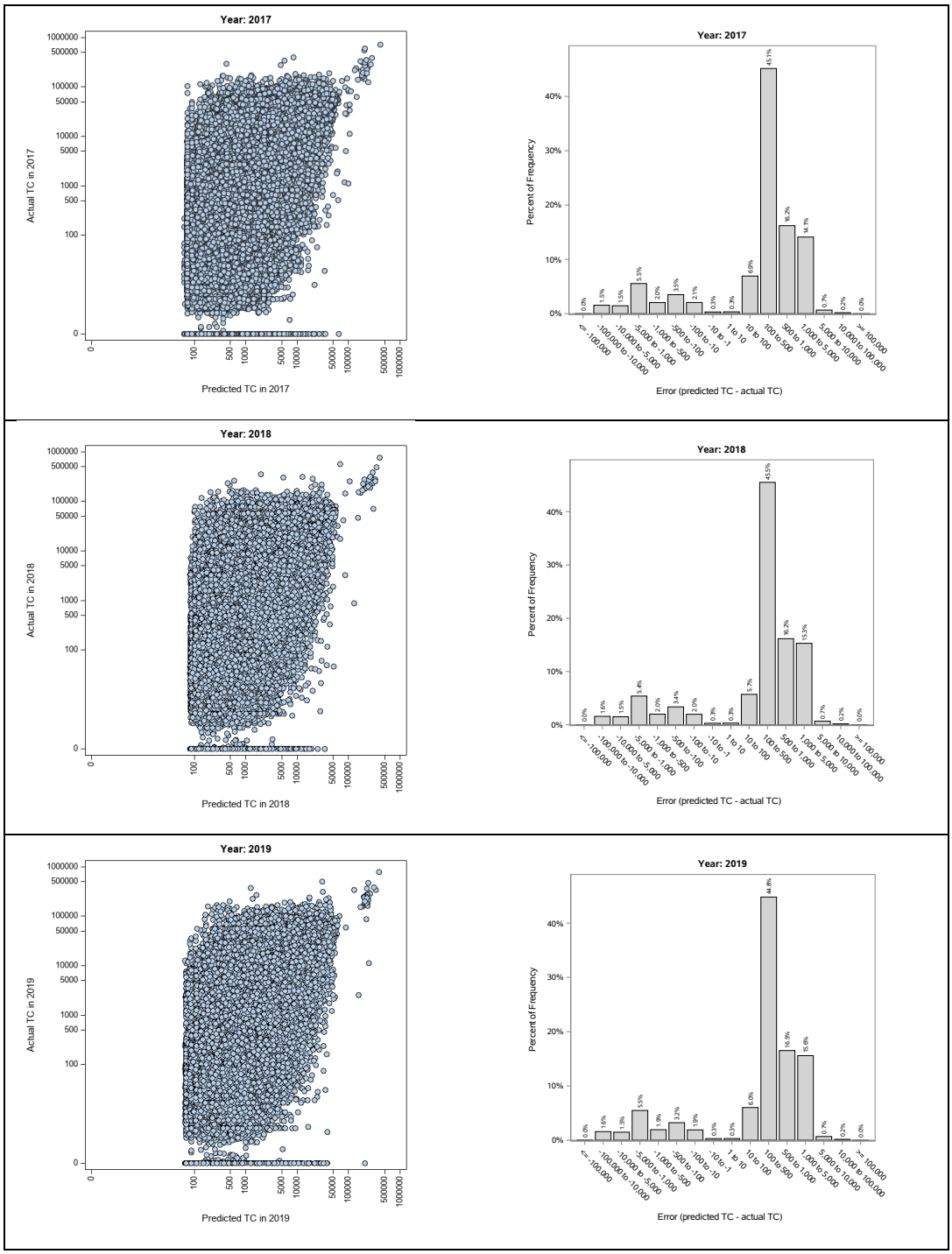
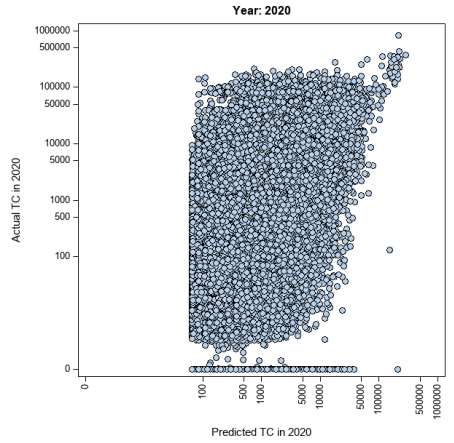
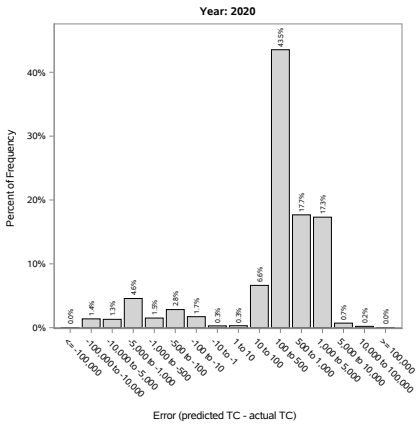
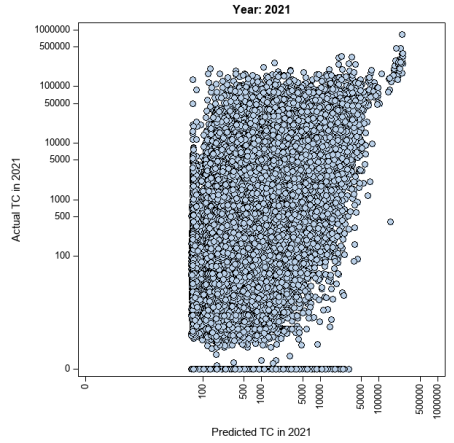

Year: 2020
Actual TC in 2020
Predicted TC in 2020
Year: 2020
Percent of Frequency
Error (predicted TC - actual TC)
Year: 2021
Actual TC in 2021
Predicted TC in 2021
Year: 2021
Percent of Frequency
Error (predicted TC - actual TC)
Year: 2022
Actual TC in 2022
Predicted TC in 2022
Year: 2022
Percent of Frequency
Error (predicted TC - actual TC)
Year: 2023
Actual TC in 2023
Predicted TC in 2023
Year: 2023
Percent of Frequency
Error (predicted TC - actual TC)

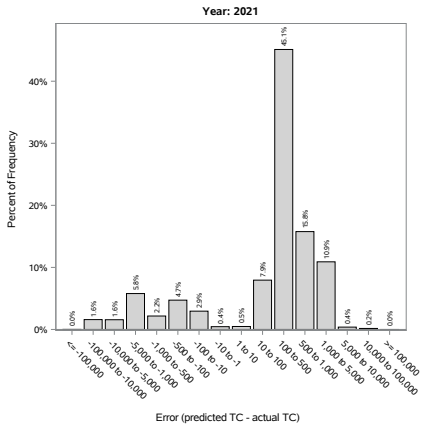
Figure 3.6 Scatter plots of actual TC vs. predicted TC (left-hand panels) and distributions of absolute errors (right-hand panels), in the whole population. Cost predictions derived from updated algorithms, including historical cost information in the predictors' set.

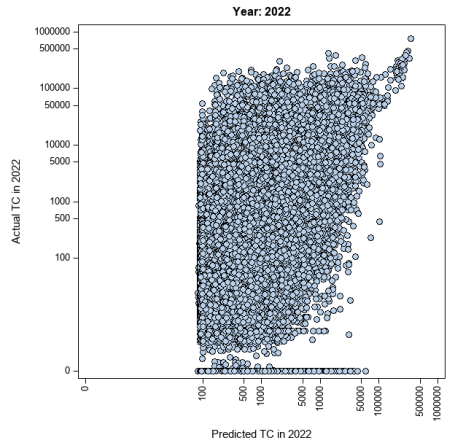


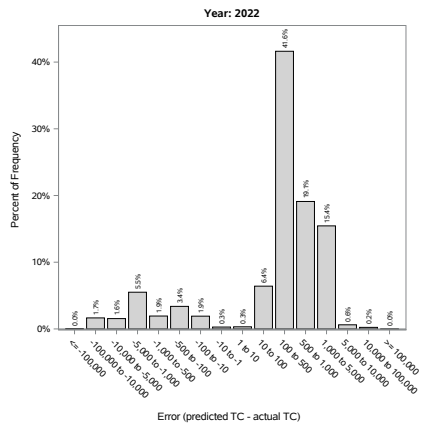

Year: 2020
Actual TC in 2020
Predicted TC in 2020

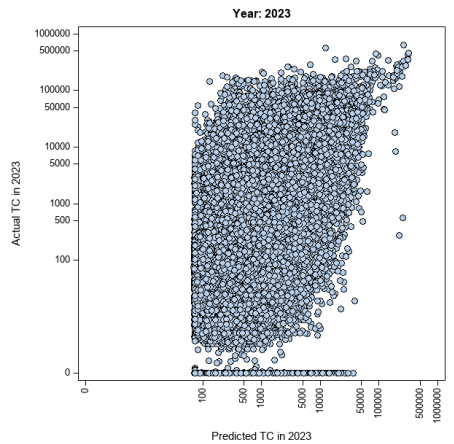

Year: 2020
Percent of Frequency
Error (predicted TC - actual TC)


Year: 2021
Actual TC in 2021
Predicted TC in 2021


Year: 2021
Percent of Frequency
Error (predicted TC - actual TC)


Year: 2022
Actual TC in 2022
Predicted TC in 2022


Year: 2022
Percent of Frequency
Error (predicted TC - actual TC)


Year: 2023
Actual TC in 2023
Predicted TC in 2023

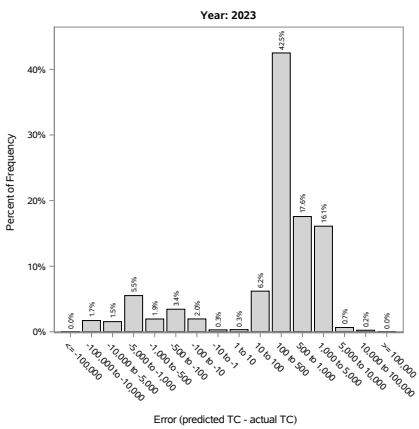

Year: 2023
Percent of Frequency
Error (predicted TC - actual TC)

Figure 3.7 Prediction errors (PE) of TC, TSC, SSC, and SC, for the whole population. Cost predictions derived from updated algorithms, excluding historical cost information from the predictors' set.

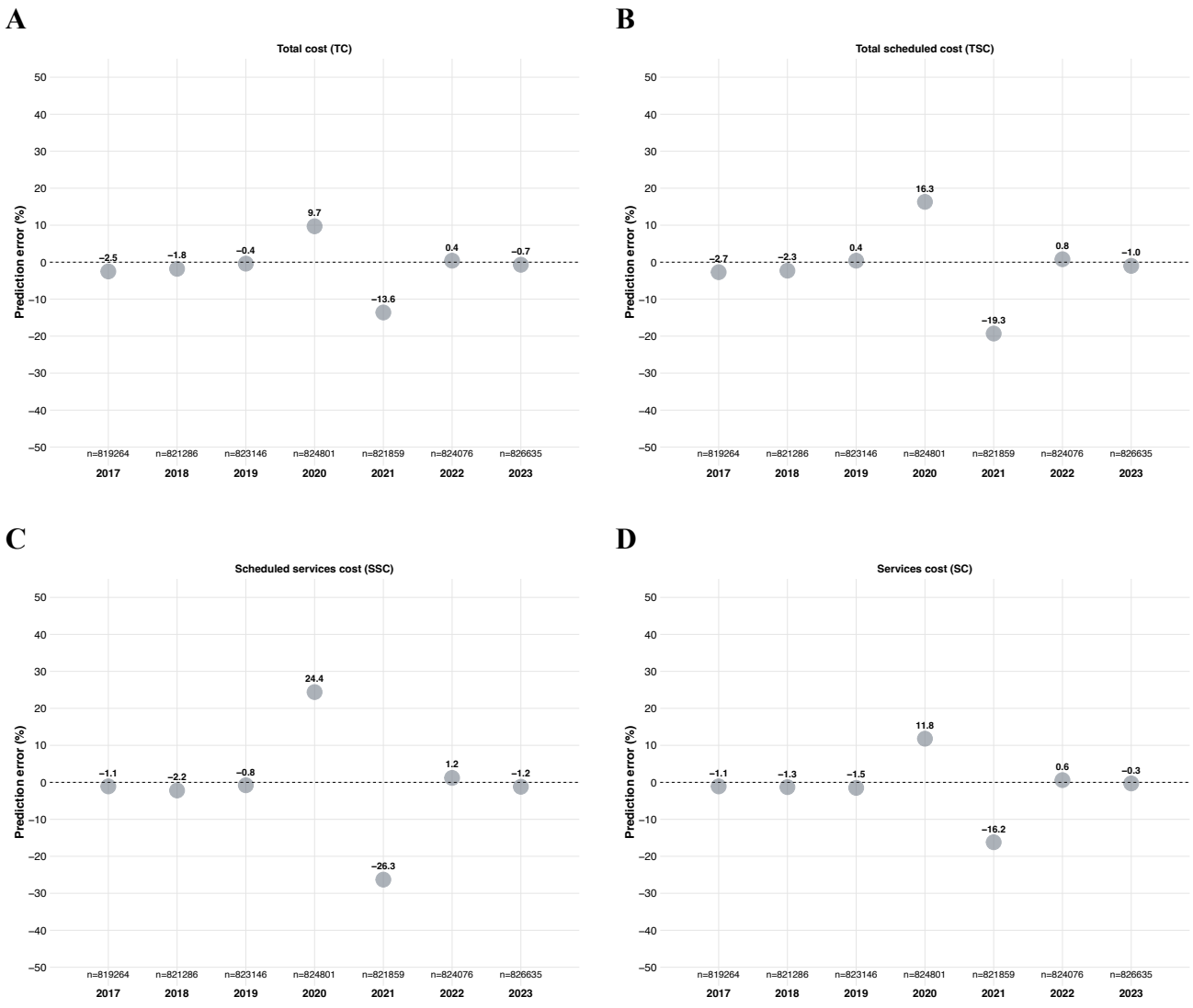


Figure 3.8 Annual predicted and actual costs (TC, TSC, SSC, and SC) for the whole population. Cost predictions derived from updated algorithms, excluding historical cost information from the predictors' set.

Figure 3.9 Prediction errors (PE) of TC, TSC, SSC, and SC, for the whole population. Cost predictions derived from updated algorithms, including historical cost information in the predictors' set.

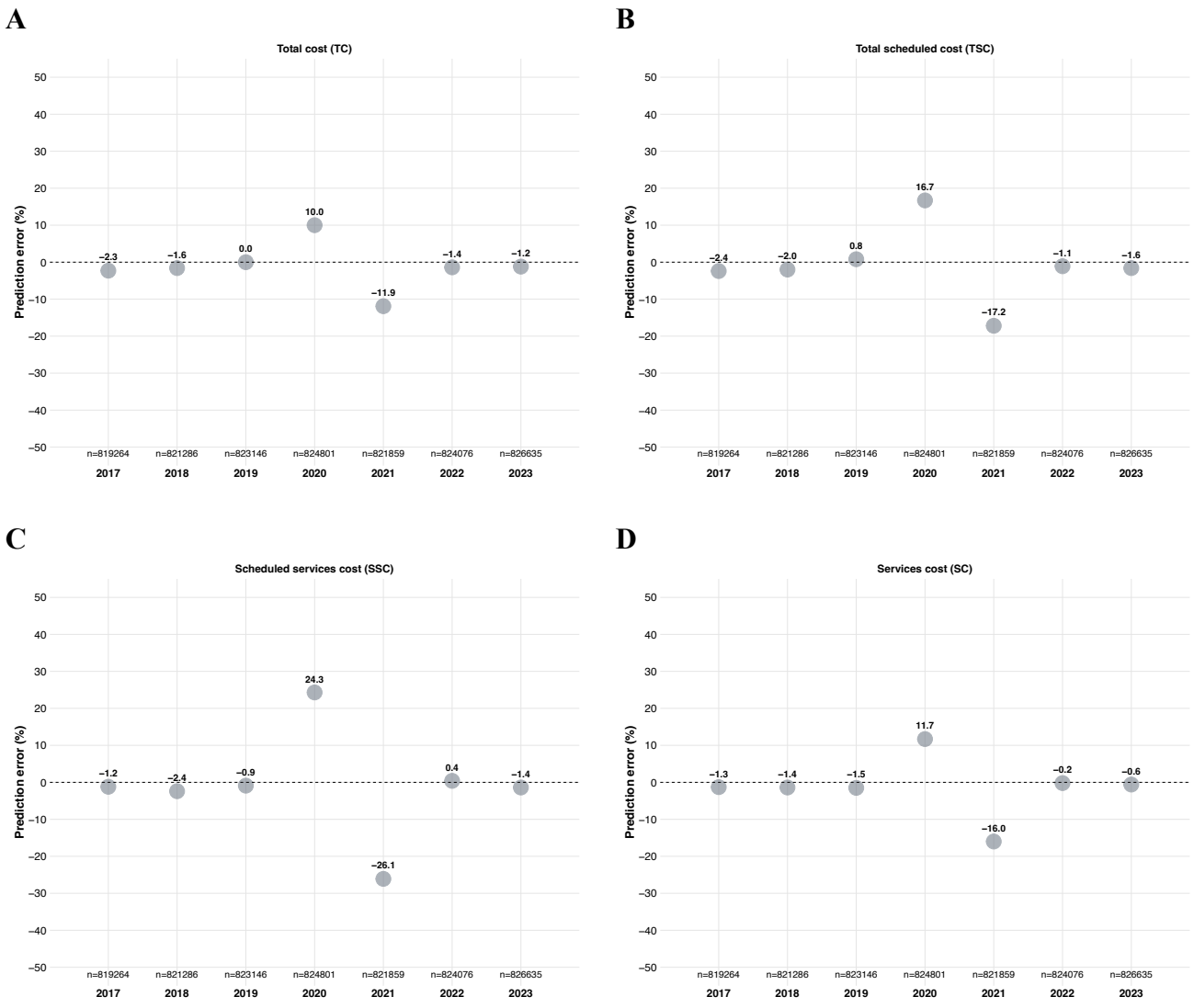


Figure 3.10 Annual predicted and actual costs (TC, TSC, SSC, and SC) for the whole population. Cost predictions derived from updated algorithms, including historical cost information in the predictors' set.

Figure 3.11 Prediction errors (PE) of TC for the CKD - Dialysis patients' groups. Cost predictions derived from updated algorithms, excluding historical cost information from the predictors' set.

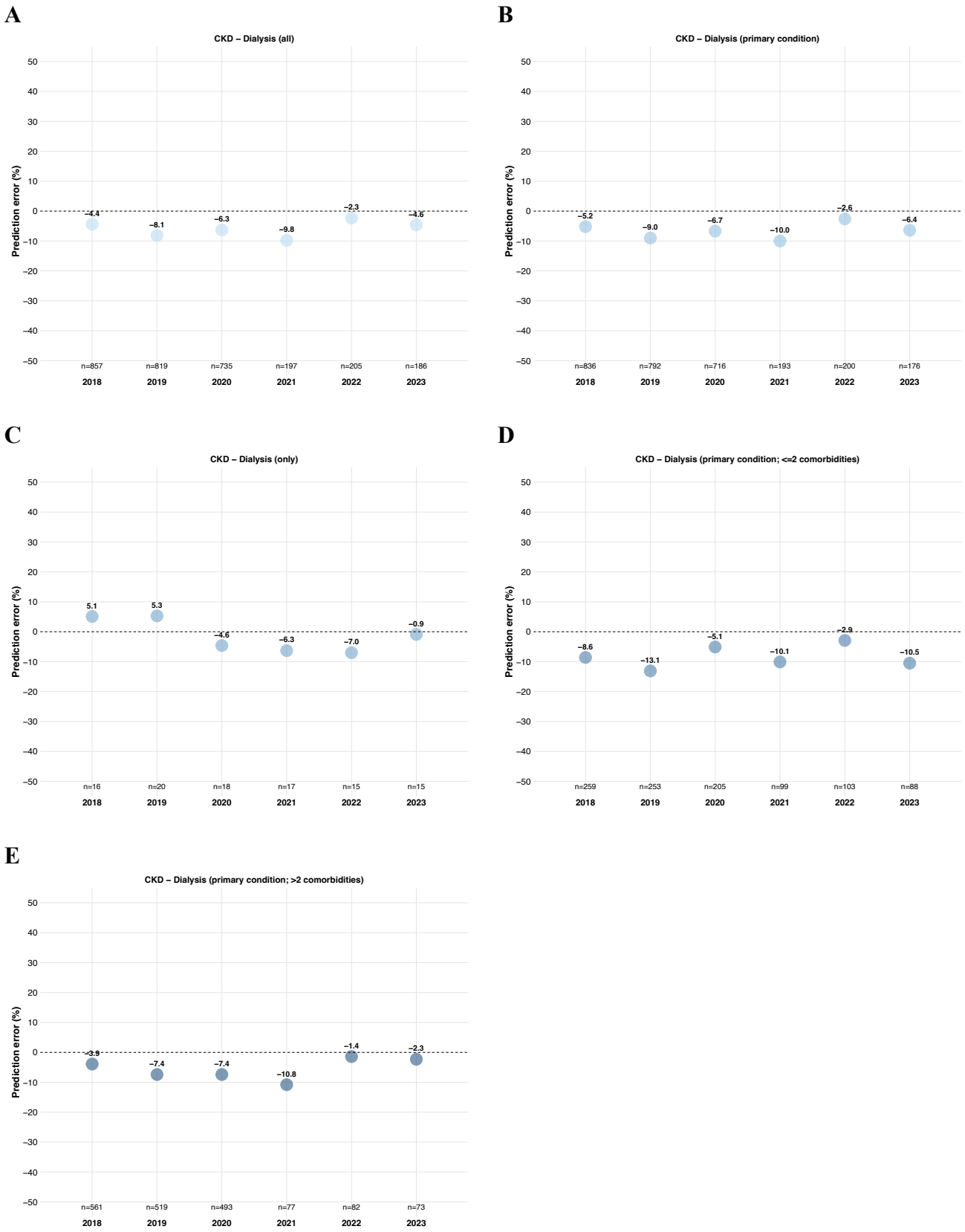


Figure 3.12 Annual predicted and actual mean TC for CKD - Dialysis patients' groups. Cost predictions derived from updated algorithms, excluding historical cost information from the predictors' set.

Figure 3.13 Prediction errors (PE) of TSC for the CKD - Dialysis patients' groups. Cost predictions derived from updated algorithms, excluding historical cost information from the predictors' set.

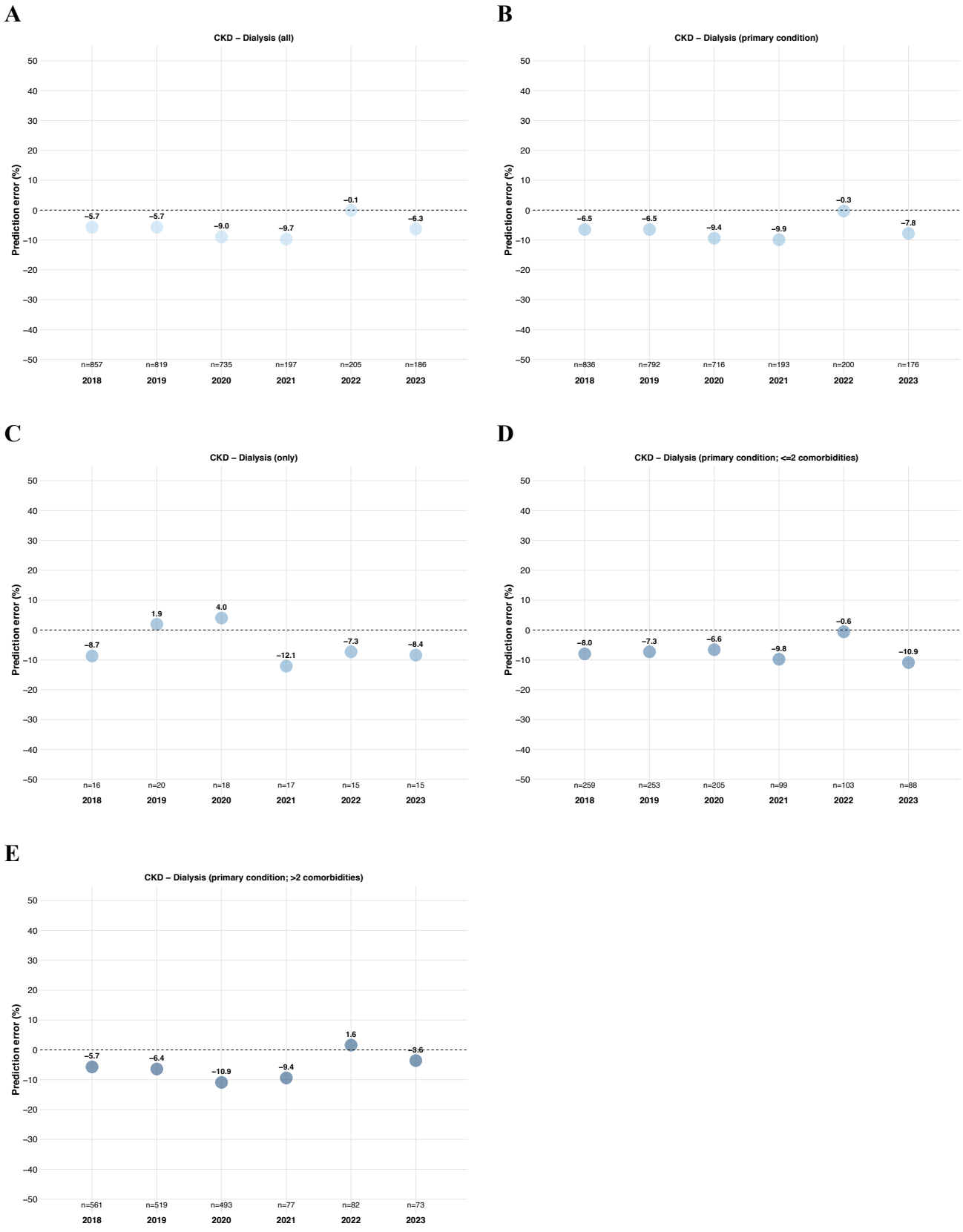


Figure 3.14 Annual predicted and actual mean TSC for CKD - Dialysis patients' groups. Cost predictions derived from updated algorithms, excluding historical cost information from the predictors' set.

Figure 3.15 Prediction errors (PE) of SSC for the CKD - Dialysis patients' groups. Cost predictions derived from updated algorithms, excluding historical cost information from the predictors' set.

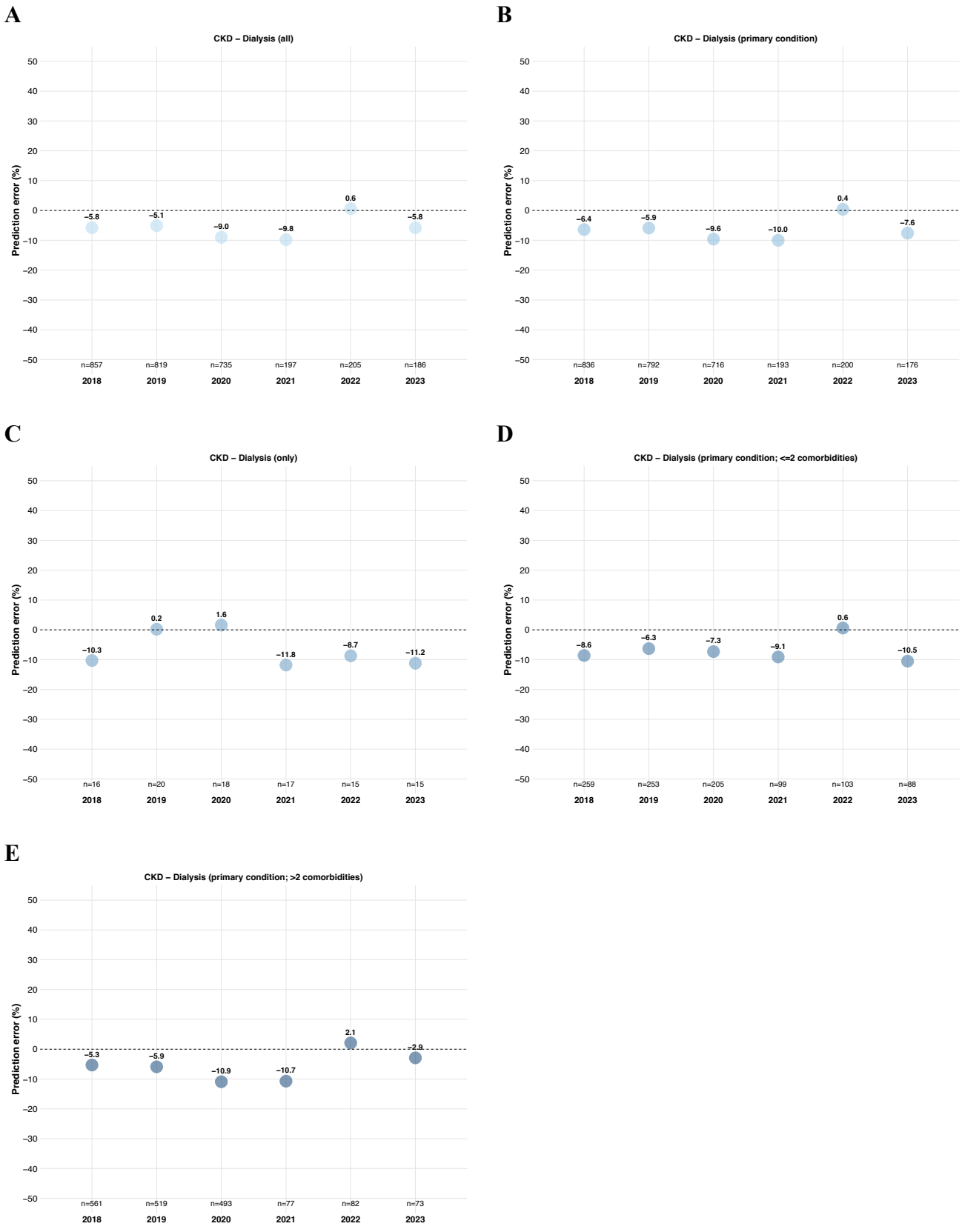


Figure 3.16 Annual predicted and actual mean SSC for CKD - Dialysis patients' groups. Cost predictions derived from updated algorithms, excluding historical cost information from the predictors' set.

Figure 3.17 Prediction errors (PE) of SC for the CKD - Dialysis patients' groups. Cost predictions derived from updated algorithms, excluding historical cost information from the predictors' set.

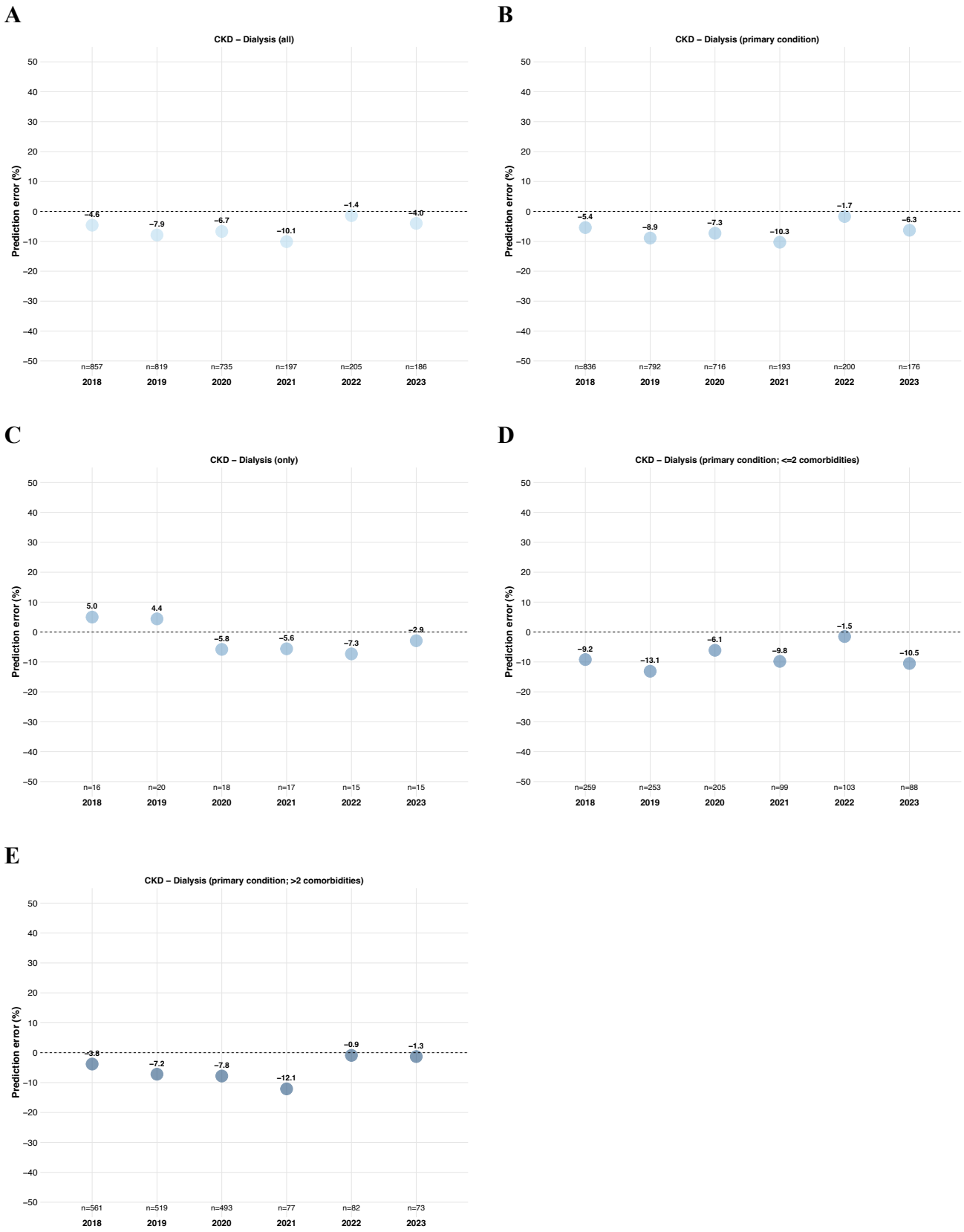


Figure 3.18 Annual predicted and actual mean SC for CKD - Dialysis patients' groups. Cost predictions derived from updated algorithms, excluding historical cost information from the predictors' set.

Figure 3.19 Prediction errors (PE) of TC for the Type 2 diabetes patients' groups. Cost predictions derived from updated algorithms, excluding historical cost information from the predictors' set.

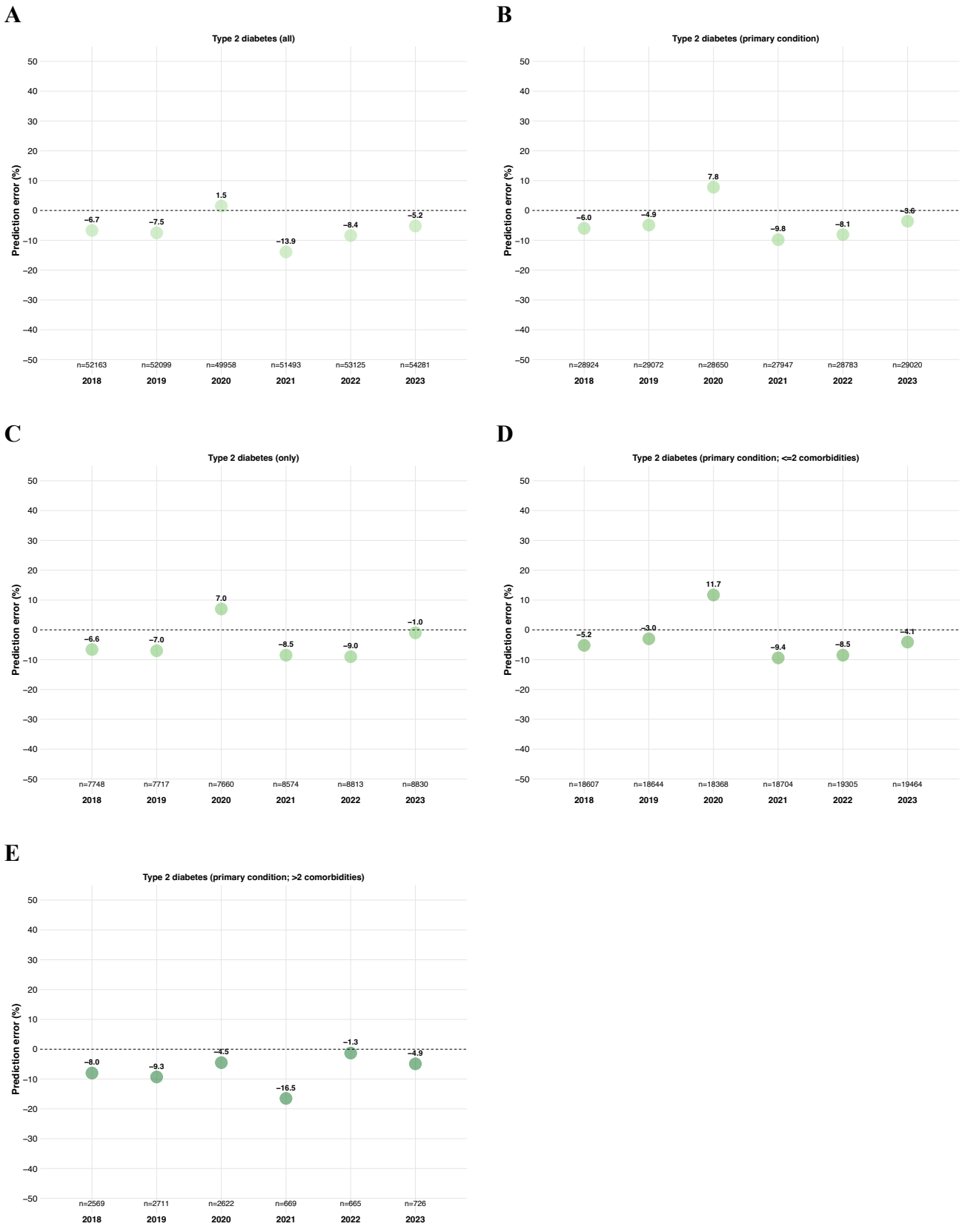


Figure 3.20 Annual predicted and actual mean TC for Type 2 diabetes patients' groups. Cost predictions derived from updated algorithms, excluding historical cost information from the predictors' set.

Figure 3.21 Prediction errors (PE) of TSC for the Type 2 diabetes patients' groups. Cost predictions derived from updated algorithms, excluding historical cost information from the predictors' set.

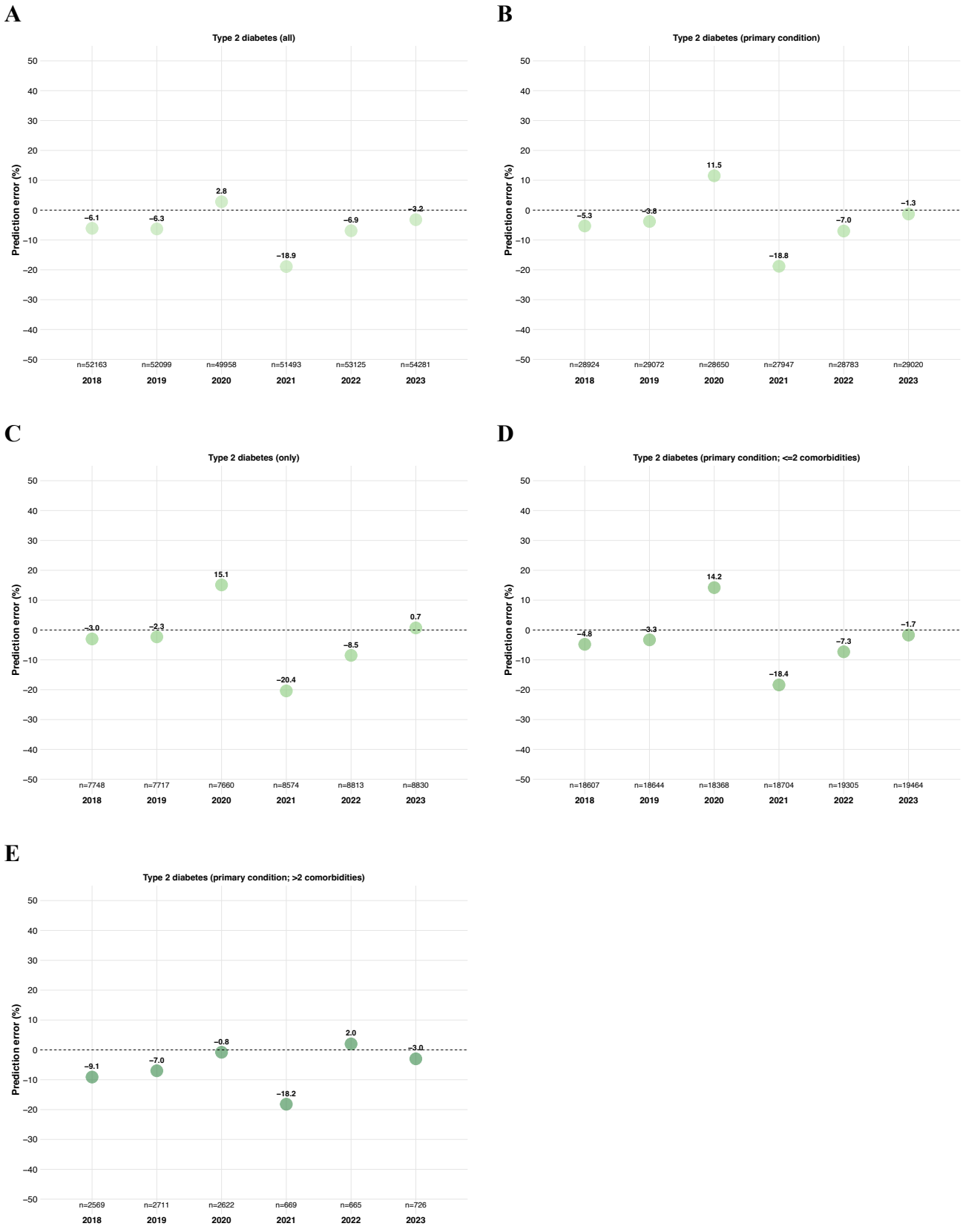


Figure 3.22 Annual predicted and actual mean TSC for Type 2 diabetes patients' groups. Cost predictions derived from updated algorithms, excluding historical cost information from the predictors' set.

Figure 3.23 Prediction errors (PE) of SSC for the Type 2 diabetes patients' groups. Cost predictions derived from updated algorithms, excluding historical cost information from the predictors' set.

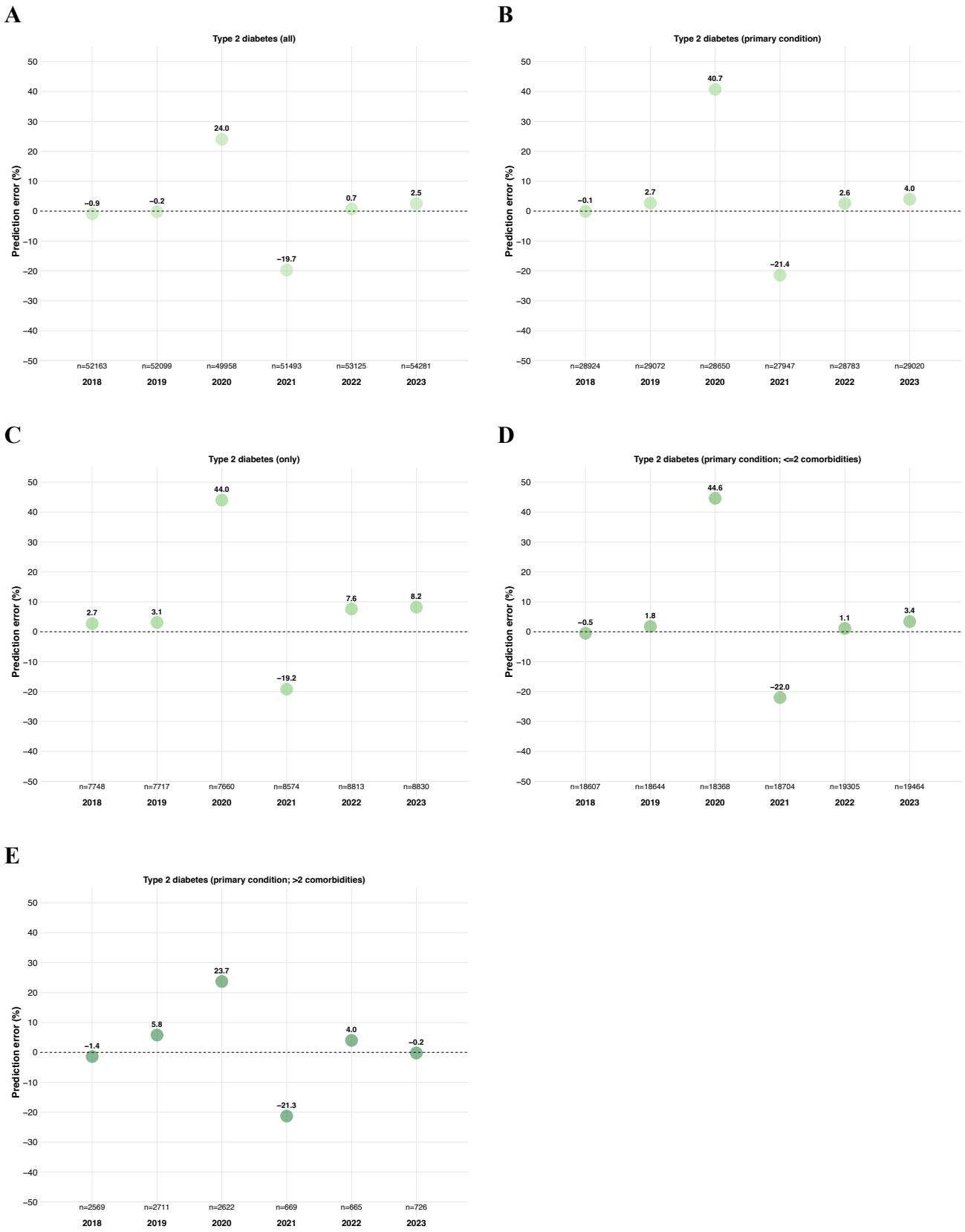


Figure 3.24 Annual predicted and actual mean SSC for Type 2 diabetes patients' groups. Cost predictions derived from updated algorithms, excluding historical cost information from the predictors' set.

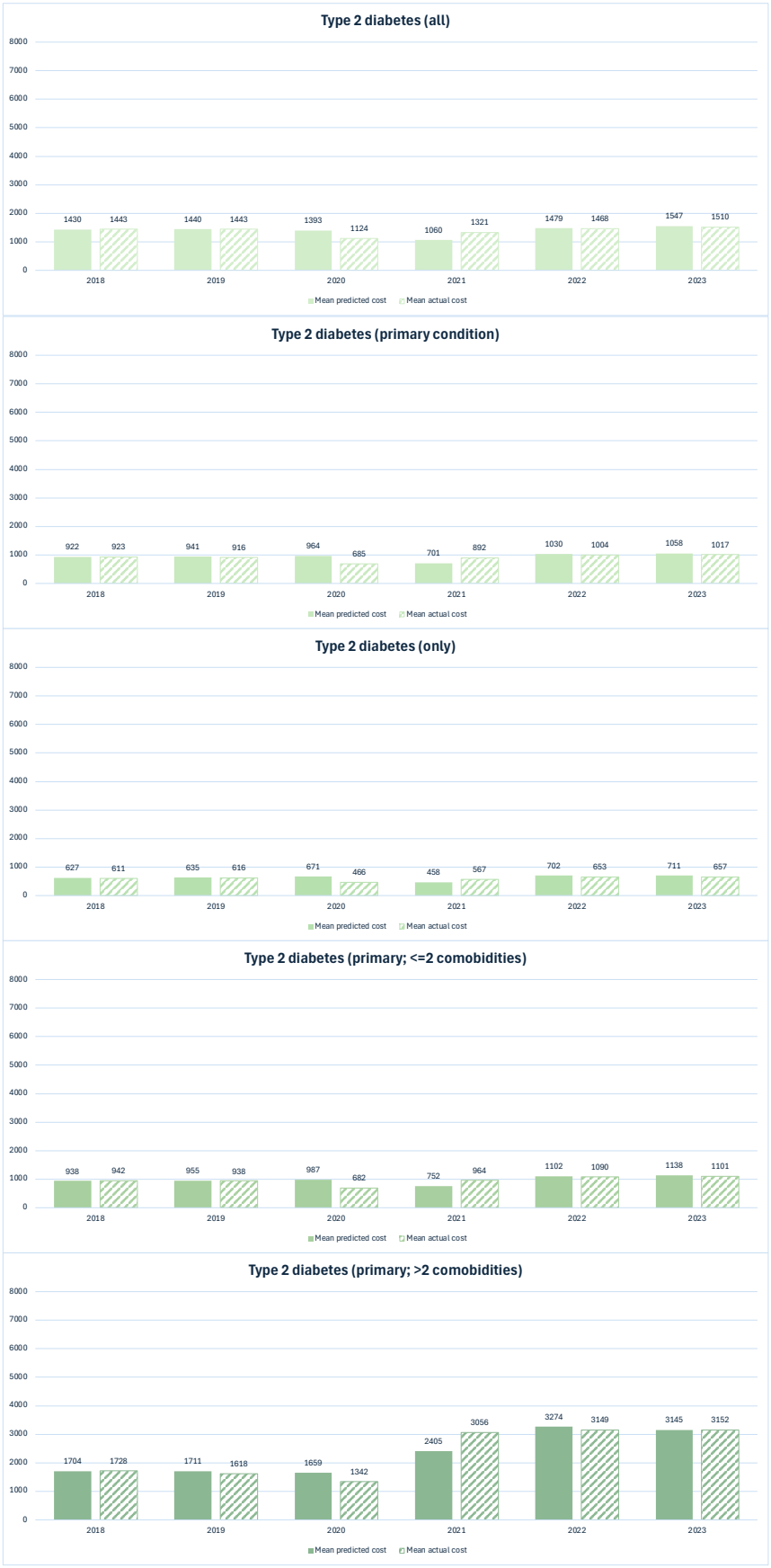


Figure 3.25 Prediction errors (PE) of SC for the Type 2 diabetes patients' groups. Cost predictions derived from updated algorithms, excluding historical cost information from the predictors' set.

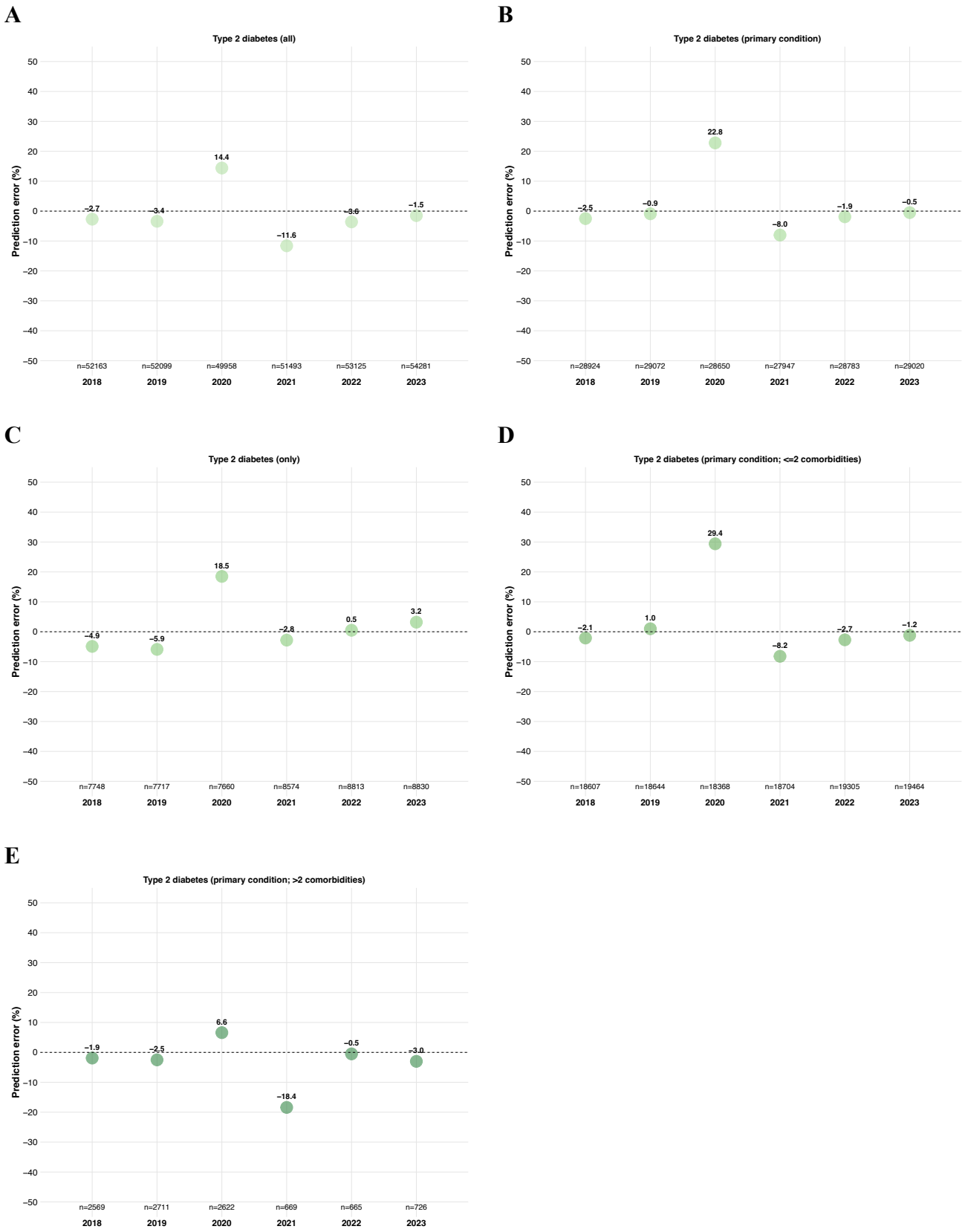


Figure 3.26 Annual predicted and actual mean SC for Type 2 diabetes patients' groups. Cost predictions derived from updated algorithms, excluding historical cost information from the predictors' set.

Figure 3.27 Prediction errors (PE) of TC for the Heart failure patients' groups. Cost predictions derived from updated algorithms, excluding historical cost information from the predictors' set.

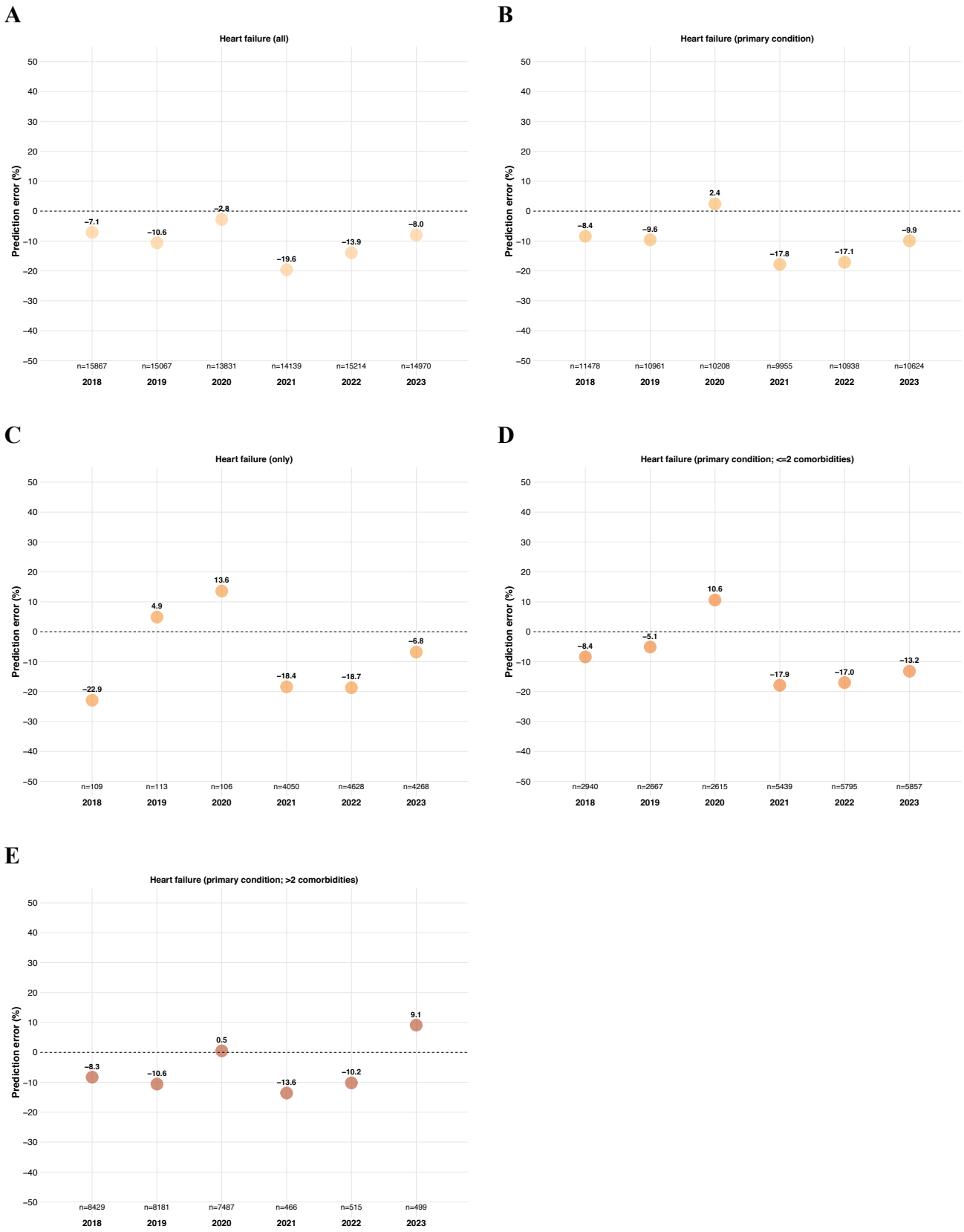


Figure 3.28 Annual predicted and actual mean TC for Heart failure patients' groups. Cost predictions derived from updated algorithms, excluding historical cost information from the predictors' set.

Figure 3.29 Prediction errors (PE) of TSC for the Heart failure patients' groups. Cost predictions derived from updated algorithms, excluding historical cost information from the predictors' set.

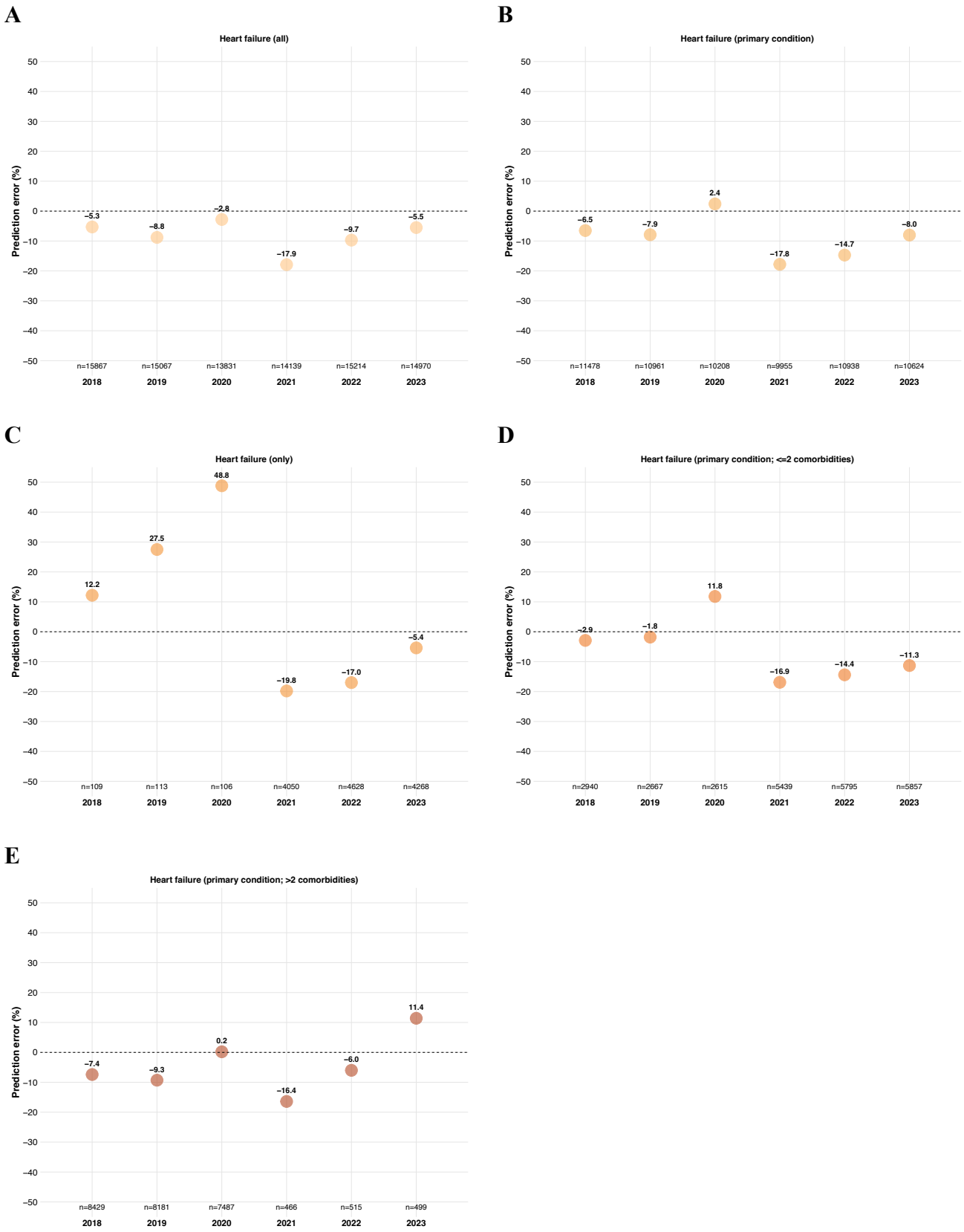


Figure 3.30 Annual predicted and actual mean TSC for Heart failure patients' groups. Cost predictions derived from updated algorithms, excluding historical cost information from the predictors' set.

Figure 3.31 Prediction errors (PE) of SSC for the Heart failure patients' groups. Cost predictions derived from updated algorithms, excluding historical cost information from the predictors' set.

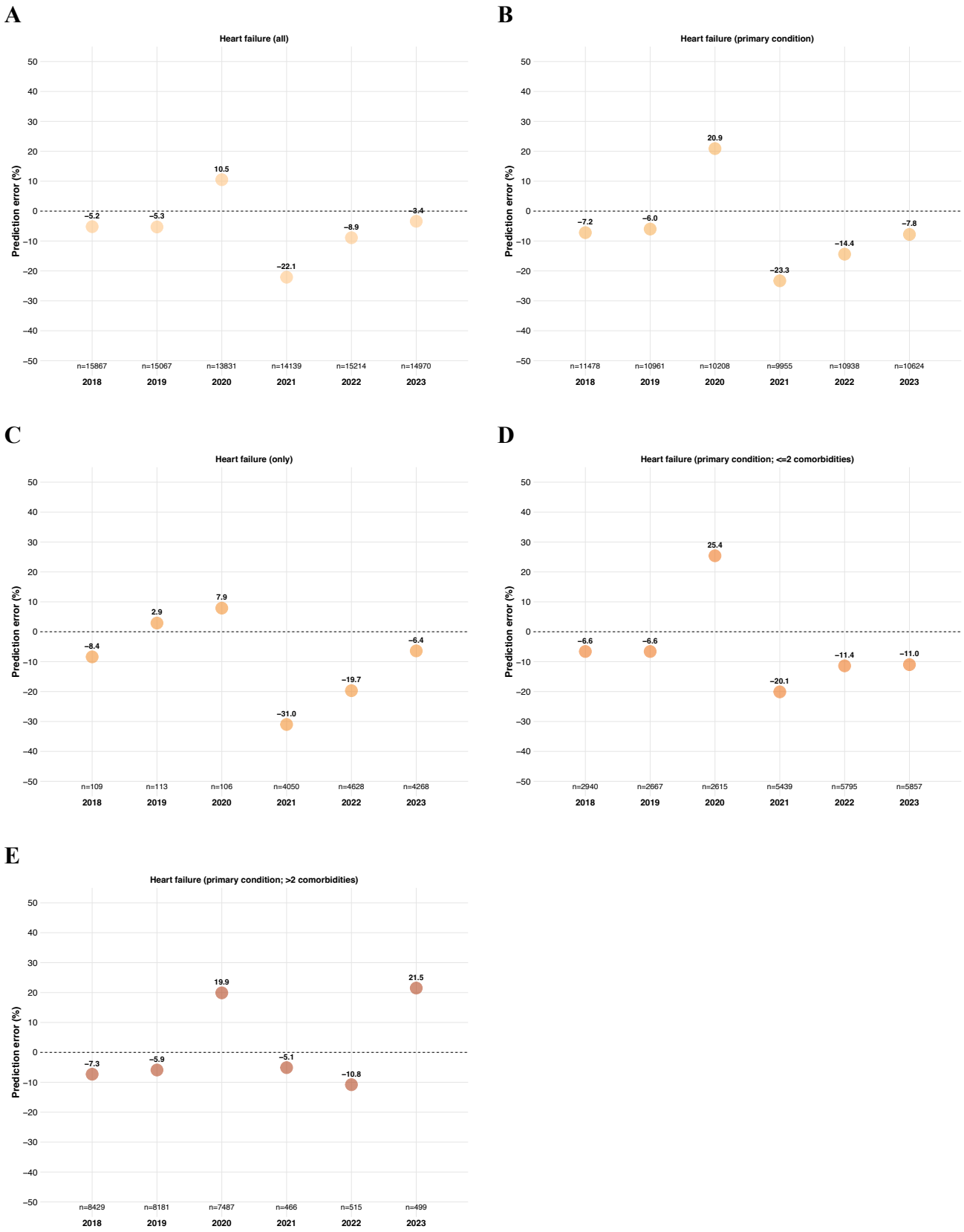


Figure 3.32 Annual predicted and actual mean SSC for Heart failure patients' groups. Cost predictions derived from updated algorithms, excluding historical cost information from the predictors' set.

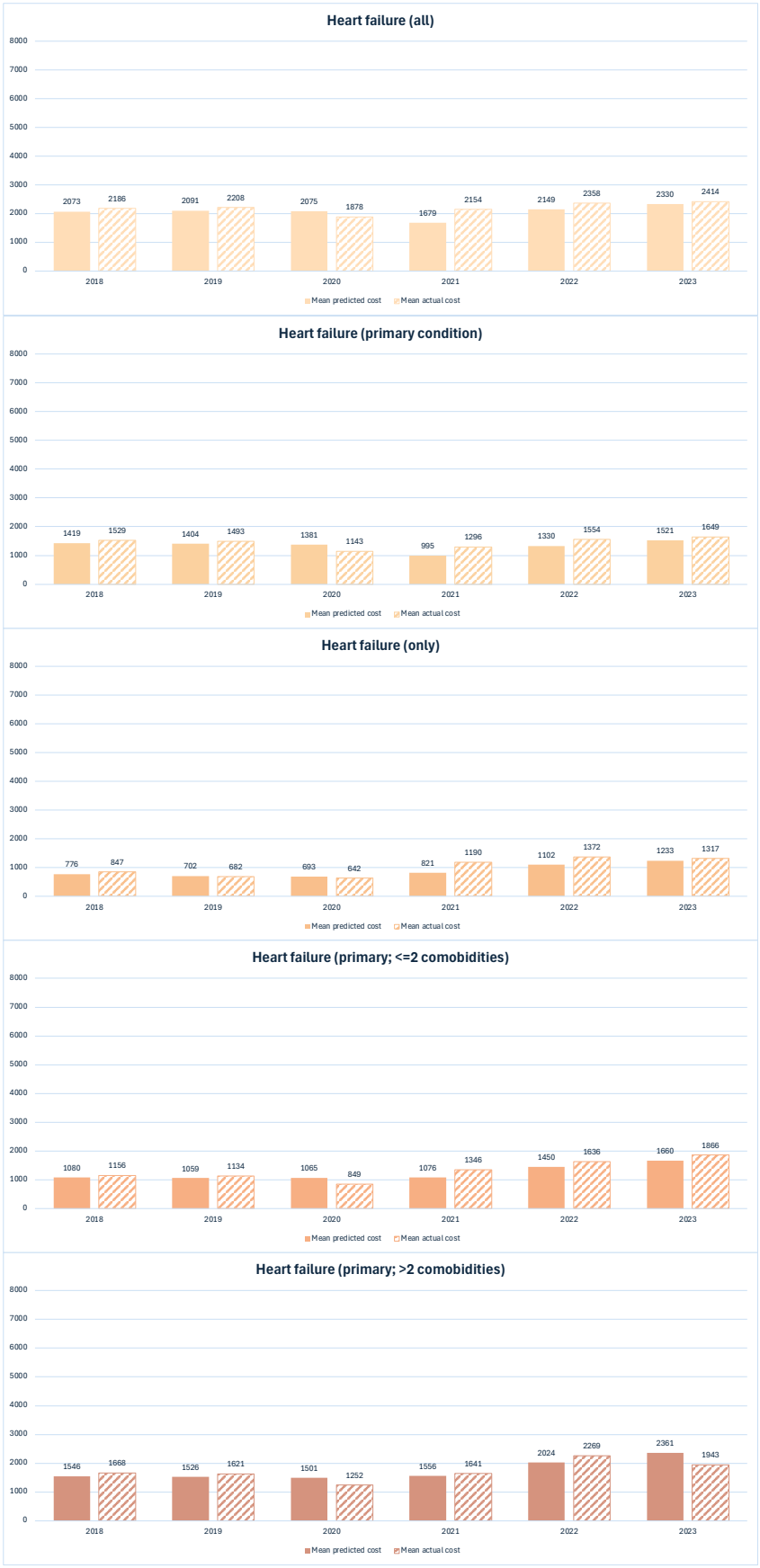


Figure 3.33 Prediction errors (PE) of SC for the Heart failure patients' groups. Cost predictions derived from updated algorithms, excluding historical cost information from the predictors' set.

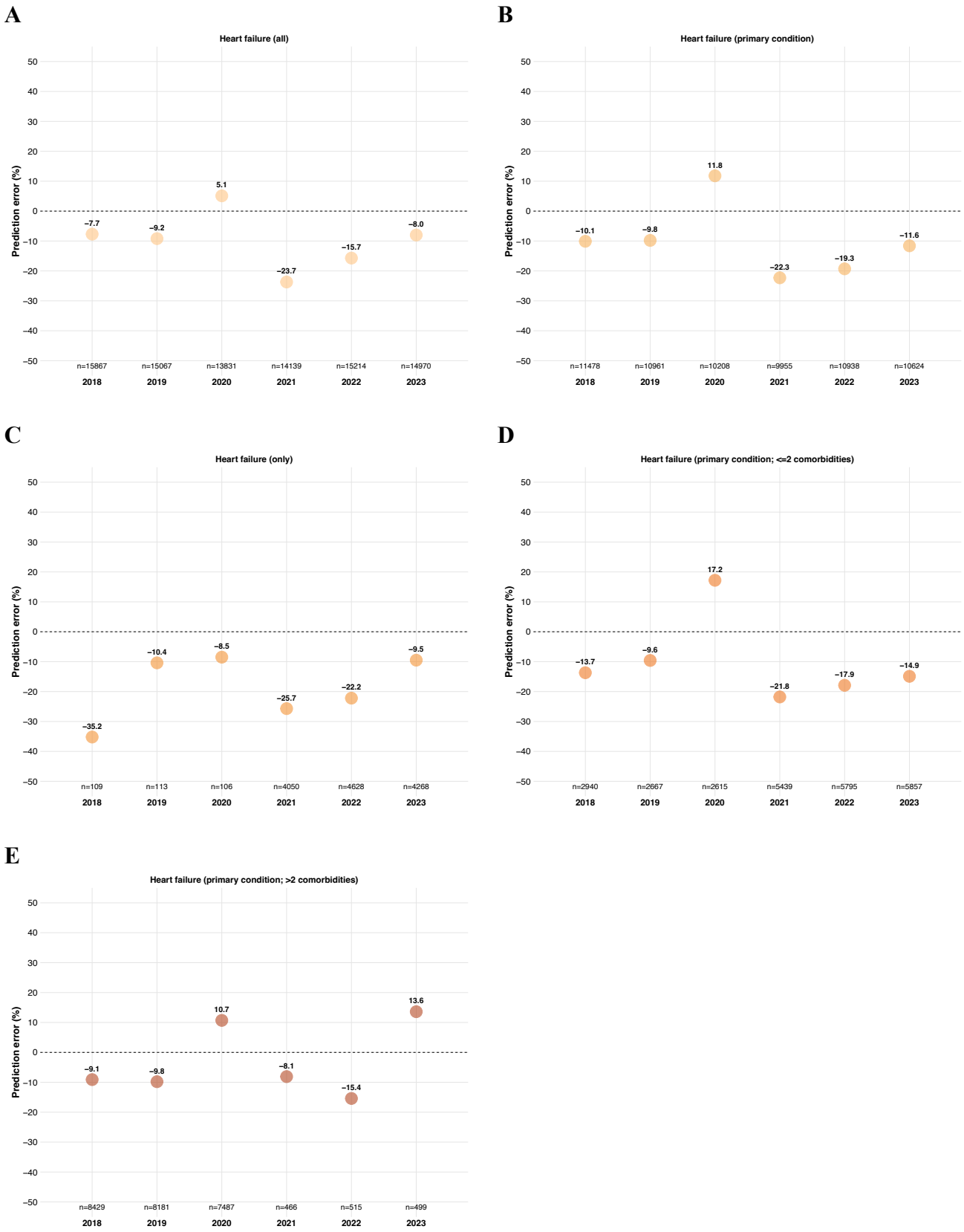


Figure 3.34 Annual predicted and actual mean SC for Heart failure patients' groups. Cost predictions derived from updated algorithms, excluding historical cost information from the predictors' set.

Figure 3.35 Prediction errors (PE) of TC for the Parkinson's disease/Parkinsonian syndromes patients' groups. Cost predictions derived from updated algorithms, excluding historical cost information from the predictors' set.

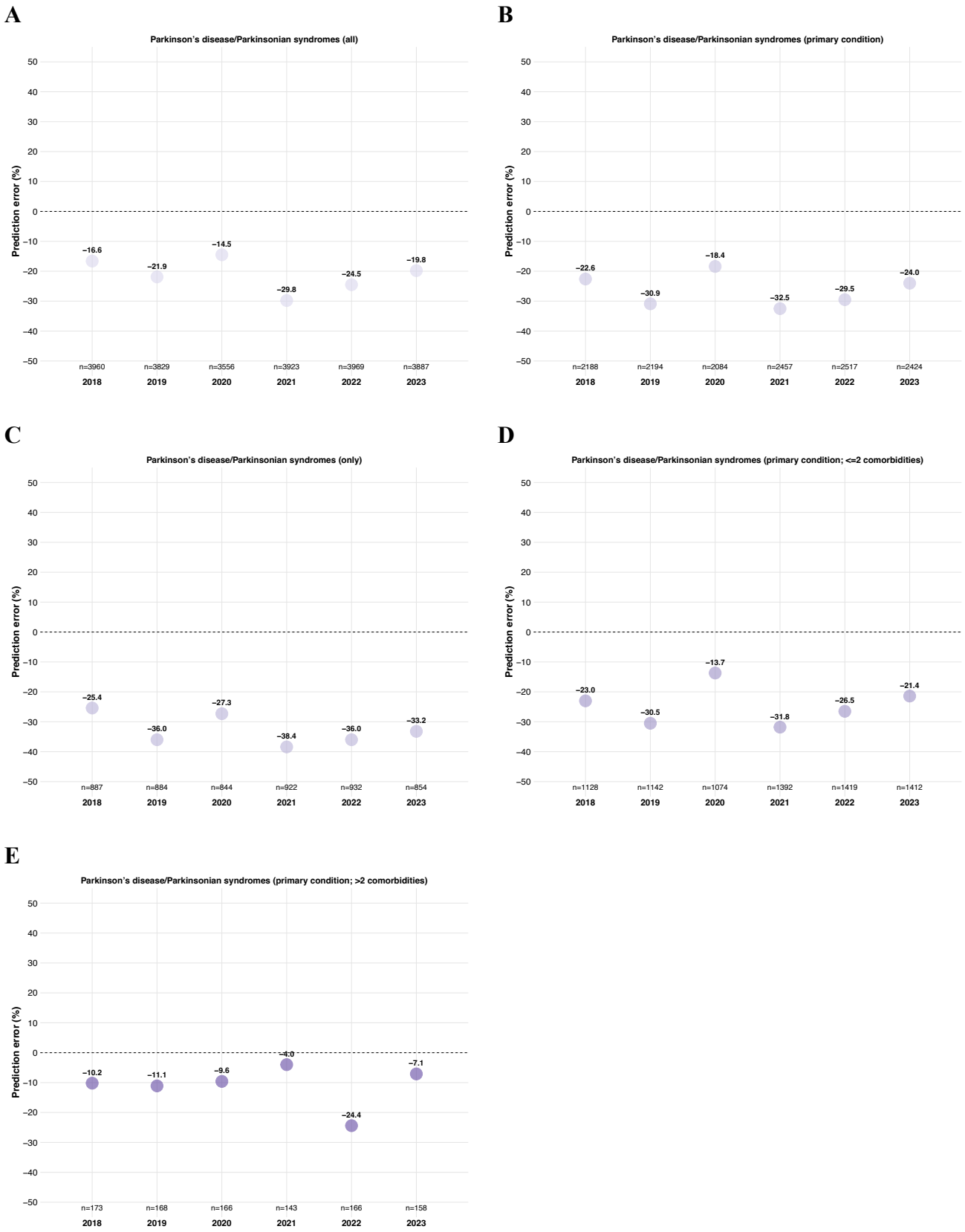


Figure 3.36 Annual predicted and actual mean TC for Parkinson's disease/Parkinsonian syndromes patients' groups. Cost predictions derived from updated algorithms, excluding historical cost information from the predictors' set.

Figure 3.37 Prediction errors (PE) of TSC for the Parkinson's disease/Parkinsonian syndromes patients' groups. Cost predictions derived from updated algorithms, excluding historical cost information from the predictors' set.

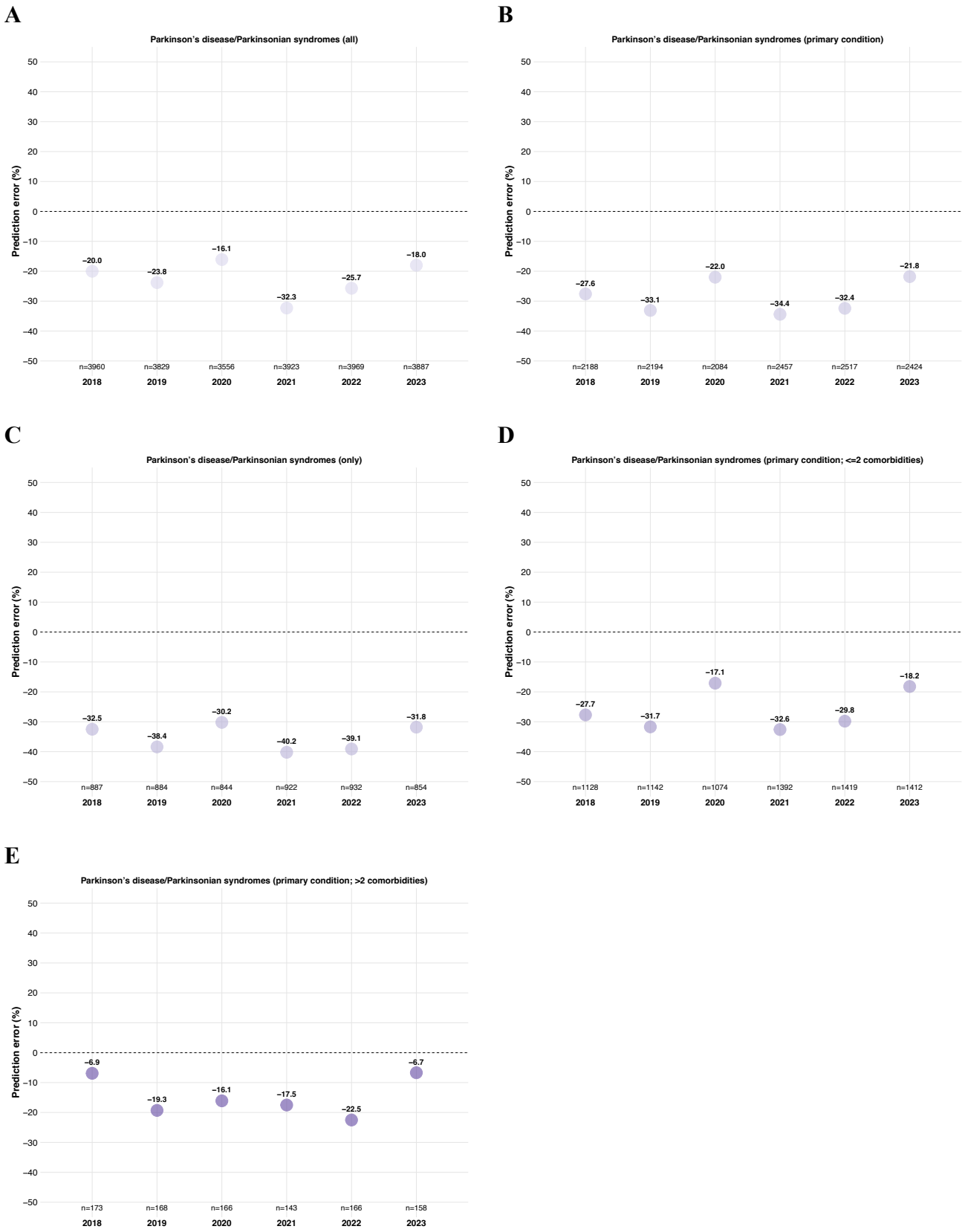


Figure 3.38 Annual predicted and actual mean TSC for Parkinson's disease/Parkinsonian syndromes patients' groups. Cost predictions derived from updated algorithms, excluding historical cost information from the predictors' set.

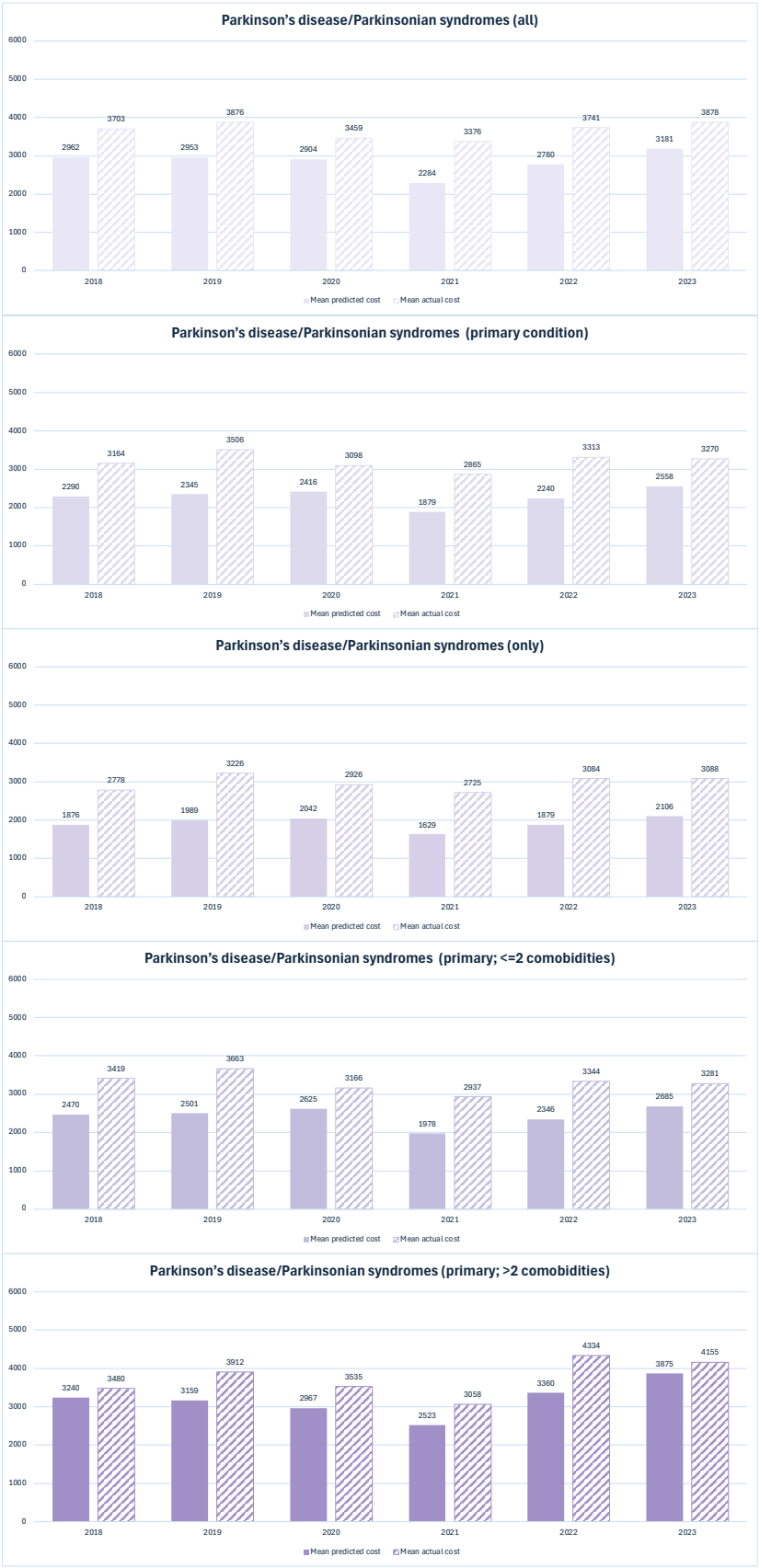


Figure 3.39 Prediction errors (PE) of SSC for the Parkinson's disease/Parkinsonian syndromes patients' groups. Cost predictions derived from updated algorithms, excluding historical cost information from the predictors' set.

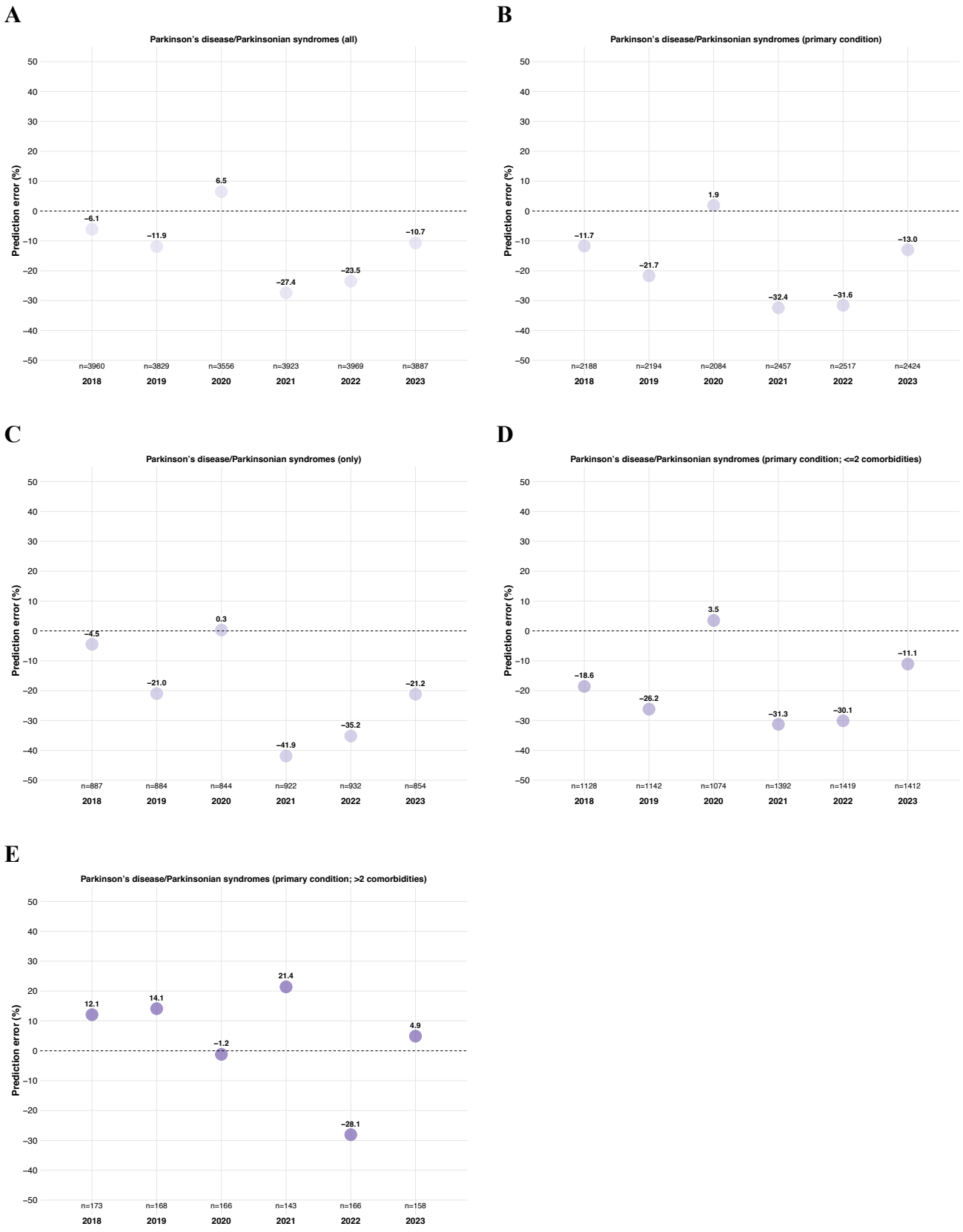


Figure 3.40 Annual predicted and actual mean SSC for Parkinson's disease/Parkinsonian syndromes patients' groups. Cost predictions derived from updated algorithms, excluding historical cost information from the predictors' set.

Figure 3.41 Prediction errors (PE) of SC for the Parkinson's disease/Parkinsonian syndromes patients' groups. Cost predictions derived from updated algorithms, excluding historical cost information from the predictors' set.

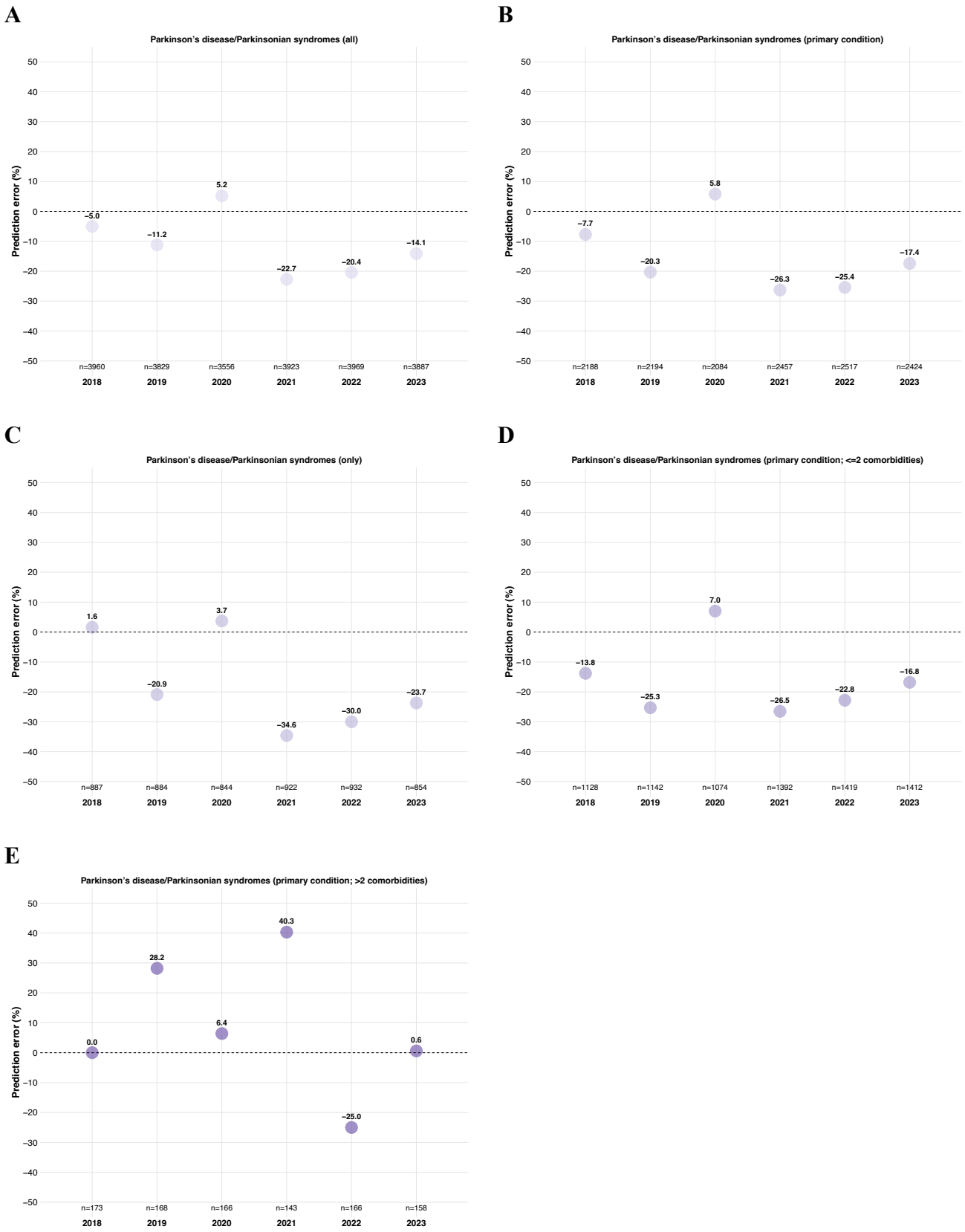


Figure 3.42 Annual predicted and actual mean SC for Parkinson's disease/Parkinsonian syndromes patients' groups. Cost predictions derived from updated algorithms, excluding historical cost information from the predictors' set.

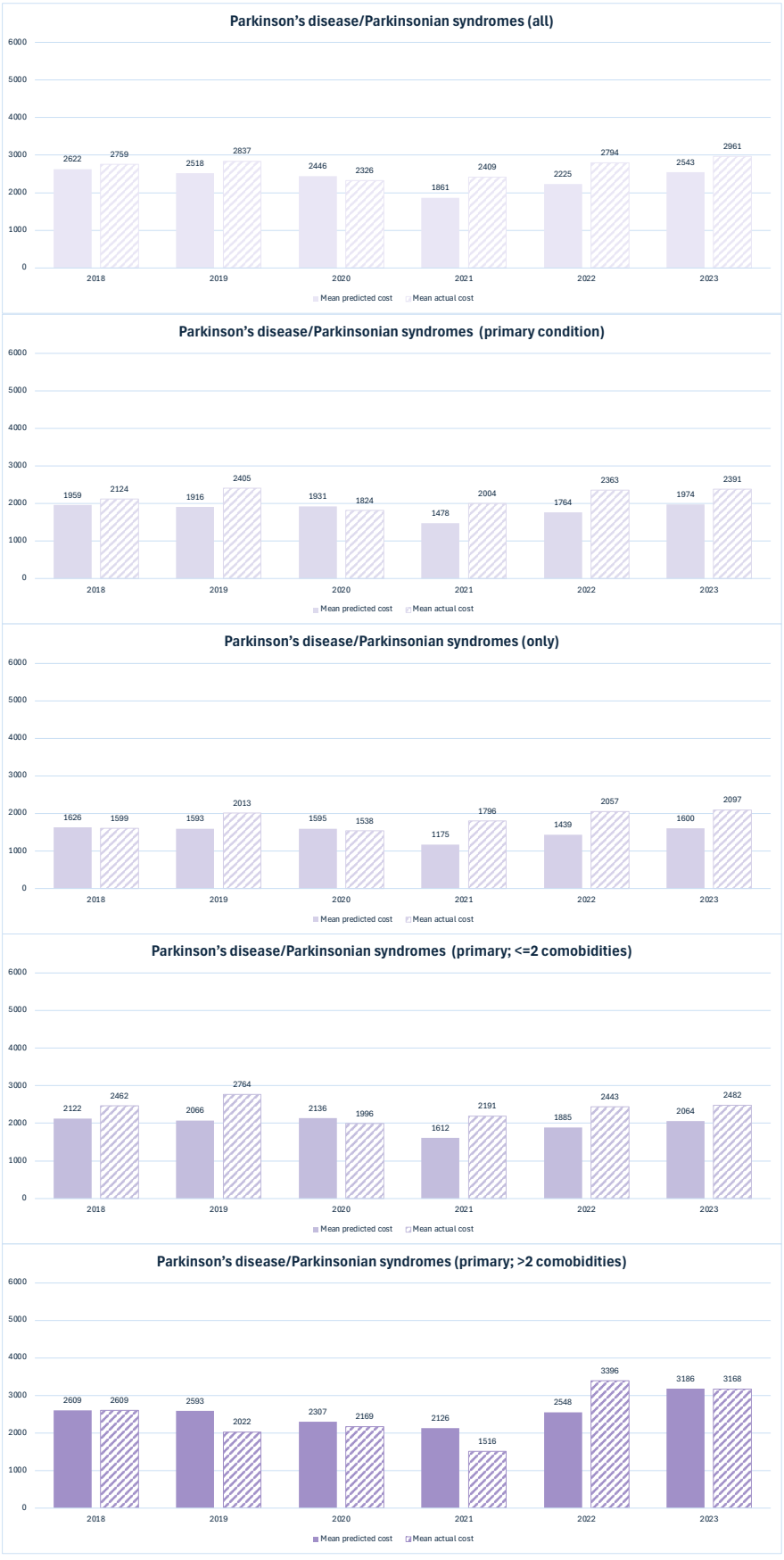


Figure 3.43 Prediction errors (PE) of TC for the Active neoplasia patients' groups. Cost predictions derived from updated algorithms, excluding historical cost information from the predictors' set.

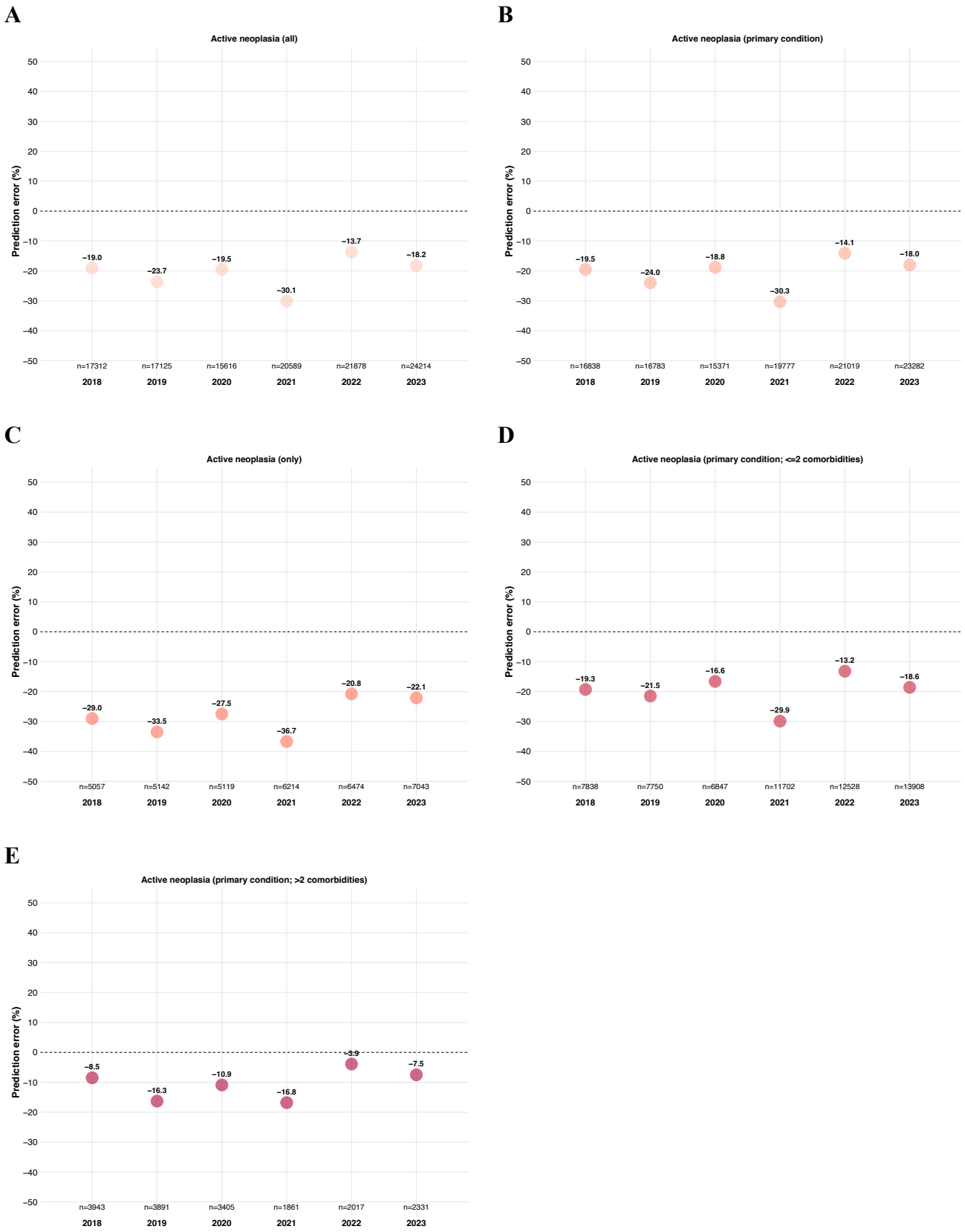


Figure 3.44 Annual predicted and actual mean TC for Active neoplasia patients' groups. Cost predictions derived from updated algorithms, excluding historical cost information from the predictors' set.

Figure 3.45 Prediction errors (PE) of TSC for the Active neoplasia patients' groups. Cost predictions derived from updated algorithms, excluding historical cost information from the predictors' set.

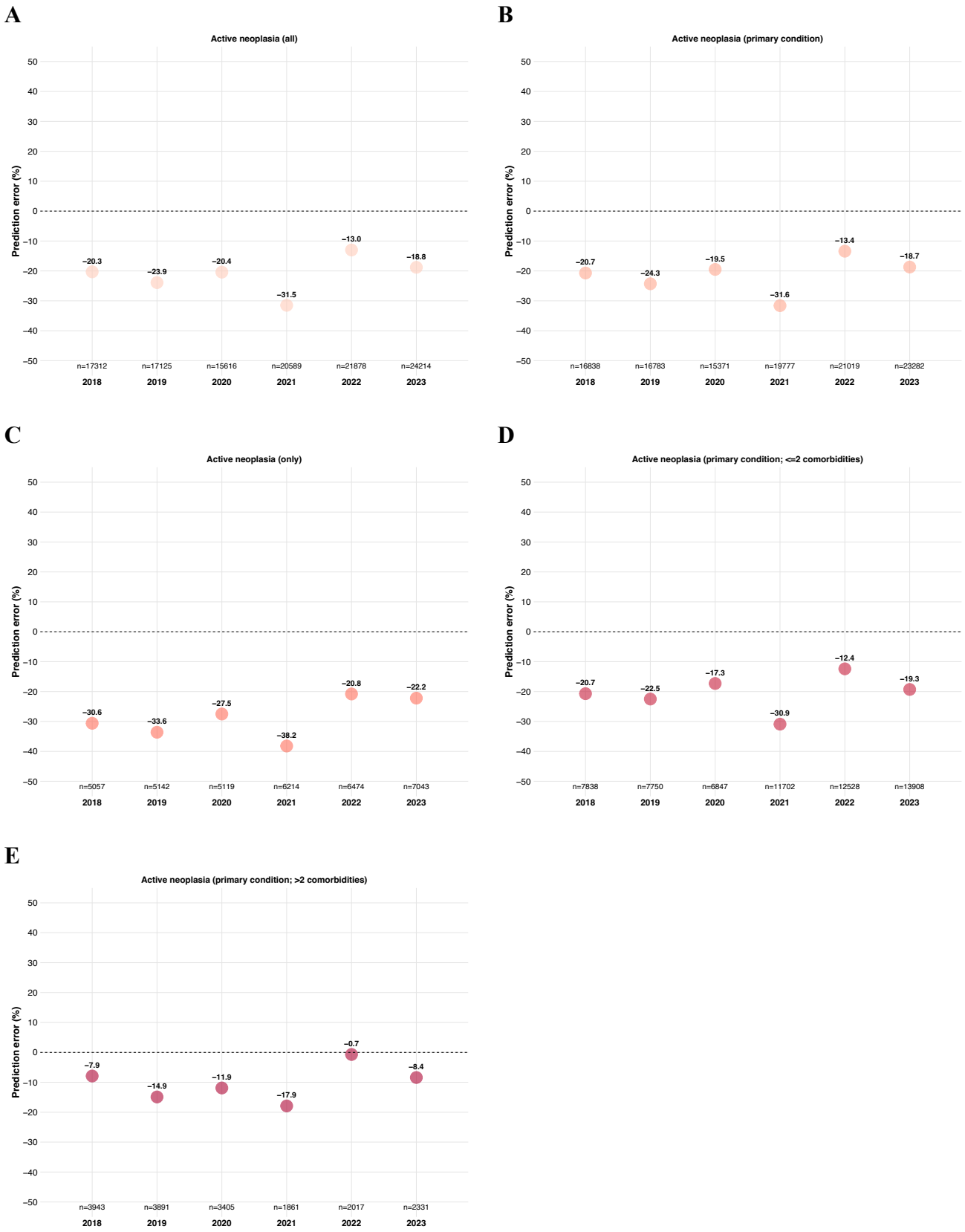


Figure 3.46 Annual predicted and actual mean TSC for Active neoplasia patients' groups. Cost predictions derived from updated algorithms, excluding historical cost information from the predictors' set.

Figure 3.47 Prediction errors (PE) of SSC for the Active neoplasia patients' groups. Cost predictions derived from updated algorithms, excluding historical cost information from the predictors' set.

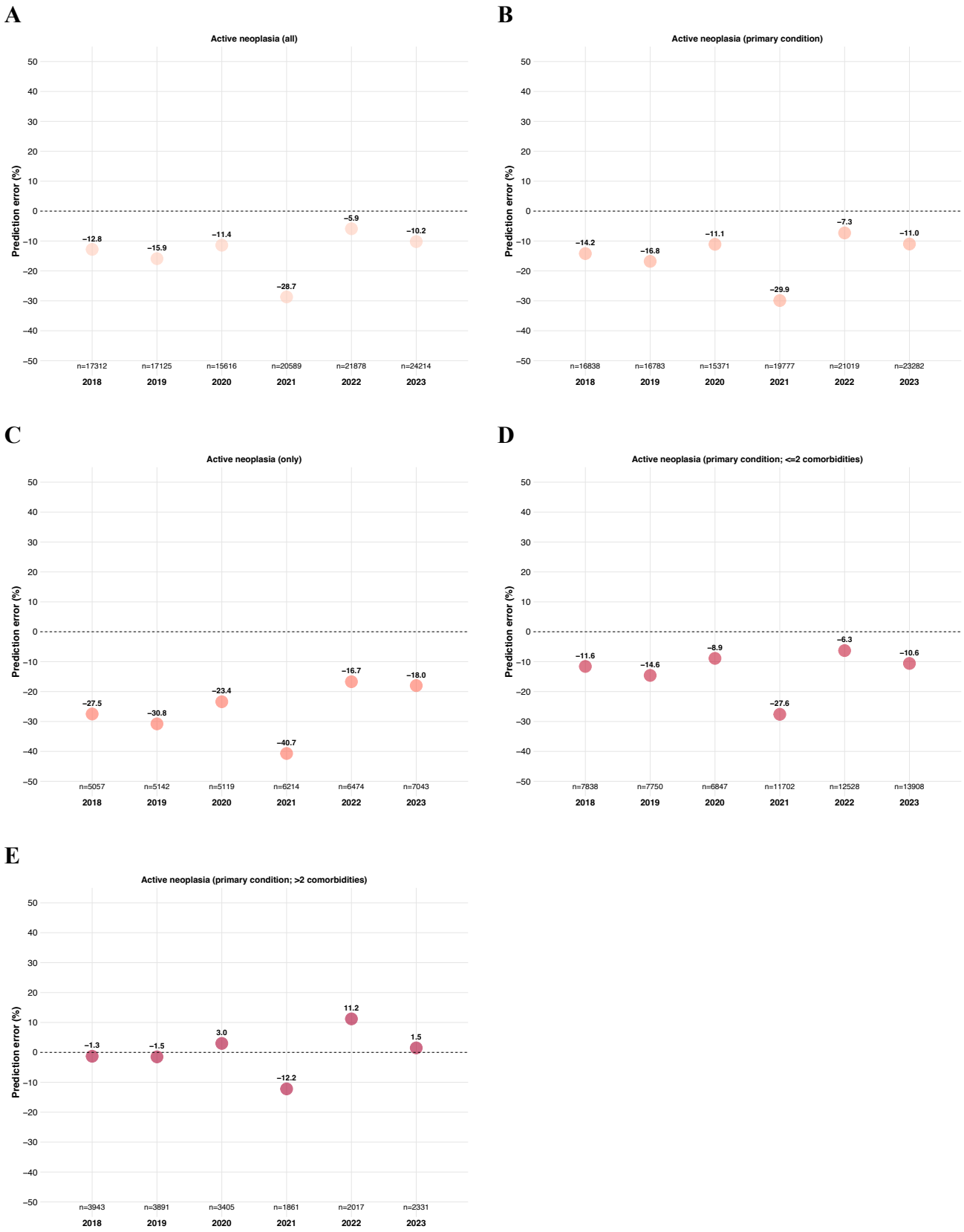


Figure 3.48 Annual predicted and actual mean SSC for Active neoplasia patients' groups. Cost predictions derived from updated algorithms, excluding historical cost information from the predictors' set.

Figure 3.49 Prediction errors (PE) of SC for the Active neoplasia patients' groups. Cost predictions derived from updated algorithms, excluding historical cost information from the predictors' set.

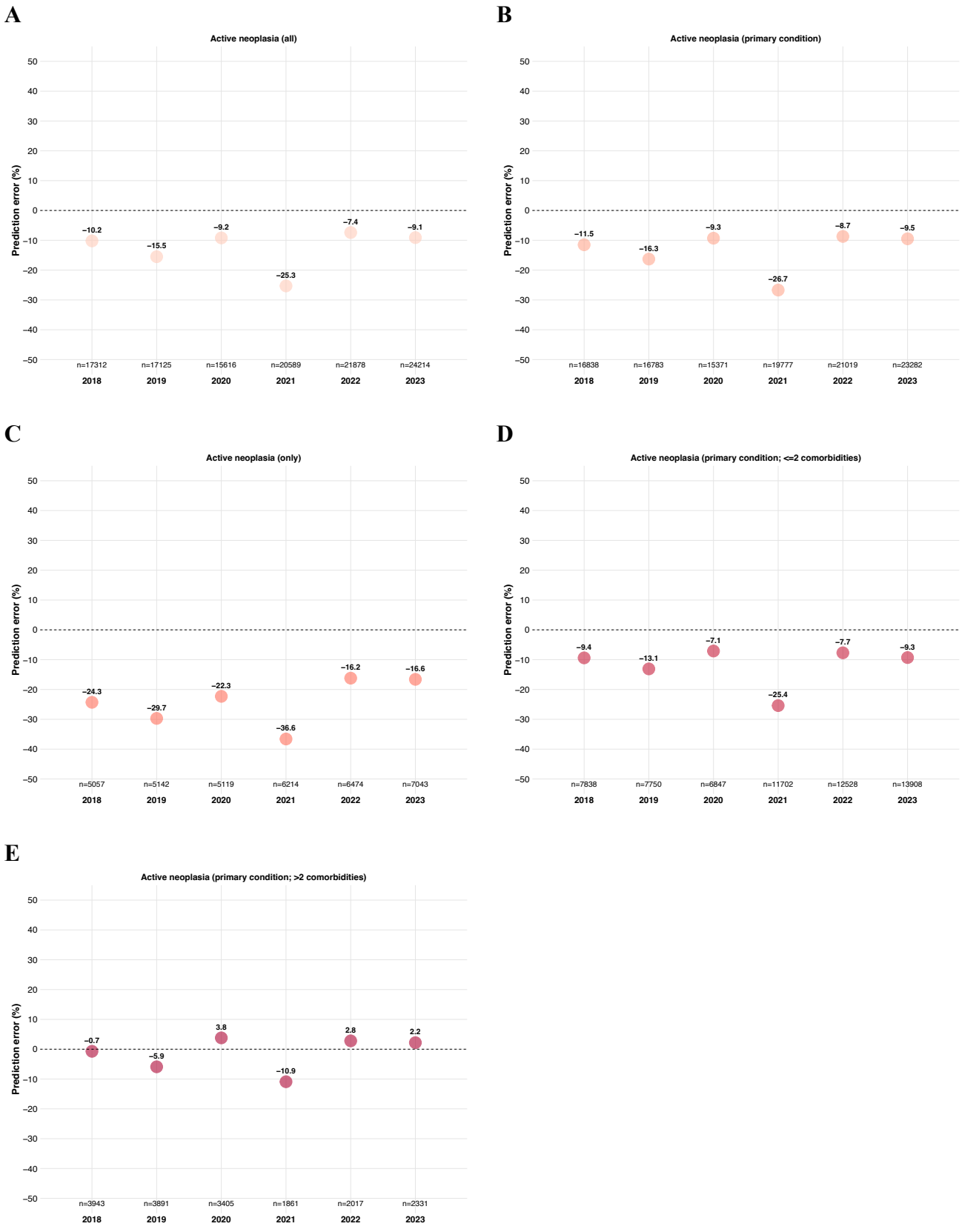


Figure 3.50 Annual predicted and actual mean SC for Active neoplasia patients' groups. Cost predictions derived from updated algorithms, excluding historical cost information from the predictors' set.

4. DISCUSSION

Predicting future healthcare expenditures is crucial for effective planning, as population ageing and the increasing prevalence of chronic conditions pose a growing challenge to the financial sustainability of healthcare systems. This necessitates the development of tools capable of estimating, with reasonable accuracy, the costs that may be incurred in the future.

The aim of the present study was to implement a predictive tool for estimating future healthcare costs for the general population and for high-impact patient groups, based on individuals' historical patterns of access to, and utilisation of, services provided by the Italian NHS. To this end, ML techniques were applied to administrative healthcare databases pertaining to all adults resident in the area of the Bergamo's Health Protection Agency from 2011 to 2023. For each subject and for each year, as well as for cumulative four-year periods, the history of access to NHS services was reconstructed (e.g., number and type of hospital admissions, outpatient visits, and laboratory tests, and pharmaceutical consumption). Random forest regression algorithms were then trained to estimate next-year healthcare costs for the entire population and for selected population segments of particular interest - namely patients on dialysis, patients with type 2 diabetes, those with heart failure, those affected by Parkinson's disease or parkinsonism, and patients with active neoplasia.

Compared with classical approaches such as linear regression, the use of predictive ML algorithms makes it possible to capture complex patterns in the data, thereby enhancing predictive performance. In recent years, many areas of medicine have benefited from the application of ML techniques, including diagnosis and outcome prediction; this also encompasses the possibility of identifying individuals at high risk for medical emergencies such as relapse. (70) For instance, ML algorithms have recently been successfully employed to predict age-specific primary ovarian insufficiency risk among long-term survivors of childhood cancer (71) and to establish which groups of patients with gastrointestinal stromal tumour should receive adjuvant therapy (72). Over the years, the application of ML algorithms to the prediction of healthcare costs has also proliferated, driven in particular by the growth in healthcare expenditures in the United States and worldwide. Healthcare cost prediction is of interest not only to health systems - for the optimal allocation of resources and cost containment - but also to insurance companies, for whom accurate forecasts of likely expenses can support general business planning, and to patients themselves, as knowing their expected expenditures for the coming year could enable more informed selection of insurance plans with appropriate premiums. (34)

Sushmita et al. (31) investigated the use of three ML algorithms - regression tree, M5 model tree, and random forest - to estimate individual patients' healthcare expenditures based on their historical medical and cost profiles, drawing on data from both the Washington State Inpatient Database, which includes hospital inpatient claims, and the Medical Expenditure Panel Survey, which comprises

information collected through household, employer, provider, and insurer surveys. Yang et al. (73) employed a longitudinal administrative claims dataset from a public insurance program to predict future expenditures for high-cost, high-need individuals, defined as those in the top 10% of annual expenditures. Rao et al. (74) analyzed patient-level data from the New York State Statewide Planning and Research Cooperative System and applied catboost regression, decision trees, and random forests to predict total costs, concluding that diagnosis codes, severity of illness, and length of stay were among the strongest predictors.

In our study, at the individual level we observed proportions of variance explained by the models below 50%, along with large residuals between predicted and actual values. Specifically, although for the general population more than 40% of absolute errors fell within the range of €100 - €500, we also detected non-negligible proportions of errors exceeding €100,000. Notably, these extreme deviations were predominantly in the direction of underestimation of actual costs. These findings indicate that these predictive algorithms cannot be considered sufficiently reliable for individual-level cost prediction. Nevertheless, in the context of our study and given its intended application, the relevance of individual-level prediction is limited. In the field of public health and healthcare planning, the primary interest does not lie in anticipating the expenditure of single individuals, but rather in estimating average costs at the population level and within specific subgroups of interest. Our methodological focus, then, was placed on minimizing group-level prediction error rather than optimizing individual-level accuracy. This population-based perspective mirrors real-world decision-making processes, which prioritize the assessment of collective needs and the allocation of resources across groups, rather than the prediction of outcomes for individual patients. Although individual-level predictive accuracy is not our primary objective, it remains relevant, as group-level cost estimates inevitably derive from individual predictions. The limited ability of the predictive algorithms to capture high-cost values - particularly their stronger tendency to underestimate high actual costs rather than to overestimate low actual costs - was consistently reflected in a systematic underestimation of mean costs, both at the population level and within the high-impact subgroups analyzed. As also discussed by Zhang et al. (75), since random forests average predictions across weak learners, they inherently tend to underestimate extremely high values and overestimate extremely low values. This characteristic is consistent with the systematic underestimation we observed at the upper tail of the cost distribution. At the same time, it is worth noting that the algorithm also failed to capture zero-cost observations, being unable to predict null costs for individuals with no healthcare expenditures. This reflects a broader limitation of tree-based methods: unless tree depth is allowed to grow sufficiently, these models struggle to effectively capture the

extremes of the distribution. This consideration suggests the future need to re-evaluate the hyperparameters, such as the minimum number of samples required to split a node, in order to enable the random forest to provide a more accurate representation of extreme values.
The systematic underestimation of predicted costs compared to observed ones might suggest that this underestimation reflects an overall increase in healthcare service costs in the subsequent year given that the predictive algorithms were trained on data from previous years. However, this is unlikely to be a driver of the observed pattern. In our context, healthcare tariffs are regulated, and their annual adjustments are generally limited and tightly controlled. Consequently, there is no evidence of a consistent year-to-year increase across all - or even most - services. Occasional tariff revisions may occur in specific years, but if the underestimation were attributable to such changes, we would expect to observe a pronounced underestimation restricted to those years, along with substantially more accurate predictions in the others. Gradual variations in service costs, on the other hand, tend to be absorbed by our predictive algorithms as they are updated over time, as evidenced by the comparison between results of non-updated algorithms trained over the 2011–2016 period and those trained on the immediately preceding time period. Finally, it is important to note that healthcare organizations typically operate within relatively stable annual budgets, which further constrains the extent of year-to-year fluctuations in overall expenditure.

It is crucial to recognize that the magnitude of PEs varied according to the specific population segment under consideration. Excluding the biennium 2020-2021 - when the accuracy of cost predictions was severely undermined by the unforeseeable impact of the COVID-19 pandemic - we observed generally negative PE values for patients on dialysis and for those with type 2 diabetes, though never exceeding -10%. Among patients with heart failure, temporal patterns of PE values appeared more irregular, yet by 2023 they tended to converge toward zero. By contrast, for patients with Parkinson's disease or parkinsonisms, as well as those with active neoplasia, we found a more pronounced and persistent underestimation compared to other groups, with PE values reaching as low as approximately -30%. It is worth noting that the more a group is defined by specific clinical characteristics, the greater the accuracy of mean cost predictions. For example, the group of dialysis patients with CKD is characterized by the need for a continuous, high-cost chronic therapy, with limited variability across individuals. Consequently, for these patients, we observed only limited differences between mean predicted and observed values, even during the period affected by COVID-19, where PE values remained broadly consistent with those recorded in other years. A similar trend was observed in another highly fragile group - patients with active neoplasia - though the deviation between predicted and actual mean costs was larger than that observed for dialysis patients. By

contrast, the impact of the pandemic on cost prediction is particularly evident for patients with type 2 diabetes. This may be attributable to the lower degree of protection provided to this patient group during the emergency period, which resulted in a limited access to inpatient and outpatient services. This reduction in service availability is reflected in the substantially larger PE values observed in 2020 and 2021, especially for the SC and SSC outcomes.
With regard to the substantial underestimation of predicted costs compared with actual costs for patients with Parkinson's disease and for those with neoplasia, relative to the other groups considered, this may be attributable to the considerable heterogeneity within these populations. For example, patients with Parkinson's disease are generally older and therefore likely to present with multiple comorbidities. Similarly, within the group of oncology patients, several discriminating factors - such as tumor type, stage, and grade - play a critical role. These elements, in turn, influence key cost drivers, including pharmacological therapy and frequency of hospital admissions.

Another interesting aspect concerns the extent to which the definition used to select patients within the same disease group affects the accuracy of mean cost predictions. For instance, in the case of patients with CKD on dialysis and those with type 2 diabetes, we observed no substantial differences in PE values across alternative definitions: individuals affected by the chronic condition, though it may not be the only one ("all" group); individuals for whom the chronic condition is classified as the primary condition, though it may not be the only one ("primary condition" group); individuals affected exclusively by the chronic condition ("only" group); individuals for whom the chronic condition is classified as the primary condition, with no more than two additional chronic conditions besides the one of interest ("primary condition; ≤2 comorbidities" group); and, finally, individuals for whom the chronic condition is classified as the primary condition, with three or more additional chronic conditions besides the one of interest ("primary condition; >2 comorbidities" group). On the contrary, for population segments affected by heart failure, Parkinson's disease or parkinsonisms, and active neoplasia, the differences observed across the various selection criteria were more pronounced. It is therefore essential to be mindful of how the group is defined and which patients are included, in order to properly interpret the accuracy of the predictions being generated.

Finally, we observed that the inclusion of historical cost information among the predictors did not substantially improve the accuracy of mean cost predictions, either at the population level or within high-impact patient subgroups. Although a significant increase in the proportion of variance explained was observed at the individual level when such information was included - rising from around 20% to over 40% - no meaningful improvement was detected in PE values. Overall, PE values

tended to be slightly lower in absolute terms, indicating a smaller difference between predicted and actual mean costs; however, this reduction was not sufficient to produce a significant improvement compared with predictive algorithms that did not include these variables among the predictors. Nevertheless, it is noteworthy that, in studies focused on individual-level healthcare cost prediction, prior healthcare expenditures have consistently been identified as important predictors. (34) For example, Sushmita et al. (31) concluded that previous healthcare costs alone can serve as a strong indicator of future healthcare costs. Bertsimas et al. (76) demonstrated that using only 22 cost-related features as inputs to a CART regression decision tree yielded nearly the same performance as models that also incorporated medical and demographic information. Similar conclusions were reached by Duncan et al. (77) and Kuo et al. (78).

A key strength of our study is its scalability and replicability. Relying exclusively on routinely collected administrative data, the approach can be readily extended to other regions and populations without requiring ad hoc data collection. This enables its systematic application within the Italian NHS, supporting evidence-based healthcare planning, resource allocation, and the evaluation of health policies at both regional and national levels. A further distinctive aspect of the project is the use of advanced ML techniques, which allow for the simultaneous analysis of numerous variables and the identification of complex patterns. Moreover, the identification of high-impact clinical subgroups provides actionable insights for targeted resource allocation and more precise population-level health planning. The project is designed to achieve concrete operational impact. To this end, the results have been continuously discussed in close collaboration with the HPA of Bergamo, allowing for the assessment of their practical relevance, the contextual interpretation of the data, and the alignment of the predictive tool's potential applications with real-world healthcare planning and policy decisions. This collaborative approach ensures that the research not only generates insights but also informs actionable strategies within the health system.

We acknowledge some limitations of our study to be considered. First, the training and validation of the algorithm on large, longitudinal datasets demand substantial computational power, which affects both processing time and resource requirements. Likewise, data preparation, which involves the integration of multiple large-scale data sources to reconstruct each individual's healthcare service utilization history, requires significant effort. However, the standardized and structured nature of administrative data enables automation of these processes, supporting the scalability and broader applicability of the approach despite the high initial resource investment. Second, a variety of ML techniques could be employed for healthcare cost prediction. In this study, we chose to use random

forest regressions without systematically comparing them to other approaches for several reasons, including lower computational demands, straightforward hyperparameter tuning, and greater interpretability of the results. Deep neural networks, for example, are widely recognize as “black boxes,” making it difficult to explain the complex interactions underlying their predictions, which is a critical consideration in healthcare decision-making. (74,79,80) Moreover, given that our primary focus was population-level cost prediction, method selection prioritized interpretability and the ease of aggregating individual predictions over maximizing individual-level accuracy. Third, to our knowledge, this is the first study on healthcare cost prediction to utilize such an extensive longitudinal set of data, covering a period of 13 years. However, the COVID-19 pandemic, which impacted at least the 2020-2021 period and subsequently led to permanent changes in the organization and management of the healthcare system, did not allow for a clear assessment of trends in cost prediction accuracy over time. Nonetheless, the observation that prediction accuracy in the post-COVID period is noticeably improved when using updated algorithms compared to non-updated ones provides reassurance regarding the robustness of our predictive algorithm.

As a next step, we plan to extend the validation of the predictive algorithm to additional high-impact clinical subgroups, identified through ongoing collaboration with the HPA of Bergamo. Given the complexity of the algorithms’ outputs, effective visualization tools are essential to facilitate its use by healthcare planners and institutional or regional stakeholders. Accordingly, we aim to develop a web-based application that enables users to easily access and interpret predicted healthcare costs, both at the population level and for specific subgroups, thereby supporting evidence-based decision-making and strategic planning in the health system.

In conclusion, we reconstructed the healthcare service utilization history of all adult residents in the HPA of Bergamo, as recorded in the main administrative health databases from 2011 to 2023. We then trained random forest algorithms to predict the healthcare costs of these individuals based on their prior five-year service utilization and access history, generating total and mean cost predictions at both the population level and for patient subgroups with high clinical and economic impact. We observed a tendency to underestimate predicted mean costs compared to actual costs, with the magnitude of underestimation varying across subpopulations. Greater accuracy in predicting mean costs was observed in relatively homogeneous high-impact subgroups, such as patients undergoing dialysis. In more heterogeneous groups, such as patients with Parkinson’s disease or active neoplasms, the underestimation of predicted mean costs was more pronounced. Nevertheless, within each analyzed subgroup, prediction accuracy remained consistent over time, confirming the robustness and validity of the predictive tool.

BIBLIOGRAPHY

1. Organisation for Economic Co-operation and Development (OECD). Health at a Glance 2023: OECD Indicators. OECD Publishing, Paris; 2023.
2. Boutayeb A, Boutayeb S. The burden of non communicable diseases in developing countries. Int J Equity Health. 2005; 4(1), 2.
3. About Chronic Diseases. Chronic Diseases in America. 2024. Available from: https://www.cdc.gov/chronic-disease/about/index.html [Last accessed on 13 February 2025].
4. Riecke F, Bauer L, Polzer H, et al. Effects of medical interventions on health-related quality of life in chronic disease – systematic review and meta-analysis of the 19 most common diagnoses. Frontiers in Public Health, 2024; 12, 1313685.
5. World Health Organization (WHO). Preventing chronic diseases. A vital investment: overview. 2005.
6. Van Dyke K. The incredible costs of chronic diseases: why they occur and possible preventions and/or treatments. J Health Educ Res Dev, 2016; 4, 182.
7. Bähler C, Huber CA, Brüngger B, Reich O. Multimorbidity, health care utilization and costs in an elderly community-dwelling population: a claims data based observational study. BMC Health Serv Res, 2015; 15(1), 23.
8. Glynn LG, Valderas JM, Healy P, et al. The prevalence of multimorbidity in primary care and its effect on health care utilization and cost. Family Practice, 2011; 28(5), 516–23.
9. Lehnert T, Heider D, Leicht H, et al. Health care utilization and costs of elderly persons with multiple chronic conditions. Medical Care Research and Review, 2011; 68(4), 387–420.
10. Safiliou-Rothschild C. Are older people responsible for high healthcare costs? In CESifo Forum 2009 (Vol. 10, No. 1, pp. 57-64). München: ifo Institut für Wirtschaftsforschung an der Universität München.
11. Istituto Nazionale di Statistica (ISTAT). Rapporto annuale 2024. La situazione del Paese. ISTAT, 2024.
12. Naghavi M, Zamagni G, Abbafati C, et al. State of health and inequalities among Italian regions from 2000 to 2021: a systematic analysis based on the Global Burden of Disease Study 2021. The Lancet Public Health, 2025; 10(4), e309-e320.
13. Melchiorre MG, Socci M, Quattrini S, et al. Frail older people ageing in place in Italy: use of health services and relationship with general practitioner. International journal of environmental research and public health, 2022; 19(15), 9063.
14. Istituto Nazionale di Statistica (ISTAT). Le condizioni di salute della popolazione anziana in Italia. Anno 2019. Statistiche Report, 2021; 16.

15. Istituto Superiore di Sanità (ISS). Chronic diseases and aging. Available from: https://www.iss.it/web/iss-en/chronic-diseases-and-aging1 [Last accessed on 13 February 2025].
16. Ricciardi W, Tarricone R. The evolution of the Italian National Health Service. The Lancet. 2021; 398(10317), 2193–2206.
17. World Health Organization (WHO). Global health and aging. 2014.
18. Ouwens M, Wollersheim H, Hermens R, et al. Integrated care programmes for chronically ill patients: a review of systemic reviews. International Journal for Quality in Health Care, 2005; 17(2), 141–146.
19. Wright J, Williams R, Wilkinson JR. Development and importance of health needs assessment. BMJ, 1998; 316(7140), 1310-1313.
20. Vuik SI, Mayer E, Darzi A. A quantitative evidence base for population health: Applying utilization-based cluster analysis to segment a patient population. Popul Health Metr, 2016; 14(1), 44.
21. Lynn J, Straube BM, Bell KM, et al. Using population segmentation to provide better health care for all: the "Bridges to Health" model. The Milbank Quarterly 2007; 85(2), 185-208.
22. Madotto F, Riva MA, Fornari C, et al. Administrative databases as a tool for identifying healthcare demand and costs in an over-one million population. Epidemiology, Biostatistics, and Public Health, 2013; 10(2).
23. Langton JM, Wong ST, Burge F, et al. Population segments as a tool for health care performance reporting: an exploratory study in the Canadian province of British Columbia. BMC Family Practice, 2020; 21(1), 98.
24. Crane SJ, Tung EE, Hanson GJ, et al. Use of an electronic administrative database to identify older community dwelling adults at high-risk for hospitalization or emergency department visits: the elders risk assessment index. BMC Health Serv Res, 2010; 10(1), 338.
25. Mantovani LG, Fornari C, Madotto F, et al. Burden of acute myocardial infarction. Int J Cardiol, 2011; 150(1), 111-112.
26. Schmitt J, Maywald U, Schmitt NM, et al. Cardiovascular comorbidity and cardiovascular risk factors in patients with chronic inflammatory skin diseases: a case-control study utilising a population-based administrative database. Italian Journal of Public Health, 2008; 5(3).
27. Baglio G, Sera F, Cardo S, et al. The validity of hospital administrative data for outcome measurement after hip replacement. Italian Journal of Public Health, 2009; 6(2).
28. Suissa S, Garbe E. Primer: administrative health databases in observational studies of drug effects - advantages and disadvantages. Nat Clin Pract Rheumatol, 2007; 3(12), 725–732.

29. Loon Chong J, Matchar DB. Benefits of population segmentation analysis for developing health policy to promote patient-centred care. Ann Acad Med Singap, 2017; 46(7), 287-9.
30. Lopreite M, Mauro M. The effects of population ageing on health care expenditure: a Bayesian VAR analysis using data from Italy. Health Policy, 2017; 121(6), 663–674.
31. Sushmita S, Newman S, Marquardt J, et al. Population cost prediction on public healthcare datasets. In Proceedings of the 5th international conference on digital health, 2015 (pp. 87-94).
32. Jones AM. Models for Health Care. In Michael P. Clements, and David F. Hendry (eds), The Oxford Handbook of Economic Forecasting, Oxford Handbooks, 2011.
33. Mihaylova B, Briggs A, O'Hagan A, Thompson SG. Review of statistical methods for analysing healthcare resources and costs. Health economics, 2011; 20(8), 897-916.
34. Morid AM, Kawamoto K, Ault T, et al. Supervised learning methods for predicting healthcare costs: systematic literature review and empirical evaluation. In: AMIA annual symposium proceedings, 2018 (Vol. 2017, p. 1312).
35. Gregori D, Petrinco M, Bo S, et al. Regression models for analyzing costs and their determinants in health care: an introductory review. International Journal for Quality in Health Care, 2011; 23(3), 331–341.
36. Diehr P, Yanez D, Ash A, et al. Methods for analyzing health care utilization and costs. Annual review of public health, 1999; 20(1), 125-144.
37. Regione Lombardia. Decreto n°4383 del 16/05/2011. Oggetto: Determinazioni in materia CREG (Chronic Related Group), in attuazione della D.G.R. N. IX/1479 del 30 marzo 2011.
38. La Banca Dati Assistito (BDA). Available from: https://www.ats-brescia.it/principali-malattie-croniche-osservate-tramite-bda [Last accessed on 12 August 2025].
39. ATS di Brescia. Banca dati assistiti integrata: presa in carico nel 2022. Available from: https://www.ats-brescia.it/system/files/pagina_standard/files/1241/5910/BDA2022.pdf [Last accessed on 12 August 2025].
40. World Health Organization (WHO). International classification of diseases: [9th] ninth revision, basic tabulation list with alphabetic index. 1978.
41. Deo RC. Machine learning in medicine. Circulation, 2015; 132(20), 1920–1930.
42. Mitchell TM. Machine Learning. McGraw-Hill Science/Engineering/Math; 1997.
43. Si J, Barto AG, Powell WB, Wunsch D. Reinforcement learning and its relationship to supervised learning. In: Handbook of learning and approximate dynamic programming. IEEE, 2004, pp.45-63.
44. Cunningham P, Cord M, Delany SJ. Supervised Learning. In: Machine Learning Techniques for Multimedia. Cognitive Technologies. Springer, Berlin, Heidelberg. 2008.

45. Bronstein A. Lecture 2: Supervised Learning. Available from: https://vistalab-technion.github.io/cs236605/lecture_notes/lecture_02/ [Last accessed on 27 August 2025].
46. James G., Witten D., Hastie T., et al. Unsupervised Learning. In: An Introduction to Statistical Learning. Springer Texts in Statistics. Springer, Cham. 2023.
47. Quinlan JR. C4. 5: programs for machine learning. Elsevier. 1993.
48. Kuhn M, Johnson K. Applied predictive modeling. (Vol. 26, p. 13). New York: Springer. 2013.
49. Quinlan JR. Induction of decision trees. Machine learning, 1986; 1(1), 81-106.
50. Breiman L, Friedman J, Olshen RA, Stone CJ. Classification and regression trees. Chapman and Hall/CRC. 1984.
51. Therneau T, Atkinson B, Ripley B,, Ripley MB. Package 'rpart'. 2025. Available from: http://cran.ma.ic.ac.uk/web/packages/rpart/rpart.pdf [Last accessed on 28 July 2025]
52. Choi HI. Lectures on Machine Learning (Fall 2017). Lecture 9: Classification and Regression Tree (CART). Seoul National University. 2017.
53. Gallic E. Machine Learning Training: Hands-on Sessions. 2021. Available from: https://egallic.fr/Enseignement/ML/ECB/book/index.html [Last accessed on 28 July 2025]
54. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning: with applications in R. Springer Texts in Statistics. Springer, Cham. 2023.
55. PennState. Elberly College of Science. Applied Data Mining and Statistical Learning. Minimal Cost-Complexity Pruning. Available from: https://online.stat.psu.edu/stat857/node/60/ [Last accessed on 28 July 2025]
56. Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences,1997; 55(1), 119-139.
57. Breiman L. Bagging Predictors. Machine Learning, 1996; 24, 123-140.
58. Sutton CD. Classification and regression trees, bagging, and boosting. In: Handbook of statistics (Vol. 24, pp. 303–329). 2005.
59. Breiman L. Random Forests. Machine Learning, 2001; 45, 5–32.
60. Breiman L, Cutler A, Liaw A, Wiener M. Package 'randomforest'. 2018. Available from: https://peerj.com/articles/9945/Supplemental_Data_S11.pdf [Last accessed on 28 July 2025]
61. Choi HI. Lectures on Machine Learning (Fall 2017). Lecture 10: Random Forests. Seoul National University. 2017.
62. Zhang C, Ma Y. Ensemble Machine Learning: Methods and Applications. Springer New York. 2012.

63. Probst P, Wright MN, Boulesteix AL. Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews: data mining and knowledge discovery, 2019; 9(3), e1301.
64. Boehmke B, Greenwell BM. Hands-on machine learning with R. Chapman and Hall/CRC. 2019.
65. Goldstein BA, Polley EC, Briggs FB. Random forests for genetic association studies. Statistical applications in genetics and molecular biology, 2011; 10(1).
66. Díaz-Uriarte R, and de Andrés A. Gene selection and classification of microarray data using random forest. BMC Bioinformatics, 2006; 7(1), 3.
67. Strobl C, Boulesteix AL, Kneib T, Augustin T, Zeileis A. Conditional variable importance for random forests. BMC Bioinformatics, 2008; 9(1), 307.
68. Hapfelmeier A, Ulm K. A new variable selection approach using random forests. Computational Statistics & Data Analysis, 2013; 60, 50-69.
69. Debeer D, Strobl C. Conditional permutation importance revisited. BMC Bioinformatics. 2020; 21(1), 307.
70. Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. BMC Med Res Methodol, 2019; 19(1), 64.
71. Im C, Lu Z, Mostoufi-Moab S, et al. Development and validation of age-specific risk prediction models for primary ovarian insufficiency in long-term survivors of childhood cancer: a report from the Childhood Cancer Survivor Study and St Jude Lifetime Cohort. Lancet Oncol, 2023; 24(12):1434–1442.
72. Bertsimas D, Margonis GA, Sujichantararat S, et al. Interpretable artificial intelligence to optimise use of imatinib after resection in patients with localised gastrointestinal stromal tumours: an observational cohort study. Lancet Oncol, 2024; 25(8), 1025–1037.
73. Yang C, Delcher C, Shenkman E, Ranka S. Machine learning approaches for predicting high cost high need patient expenditures in health care. Biomed Eng Online, 2018; 17(Suppl 1), 131.
74. Rao AR, Jain R, Singh M, Garg R. Predictive interpretable analytics models for forecasting healthcare costs using open healthcare data. Healthcare Analytics, 2024; 6, 100351.
75. Zhang G, Lu Y. Bias-corrected random forests in regression. Journal of Applied Statistics. 2012; 39(1), 151-160.
76. Bertsimas D, Bjarnadóttir MV, Kane MA, et al. Algorithmic prediction of health-care costs. Operations Research, 2008; 56(6), 1382–1392.

77. Duncan I, Loginov M, Ludkovski M. Testing alternative regression frameworks for predictive modeling of health care costs. North American Actuarial Journal, 2016; 20(1), 65–87.
78. Kuo RN, Dong YH, Liu JP, et al. Predicting healthcare utilization using a pharmacy-based metric with the WHO's anatomic therapeutic chemical algorithm. Med Care, 2011; 49(11):1031–1039.
79. Yang G, Ye Q, Xia J. Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: a mini-review, two showcases and beyond. Information Fusion, 2022; 77, 29–52.
80. Rubinger L., Gazendam A., Ekhtiari S., Bhandari M. Machine learning and artificial intelligence in research and healthcare. Injury, 2023; 54, S69–S73.

SUPPLEMENTARY MATERIALS

Table S1. List of predictors

Predictor	Description
1	Sex
2	Age class
3	Number of hospitalizations in the last year
4	Number of scheduled hospitalizations in the last year
5	Number of urgent hospitalizations in the last year
6	Number of day-hospital hospitalizations in the last year
7	Frequency of hospitalizations over the previous four years
8	Frequency of scheduled hospitalizations over the previous four years
9	Number of hospitalizations in the last year: class of diagnosis "Infectious and parasitic diseases (001-139)"
10	Number of scheduled hospitalizations in the last year: class of diagnosis "Infectious and parasitic diseases (001-139)"
11	Number of urgent hospitalizations in the last year: class of diagnosis "Infectious and parasitic diseases (001-139)"
12	Number of day-hospital hospitalizations in the last year: class of diagnosis "Infectious and parasitic diseases (001-139)"
13	Frequency of hospitalizations over the previous four years: class of diagnosis "Infectious and parasitic diseases (001-139)"
14	Frequency of scheduled hospitalizations over the previous four years: class of diagnosis "Infectious and parasitic diseases (001-139)"
15	Number of hospitalizations in the last year: class of diagnosis "Neoplasms (140-239)"
16	Number of scheduled hospitalizations in the last year: class of diagnosis "Neoplasms (140-239)"
17	Number of urgent hospitalizations in the last year: class of diagnosis "Neoplasms (140-239)"
18	Number of day-hospital hospitalizations in the last year: class of diagnosis "Neoplasms (140-239)"
19	Frequency of hospitalizations over the previous four years: class of diagnosis "Neoplasms (140-239)"
20	Frequency of scheduled hospitalizations over the previous four years: class of diagnosis "Neoplasms (140-239)"
21	Number of hospitalizations in the last year: class of diagnosis "Endocrine, nutritional and metabolic diseases, and immunity disorders (240-279)"
22	Number of scheduled hospitalizations in the last year: class of diagnosis "Endocrine, nutritional and metabolic diseases, and immunity disorders (240-279)"
23	Number of urgent hospitalizations in the last year: class of diagnosis "Endocrine, nutritional and metabolic diseases, and immunity disorders (240-279)"
24	Number of day-hospital hospitalizations in the last year: class of diagnosis "Endocrine, nutritional and metabolic diseases, and immunity disorders (240-279)"
25	Frequency of hospitalizations over the previous four years: class of diagnosis "Endocrine, nutritional and metabolic diseases, and immunity disorders (240-279)"
26	Frequency of scheduled hospitalizations over the previous four years: class of diagnosis "Endocrine, nutritional and metabolic diseases, and immunity disorders (240-279)"
27	Number of hospitalizations in the last year: class of diagnosis "Diseases of the blood and blood-forming organs (280-289)"
28	Number of scheduled hospitalizations in the last year: class of diagnosis "Diseases of the blood and blood-forming organs (280-289)"
29	Number of urgent hospitalizations in the last year: class of diagnosis "Diseases of the blood and blood-forming organs (280-289)"
30	Number of day-hospital hospitalizations in the last year: class of diagnosis "Diseases of the blood and blood-forming organs (280-289)"
31	Frequency of hospitalizations over the previous four years: class of diagnosis "Diseases of the blood and blood-forming organs (280-289)"
32	Frequency of scheduled hospitalizations over the previous four years: class of diagnosis "Diseases of the blood and blood-forming organs (280-289)"
33	Number of hospitalizations in the last year: class of diagnosis "Mental, behavioral and neurodevelopmental disorders (290-319)"
34	Number of scheduled hospitalizations in the last year: class of diagnosis "Mental, behavioral and neurodevelopmental disorders (290-319)"
35	Number of urgent hospitalizations in the last year: class of diagnosis "Mental, behavioral and neurodevelopmental disorders (290-319)"
36	Number of day-hospital hospitalizations in the last year: class of diagnosis "Mental, behavioral and neurodevelopmental disorders (290-319)"
37	Frequency of hospitalizations over the previous four years: class of diagnosis "Mental, behavioral and neurodevelopmental disorders (290-319)"
38	Frequency of scheduled hospitalizations over the previous four years: class of diagnosis "Mental, behavioral and neurodevelopmental disorders (290-319)"
39	Number of hospitalizations in the last year: class of diagnosis "Diseases of the nervous system and sense organs (320-389)"
40	Number of scheduled hospitalizations in the last year: class of diagnosis "Diseases of the nervous system and sense organs (320-389)"
41	Number of urgent hospitalizations in the last year: class of diagnosis "Diseases of the nervous system and sense organs (320-389)"
42	Number of day-hospital hospitalizations in the last year: class of diagnosis "Diseases of the nervous system and sense organs (320-389)"

43	Frequency of hospitalizations over the previous four years: class of diagnosis "Diseases of the nervous system and sense organs (320-389)"
44	Frequency of scheduled hospitalizations over the previous four years: class of diagnosis "Diseases of the nervous system and sense organs (320-389)"
45	Number of hospitalizations in the last year: class of diagnosis "Diseases of the circulatory system (390-459)"
46	Number of scheduled hospitalizations in the last year: class of diagnosis "Diseases of the circulatory system (390-459)"
47	Number of urgent hospitalizations in the last year: class of diagnosis "Diseases of the circulatory system (390-459)"
48	Number of day-hospital hospitalizations in the last year: class of diagnosis "Diseases of the circulatory system (390-459)"
49	Frequency of hospitalizations over the previous four years: class of diagnosis "Diseases of the circulatory system (390-459)"
50	Frequency of scheduled hospitalizations over the previous four years: class of diagnosis "Diseases of the circulatory system (390-459)"
51	Number of hospitalizations in the last year: class of diagnosis "Diseases of the respiratory system (460-519)"
52	Number of scheduled hospitalizations in the last year: class of diagnosis "Diseases of the respiratory system (460-519)"
53	Number of urgent hospitalizations in the last year: class of diagnosis "Diseases of the respiratory system (460-519)"
54	Number of day-hospital hospitalizations in the last year: class of diagnosis "Diseases of the respiratory system (460-519)"
55	Frequency of hospitalizations over the previous four years: class of diagnosis "Diseases of the respiratory system (460-519)"
56	Frequency of scheduled hospitalizations over the previous four years: class of diagnosis "Diseases of the respiratory system (460-519)"
57	Number of hospitalizations in the last year: class of diagnosis "Diseases of the digestive system (520-579)"
58	Number of scheduled hospitalizations in the last year: class of diagnosis "Diseases of the digestive system (520-579)"
59	Number of urgent hospitalizations in the last year: class of diagnosis "Diseases of the digestive system (520-579)"
60	Number of day-hospital hospitalizations in the last year: class of diagnosis "Diseases of the digestive system (520-579)"
61	Frequency of hospitalizations over the previous four years: class of diagnosis "Diseases of the digestive system (520-579)"
62	Frequency of scheduled hospitalizations over the previous four years: class of diagnosis "Diseases of the digestive system (520-579)"
63	Number of hospitalizations in the last year: class of diagnosis "Diseases of the genitourinary system (580-629)"
64	Number of scheduled hospitalizations in the last year: class of diagnosis "Diseases of the genitourinary system (580-629)"
65	Number of urgent hospitalizations in the last year: class of diagnosis "Diseases of the genitourinary system (580-629)"
66	Number of day-hospital hospitalizations in the last year: class of diagnosis "Diseases of the genitourinary system (580-629)"
67	Frequency of hospitalizations over the previous four years: class of diagnosis "Diseases of the genitourinary system (580-629)"
68	Frequency of scheduled hospitalizations over the previous four years: class of diagnosis "Diseases of the genitourinary system (580-629)"
69	Number of hospitalizations in the last year: class of diagnosis "Complications of pregnancy, childbirth, and the puerperium (630-679)"
70	Number of scheduled hospitalizations in the last year: class of diagnosis "Complications of pregnancy, childbirth, and the puerperium (630-679)"
71	Number of urgent hospitalizations in the last year: class of diagnosis "Complications of pregnancy, childbirth, and the puerperium (630-679)"
72	Number of day-hospital hospitalizations in the last year: class of diagnosis "Complications of pregnancy, childbirth, and the puerperium (630-679)"
73	Frequency of hospitalizations over the previous four years: class of diagnosis "Complications of pregnancy, childbirth, and the puerperium (630-679)"
74	Frequency of scheduled hospitalizations over the previous four years: class of diagnosis "Complications of pregnancy, childbirth, and the puerperium (630-679)"
75	Number of hospitalizations in the last year: class of diagnosis "Diseases of the skin and subcutaneous tissue (680-709)"
76	Number of scheduled hospitalizations in the last year: class of diagnosis "Diseases of the skin and subcutaneous tissue (680-709)"
77	Number of urgent hospitalizations in the last year: class of diagnosis "Diseases of the skin and subcutaneous tissue (680-709)"
78	Number of day-hospital hospitalizations in the last year: class of diagnosis "Diseases of the skin and subcutaneous tissue (680-709)"
79	Frequency of hospitalizations over the previous four years: class of diagnosis "Diseases of the skin and subcutaneous tissue (680-709)"
80	Frequency of scheduled hospitalizations over the previous four years: class of diagnosis "Diseases of the skin and subcutaneous tissue (680-709)"
81	Number of hospitalizations in the last year: class of diagnosis "Diseases of the musculoskeletal system and connective tissue (710-739)"
82	Number of scheduled hospitalizations in the last year: class of diagnosis "Diseases of the musculoskeletal system and connective tissue (710-739)"
83	Number of urgent hospitalizations in the last year: class of diagnosis "Diseases of the musculoskeletal system and connective tissue (710-739)"
84	Number of day-hospital hospitalizations in the last year: class of diagnosis "Diseases of the musculoskeletal system and connective tissue (710-739)"
85	Frequency of hospitalizations over the previous four years: class of diagnosis "Diseases of the musculoskeletal system and connective tissue (710-739)"
86	Frequency of scheduled hospitalizations over the previous four years: class of diagnosis "Diseases of the musculoskeletal system and connective tissue (710-739)"
87	Number of hospitalizations in the last year: class of diagnosis "Congenital anomalies (740-759)"

88	Number of scheduled hospitalizations in the last year: class of diagnosis "Congenital anomalies (740-759)"
89	Number of urgent hospitalizations in the last year: class of diagnosis "Congenital anomalies (740-759)"
90	Number of day-hospital hospitalizations in the last year: class of diagnosis "Congenital anomalies (740-759)"
91	Frequency of hospitalizations over the previous four years: class of diagnosis "Congenital anomalies (740-759)"
92	Frequency of scheduled hospitalizations over the previous four years: class of diagnosis "Congenital anomalies (740-759)"
93	Number of hospitalizations in the last year: class of diagnosis "Injury and poisoning (800-999)"
94	Number of scheduled hospitalizations in the last year: class of diagnosis "Injury and poisoning (800-999)"
95	Number of urgent hospitalizations in the last year: class of diagnosis "Injury and poisoning (800-999)"
96	Number of day-hospital hospitalizations in the last year: class of diagnosis "Injury and poisoning (800-999)"
97	Frequency of hospitalizations over the previous four years: class of diagnosis "Injury and poisoning (800-999)"
98	Frequency of scheduled hospitalizations over the previous four years: class of diagnosis "Injury and poisoning (800-999)"
99	Number of outpatient visits in the last year
100	Frequency of previous outpatient visits
101	Number of outpatient visits in the last year: Anesthesiology (01)
102	Frequency of outpatient visits over the previous four years: Anesthesiology (01)
103	Number of outpatient visits in the last year: Cardiology (02)
104	Frequency of outpatient visits over the previous four years: Cardiology (02)
105	Number of outpatient visits in the last year: General surgery (03)
106	Frequency of outpatient visits over the previous four years: General surgery (03)
107	Number of outpatient visits in the last year: Plastic surgery (04)
108	Frequency of outpatient visits over the previous four years: Plastic surgery (04)
109	Number of outpatient visits in the last year: Vascular surgery - Angiology (05)
110	Frequency of outpatient visits over the previous four years: Vascular surgery - Angiology (05)
111	Number of outpatient visits in the last year: Dermatology and Venereology (06)
112	Frequency of outpatient visits over the previous four years: Dermatology and Venereology (06)
113	Number of outpatient visits in the last year: Diagnostic imaging: Nuclear Medicine (07)
114	Frequency of outpatient visits over the previous four years: Diagnostic imaging: Nuclear Medicine (07)
115	Number of outpatient visits in the last year: Diagnostic imaging: Diagnostic Radiology (081)
116	Frequency of outpatient visits over the previous four years: Diagnostic imaging: Diagnostic Radiology (081)
117	Number of outpatient visits in the last year: Diagnostic imaging: Interventional (082)
118	Frequency of outpatient visits over the previous four years: Diagnostic imaging: Interventional (082)
119	Number of outpatient visits in the last year: Diagnostic imaging: Ultrasound (083)
120	Frequency of outpatient visits over the previous four years: Diagnostic imaging: Ultrasound (083)
121	Number of outpatient visits in the last year: Diagnostic imaging: MRI (084)
122	Frequency of outpatient visits over the previous four years: Diagnostic imaging: MRI (084)
123	Number of outpatient visits in the last year: Diagnostic imaging: CT Scan (085)
124	Frequency of outpatient visits over the previous four years: Diagnostic imaging: CT Scan (085)
125	Number of outpatient visits in the last year: Endocrinology (09)
126	Frequency of outpatient visits over the previous four years: Endocrinology (09)
127	Number of outpatient visits in the last year: Gastroenterology - Digestive surgery and Endoscopy (010)
128	Frequency of outpatient visits over the previous four years: Gastroenterology - Digestive surgery and Endoscopy (010)
129	Number of outpatient visits in the last year: Physical medicine and Rehabilitation (012)
130	Frequency of outpatient visits over the previous four years: Physical medicine and Rehabilitation (012)
131	Number of outpatient visits in the last year: Nephrology (0131)
132	Frequency of outpatient visits over the previous four years: Nephrology (0131)
133	Number of outpatient visits in the last year: Dialysis and inpatient dialysis (0132 e 0133)
134	Frequency of outpatient visits over the previous four years: Dialysis and inpatient dialysis (0132 e 0133)
135	Number of outpatient visits in the last year: Neurosurgery (014)
136	Frequency of outpatient visits over the previous four years: Neurosurgery (014)
137	Number of outpatient visits in the last year: Neurology (015)
138	Frequency of outpatient visits over the previous four years: Neurology (015)
139	Number of outpatient visits in the last year: Ophthalmology (016)

140	Frequency of outpatient visits over the previous four years: Ophthalmology (016)
141	Number of outpatient visits in the last year: Dentistry and Maxillofacial surgery (017)
142	Frequency of outpatient visits over the previous four years: Dentistry and Maxillofacial surgery (017)
143	Number of outpatient visits in the last year: Oncology (018)
144	Frequency of outpatient visits over the previous four years: Oncology (018)
145	Number of outpatient visits in the last year: Orthopedics and Traumatology (019)
146	Frequency of outpatient visits over the previous four years: Orthopedics and Traumatology (019)
147	Number of outpatient visits in the last year: Obstetrics and Gynecology (020)
148	Frequency of outpatient visits over the previous four years: Obstetrics and Gynecology (020)
149	Number of outpatient visits in the last year: Otorhinolaryngology (021)
150	Frequency of outpatient visits over the previous four years: Otorhinolaryngology (021)
151	Number of outpatient visits in the last year: Pulmonology (022)
152	Frequency of outpatient visits over the previous four years: Pulmonology (022)
153	Number of outpatient visits in the last year: Psychiatry (023)
154	Frequency of outpatient visits over the previous four years: Psychiatry (023)
155	Number of outpatient visits in the last year: Radiotherapy (024)
156	Frequency of outpatient visits over the previous four years: Radiotherapy (024)
157	Number of outpatient visits in the last year: Urology (025)
158	Frequency of outpatient visits over the previous four years: Urology (025)
159	Number of outpatient visits in the last year: Follow-up visit (0264)
160	Frequency of outpatient visits over the previous four years Follow-up visit (0264)
161	Number of outpatient visits in the last year: General outpatient visit (0265)
162	Frequency of outpatient visits over the previous four years General outpatient visit (0265)
163	Number of laboratory test bundles in the last year
164	Frequency of laboratory test bundles over the previous four years
165	Number of laboratory test bundles in the last year: class "Lipid profile"
166	Frequency of laboratory test bundles over the previous four years: class "Lipid profile"
167	Number of laboratory test bundles in the last year: class "Glycemic profile"
168	Frequency of laboratory test bundles over the previous four years: class "Glycemic profile"
169	Number of laboratory test bundles in the last year: class "Metabolic profile"
170	Frequency of laboratory test bundles over the previous four years: class "Metabolic profile"
171	Number of laboratory test bundles in the last year: class "Protein profile"
172	Frequency of laboratory test bundles over the previous four years: class "Protein profile"
173	Number of laboratory test bundles in the last year: class "Electrolyte profile"
174	Frequency of laboratory test bundles over the previous four years: class "Electrolyte profile"
175	Number of laboratory test bundles in the last year: class "Bone metabolism"
176	Frequency of laboratory test bundles over the previous four years: class "Bone metabolism"
177	Number of laboratory test bundles in the last year: class "Renal function"
178	Frequency of laboratory test bundles over the previous four years: class "Renal function"
179	Number of laboratory test bundles in the last year: class "Renal damage indices"
180	Frequency of laboratory test bundles over the previous four years: class "Renal damage indices"
181	Number of laboratory test bundles in the last year: class "Liver function"
182	Frequency of laboratory test bundles over the previous four years: class "Liver function"
183	Number of laboratory test bundles in the last year: class "Liver markers"
184	Frequency of laboratory test bundles over the previous four years: class "Liver markers"
185	Number of laboratory test bundles in the last year: class "Cholestasis indices"
186	Frequency of laboratory test bundles over the previous four years: class "Cholestasis indices"
187	Number of laboratory test bundles in the last year: class "Pancreatic function"
188	Frequency of laboratory test bundles over the previous four years: class "Pancreatic function"
189	Number of laboratory test bundles in the last year: class "Thyroid function"
190	Frequency of laboratory test bundles over the previous four years: class "Thyroid function"
191	Number of laboratory test bundles in the last year: class "Thyroid antibody testing"

192	Frequency of laboratory test bundles over the previous four years: class "Thyroid antibody testing"
193	Number of laboratory test bundles in the last year: class "Autoantibody screening"
194	Frequency of laboratory test bundles over the previous four years: class "Autoantibody screening"
195	Number of laboratory test bundles in the last year: class "Inflammatory markers"
196	Frequency of laboratory test bundles over the previous four years: class "Inflammatory markers"
197	Number of laboratory test bundles in the last year: class "Coagulation profile"
198	Frequency of laboratory test bundles over the previous four years: class "Coagulation profile"
199	Number of laboratory test bundles in the last year: class "Female hormonal profile"
200	Frequency of laboratory test bundles over the previous four years: class "Female hormonal profile"
201	Number of laboratory test bundles in the last year: class "Lymphocyte typing"
202	Frequency of laboratory test bundles over the previous four years: class "Lymphocyte typing"
203	Number of laboratory test bundles in the last year: class "Infectious agent detection"
204	Frequency of laboratory test bundles over the previous four years: class "Infectious agent detection"
205	Number of laboratory test bundles in the last year: class "Infectious disease antibody testing"
206	Frequency of laboratory test bundles over the previous four years: class "Infectious disease antibody testing"
207	Number of laboratory test bundles in the last year: class "Cell damage markers"
208	Frequency of laboratory test bundles over the previous four years: class "Cell damage markers"
209	Number of laboratory test bundles in the last year: class "Cardiac damage markers"
210	Frequency of laboratory test bundles over the previous four years: class "Cardiac damage markers"
211	Number of laboratory test bundles in the last year: class "Complete blood count"
212	Frequency of laboratory test bundles over the previous four years: class "Complete blood count"
213	Number of laboratory test bundles in the last year: class "Tumor markers"
214	Frequency of laboratory test bundles over the previous four years: class "Tumor markers"
215	Drug utilization in the last year: Antacids or anti-reflux drugs (A02)
216	Drug utilization over the previous four years: Antacids or anti-reflux drugs (A02)
217	Drug utilization in the last year: Antidiarrheals (A07)
218	Drug utilization over the previous four years: Antidiarrheals (A07)
219	Drug utilization in the last year: Antidiabetics: insulins (A10A)
220	Drug utilization over the previous four years: Antidiabetics: insulins (A10A)
221	Drug utilization in the last year: Oral antidiabetics (A10B)
222	Drug utilization over the previous four years: Oral antidiabetics (A10B)
223	Drug utilization in the last year: New generation oral antidiabetics - GLT1/SGLT2 (A10BJ, A10BK)
224	Drug utilization over the previous four years: New generation oral antidiabetics - GLT1/SGLT2 (A10BJ, A10BK)
225	Drug utilization in the last year: Anticoagulants or antithrombotics (B01)
226	Drug utilization over the previous four years: Anticoagulants or antithrombotics (B01)
227	Drug utilization in the last year: Antihemorrhagics (B02)
228	Drug utilization over the previous four years: Antihemorrhagics (B02)
229	Drug utilization in the last year: Antianemics (B03)
230	Drug utilization over the previous four years: Antianemics (B03)
231	Drug utilization in the last year: Cardiac glycosides (C01A)
232	Drug utilization over the previous four years: Cardiac glycosides (C01A)
233	Drug utilization in the last year: Antiarrhythmics (C01B)
234	Drug utilization over the previous four years: Antiarrhythmics (C01B)
235	Drug utilization in the last year: Vasodilators (C01D)
236	Drug utilization over the previous four years: Vasodilators (C01D)
237	Drug utilization in the last year: Antihypertensives (C02)
238	Drug utilization over the previous four years: Antihypertensives (C02)
239	Drug utilization in the last year: Diuretics (C03)
240	Drug utilization over the previous four years: Diuretics (C03)
241	Drug utilization in the last year: Beta blockers (C07)
242	Drug utilization over the previous four years: Beta blockers (C07)
243	Drug utilization in the last year: Calcium channel blockers (C08)

244	Drug utilization over the previous four years: Calcium channel blockers (C08)
245	Drug utilization in the last year: Renin-angiotensin system agents (C09)
246	Drug utilization over the previous four years: Renin-angiotensin system agents (C09)
247	Drug utilization in the last year: Lipid-lowering agents (C10)
248	Drug utilization over the previous four years: Lipid-lowering agents (C10)
249	Drug utilization in the last year: Antipsoriatics (D05)
250	Drug utilization over the previous four years: Antipsoriatics (D05)
251	Drug utilization in the last year: Drugs for prostatic hypertrophy (G04C)
252	Drug utilization over the previous four years: Drugs for prostatic hypertrophy (G04C)
253	Drug utilization in the last year: Systemic corticosteroids (H02)
254	Drug utilization over the previous four years: Systemic corticosteroids (H02)
255	Drug utilization in the last year: Thyroid medications (H03)
256	Drug utilization over the previous four years: Thyroid medications (H03)
257	Drug utilization in the last year: Antibiotics (J01)
258	Drug utilization over the previous four years: Antibiotics (J01)
259	Drug utilization in the last year: Antifungals (J02)
260	Drug utilization over the previous four years: Antifungals (J02)
261	Drug utilization in the last year: Antivirals (J05)
262	Drug utilization over the previous four years: Antivirals (J05)
263	Drug utilization in the last year: Antineoplastic agents (L01AA)
264	Drug utilization over the previous four years: Antineoplastic agents (L01AA)
265	Drug utilization in the last year: Antineoplastic agents (L01AX)
266	Drug utilization over the previous four years: Antineoplastic agents (L01AX)
267	Drug utilization in the last year: Antineoplastic agents (L01BA)
268	Drug utilization over the previous four years: Antineoplastic agents (L01BA)
269	Drug utilization in the last year: Antineoplastic agents (L01BB)
270	Drug utilization over the previous four years: Antineoplastic agents (L01BB)
271	Drug utilization in the last year: Antineoplastic agents (L01BC)
272	Drug utilization over the previous four years: Antineoplastic agents (L01BC)
273	Drug utilization in the last year: Antineoplastic agents (L01CA)
274	Drug utilization over the previous four years: Antineoplastic agents (L01CA)
275	Drug utilization in the last year: Antineoplastic agents (L01CD)
276	Drug utilization over the previous four years: Antineoplastic agents (L01CD)
277	Drug utilization in the last year: Antineoplastic agents (L01CE)
278	Drug utilization over the previous four years: Antineoplastic agents (L01CE)
279	Drug utilization in the last year: Antineoplastic agents (L01DA)
280	Drug utilization over the previous four years: Antineoplastic agents (L01DA)
281	Drug utilization in the last year: Antineoplastic agents (L01DB)
282	Drug utilization over the previous four years: Antineoplastic agents (L01DB)
283	Drug utilization in the last year: Antineoplastic agents (L01EA)
284	Drug utilization over the previous four years: Antineoplastic agents (L01EA)
285	Drug utilization in the last year: Antineoplastic agents (L01EB)
286	Drug utilization over the previous four years: Antineoplastic agents (L01EB)
287	Drug utilization in the last year: Antineoplastic agents (L01EC)
288	Drug utilization over the previous four years: Antineoplastic agents (L01EC)
289	Drug utilization in the last year: Antineoplastic agents (L01ED)
290	Drug utilization over the previous four years: Antineoplastic agents (L01ED)
291	Drug utilization in the last year: Antineoplastic agents (L01EE)
292	Drug utilization over the previous four years: Antineoplastic agents (L01EE)
293	Drug utilization in the last year: Antineoplastic agents (L01EF)
294	Drug utilization over the previous four years: Antineoplastic agents (L01EF)
295	Drug utilization in the last year: Antineoplastic agents (L01EG)

296	Drug utilization over the previous four years: Antineoplastic agents (L01EG)
297	Drug utilization in the last year: Antineoplastic agents (L01EH)
298	Drug utilization over the previous four years: Antineoplastic agents (L01EH)
299	Drug utilization in the last year: Antineoplastic agents (L01EJ)
300	Drug utilization over the previous four years: Antineoplastic agents (L01EJ)
301	Drug utilization in the last year: Antineoplastic agents (L01EK)
302	Drug utilization over the previous four years: Antineoplastic agents (L01EK)
303	Drug utilization in the last year: Antineoplastic agents (L01EL)
304	Drug utilization over the previous four years: Antineoplastic agents (L01EL)
305	Drug utilization in the last year: Antineoplastic agents (L01EM)
306	Drug utilization over the previous four years: Antineoplastic agents (L01EM)
307	Drug utilization in the last year: Antineoplastic agents (L01EX)
308	Drug utilization over the previous four years: Antineoplastic agents (L01EX)
309	Drug utilization in the last year: Antineoplastic agents (L01FA)
310	Drug utilization over the previous four years: Antineoplastic agents (L01FA)
311	Drug utilization in the last year: Antineoplastic agents (L01FB)
312	Drug utilization over the previous four years: Antineoplastic agents (L01FB)
313	Drug utilization in the last year: Antineoplastic agents (L01FC)
314	Drug utilization over the previous four years: Antineoplastic agents (L01FC)
315	Drug utilization in the last year: Antineoplastic agents (L01FD)
316	Drug utilization over the previous four years: Antineoplastic agents (L01FD)
317	Drug utilization in the last year: Antineoplastic agents (L01FE)
318	Drug utilization over the previous four years: Antineoplastic agents (L01FE)
319	Drug utilization in the last year: Antineoplastic agents (L01FF)
320	Drug utilization over the previous four years: Antineoplastic agents (L01FF)
321	Drug utilization in the last year: Antineoplastic agents (L01FG)
322	Drug utilization over the previous four years: Antineoplastic agents (L01FG)
323	Drug utilization in the last year: Antineoplastic agents (L01FX)
324	Drug utilization over the previous four years: Antineoplastic agents (L01FX)
325	Drug utilization in the last year: Antineoplastic agents (L01XA)
326	Drug utilization over the previous four years: Antineoplastic agents (L01XA)
327	Drug utilization in the last year: Antineoplastic agents (L01XB)
328	Drug utilization over the previous four years: Antineoplastic agents (L01XB)
329	Drug utilization in the last year: Antineoplastic agents (L01XD)
330	Drug utilization over the previous four years: Antineoplastic agents (L01XD)
331	Drug utilization in the last year: Antineoplastic agents (L01XF)
332	Drug utilization over the previous four years: Antineoplastic agents (L01XF)
333	Drug utilization in the last year: Antineoplastic agents (L01XG)
334	Drug utilization over the previous four years: Antineoplastic agents (L01XG)
335	Drug utilization in the last year: Antineoplastic agents (L01XJ)
336	Drug utilization over the previous four years: Antineoplastic agents (L01XJ)
337	Drug utilization in the last year: Antineoplastic agents (L01XK)
338	Drug utilization over the previous four years: Antineoplastic agents (L01XK)
339	Drug utilization in the last year: Antineoplastic agents (L01XL)
340	Drug utilization over the previous four years: Antineoplastic agents (L01XL)
341	Drug utilization in the last year: Antineoplastic agents (L01XX)
342	Drug utilization over the previous four years: Antineoplastic agents (L01XX)
343	Drug utilization in the last year: Antineoplastic agents (L01XY)
344	Drug utilization over the previous four years: Antineoplastic agents (L01XY)
345	Drug utilization in the last year: Immunomodulatory antineoplastics: endocrine therapies (L02)
346	Drug utilization over the previous four years: Immunomodulatory antineoplastics: endocrine therapies (L02)
347	Drug utilization in the last year: Immunostimulants (L03)

348	Drug utilization over the previous four years: Immunostimulants (L03)
349	Drug utilization in the last year: Immunosuppressants (L04)
350	Drug utilization over the previous four years: Immunosuppressants (L04)
351	Drug utilization in the last year: Anti-inflammatory drugs (M01)
352	Drug utilization over the previous four years: Anti-inflammatory drugs (M01)
353	Drug utilization in the last year: Anti-gout medications (M04)
354	Drug utilization over the previous four years: Anti-gout medications (M04)
355	Drug utilization in the last year: Drugs for the musculoskeletal system (M05)
356	Drug utilization over the previous four years: Drugs for the musculoskeletal system (M05)
357	Drug utilization in the last year: Anesthetics (N01)
358	Drug utilization over the previous four years: Anesthetics (N01)
359	Drug utilization in the last year: Analgesics (N02A, N02B)
360	Drug utilization over the previous four years: Analgesics (N02A, N02B)
361	Drug utilization in the last year: Anti-migraine medications (N02C)
362	Drug utilization over the previous four years: Anti-migraine medications (N02C)
363	Drug utilization in the last year: Antiepileptics (N03)
364	Drug utilization over the previous four years: Antiepileptics (N03)
365	Drug utilization in the last year: Anti-Parkinson drugs (N04)
366	Drug utilization over the previous four years: Anti-Parkinson drugs (N04)
367	Drug utilization in the last year: Psycholeptics (N05)
368	Drug utilization over the previous four years: Psycholeptics (N05)
369	Drug utilization in the last year: Antidepressants (N06A)
370	Drug utilization over the previous four years: Antidepressants (N06A)
371	Drug utilization in the last year: Anti-dementia drugs (N06D)
372	Drug utilization over the previous four years: Anti-dementia drugs (N06D)
373	Drug utilization in the last year: Respiratory system drugs (R01, R02, R03, R04, R05, R07)
374	Drug utilization over the previous four years: Respiratory system drugs (R01, R02, R03, R04, R05, R07)
375	Drug utilization in the last year: Systemic antihistamines (R06)
376	Drug utilization over the previous four years: Systemic antihistamines (R06)
377	Drug utilization in the last year: Anti-glaucoma drugs (S01E)
378	Drug utilization over the previous four years: Anti-glaucoma drugs (S01E)
379	Mean total healthcare cost over the five-year period
380	Zero-costs subject in the last year (yes/no)

Table S2. Predicted and actual costs, and prediction errors, of TC, TSC, SSC, and SC, for the whole population. Cost predictions derived from updated algorithms, excluding historical cost information from the predictors' set.

Outcome	Year	N	Predicted cost, total	Actual cost, total	Predicted cost, mean (variability interval)	Actual cost, mean	PE (%)
TC	2017	819264	977616321	1003146952	1193 (1178 - 1208)	1224	-2.5
TC	2018	821286	1007378280	1026029933	1227 (1211 - 1242)	1249	-1.8
TC	2019	823146	1027433125	1031200509	1248 (1232 - 1265)	1253	-0.4
TC	2020	824801	1036195465	944632707	1256 (1241 - 1272)	1145	9.7
TC	2021	821859	866133749	1002438689	1054 (1036 - 1070)	1220	-13.6
TC	2022	824076	1089781704	1085121065	1322 (1304 - 1343)	1317	0.4
TC	2023	826635	1103322372	1111657382	1335 (1315 - 1354)	1345	-0.7
TSC	2017	819264	810625712	832787889	989 (978 - 1002)	1017	-2.7
TSC	2018	821286	840074242	859690533	1023 (1010 - 1036)	1047	-2.3
TSC	2019	823146	864651173	861471446	1050 (1035 - 1065)	1047	0.4
TSC	2020	824801	869340955	747558897	1054 (1041 - 1068)	906	16.3
TSC	2021	821859	669904607	829900146	815 (802 - 829)	1010	-19.3
TSC	2022	824076	913246627	905777298	1108 (1090 - 1127)	1099	0.8
TSC	2023	826635	929510281	939009798	1124 (1105 - 1144)	1136	-1
SSC	2017	819264	491374858	496831125	600 (591 - 607)	606	-1.1
SSC	2018	821286	497827823	509235221	606 (598 - 614)	620	-2.2
SSC	2019	823146	510996956	515269046	621 (612 - 630)	626	-0.8
SSC	2020	824801	517908663	416161702	628 (619 - 637)	505	24.4
SSC	2021	821859	369899289	502184011	450 (443 - 458)	611	-26.3
SSC	2022	824076	552196685	545591032	670 (660 - 679)	662	1.2
SSC	2023	826635	556551420	563163596	673 (664 - 683)	681	-1.2
SC	2017	819264	659879713	667190188	805 (794 - 816)	814	-1.1
SC	2018	821286	666765974	675574621	812 (800 - 823)	823	-1.3
SC	2019	823146	674705715	684998108	820 (809 - 832)	832	-1.5
SC	2020	824801	685462959	613235512	831 (820 - 842)	743	11.8
SC	2021	821859	565598219	674722554	688 (675 - 702)	821	-16.2
SC	2022	824076	729102796	724934799	885 (872 - 898)	880	0.6
SC	2023	826635	733255959	735811181	887 (874 - 900)	890	-0.3

Table S3. Predicted and actual costs, and prediction errors, of TC, TSC, SSC, and SC, for the CKD - Dialysis patients' groups. Cost predictions derived from updated algorithms, excluding historical cost information from the predictors' set.

Group	Outcome	Year	N	Predicted cost, total	Actual cost, total	Predicted cost, mean (variability interval)	Actual cost, mean	PE (%)
CKD - Dialysis (all)	TC	2018	857	32723855	34230830	38184 (35335 - 41115)	39943	-4.4
CKD - Dialysis (all)	TC	2019	819	31846459	34671227	38885 (36434 - 40951)	42334	-8.1
CKD - Dialysis (all)	TC	2020	735	29390554	31371954	39987 (37318 - 43801)	42683	-6.3
CKD - Dialysis (all)	TC	2021	197	8477223	9397816	43032 (38153 - 46841)	47705	-9.8
CKD - Dialysis (all)	TC	2022	205	9366430	9585626	45690 (41663 - 50143)	46759	-2.3
CKD - Dialysis (all)	TC	2023	186	8207809	8603295	44128 (40421 - 47605)	46254	-4.6
CKD - Dialysis (primary)	TC	2018	836	32208969	33981409	38527 (35696 - 41490)	40648	-5.2
CKD - Dialysis (primary)	TC	2019	792	31133316	34213155	39310 (36882 - 41324)	43198	-9
CKD - Dialysis (primary)	TC	2020	716	28898002	30963607	40360 (37614 - 44226)	43245	-6.7
CKD - Dialysis (primary)	TC	2021	193	8434889	9373107	43704 (38806 - 47536)	48565	-10
CKD - Dialysis (primary)	TC	2022	200	9229888	9476906	46149 (41929 - 50470)	47385	-2.6
CKD - Dialysis (primary)	TC	2023	176	7887523	8429854	44815 (40966 - 48175)	47897	-6.4
CKD - Dialysis (only)	TC	2018	16	619705	589859	38732 (28750 - 58537)	36866	5.1
CKD - Dialysis (only)	TC	2019	20	760198	721603	38010 (31890 - 44155)	36080	5.3
CKD - Dialysis (only)	TC	2020	18	674251	706893	37458 (30462 - 44074)	39272	-4.6
CKD - Dialysis (only)	TC	2021	17	817101	871615	48065 (37672 - 57967)	51271	-6.3
CKD - Dialysis (only)	TC	2022	15	715327	768949	47688 (38874 - 55000)	51263	-7
CKD - Dialysis (only)	TC	2023	15	703164	709694	46878 (39174 - 54200)	47313	-0.9
CKD - Dialysis (primary; <=2 comorbidities)	TC	2018	259	9956470	10891394	38442 (34638 - 42902)	42052	-8.6
CKD - Dialysis (primary; <=2 comorbidities)	TC	2019	253	9871736	11353704	39019 (36032 - 41573)	44876	-13.1
CKD - Dialysis (primary; <=2 comorbidities)	TC	2020	205	8320493	8763652	40588 (37282 - 44650)	42750	-5.1
CKD - Dialysis (primary; <=2 comorbidities)	TC	2021	99	4332703	4818447	43765 (38501 - 48524)	48671	-10.1
CKD - Dialysis (primary; <=2 comorbidities)	TC	2022	103	4710349	4849595	45732 (40484 - 50638)	47083	-2.9
CKD - Dialysis (primary; <=2 comorbidities)	TC	2023	88	3938258	4399076	44753 (40350 - 48707)	49990	-10.5
CKD - Dialysis (primary; >2 comorbidities)	TC	2018	561	21632794	22500156	38561 (35673 - 41491)	40107	-3.9
CKD - Dialysis (primary; >2 comorbidities)	TC	2019	519	20501382	22137849	39502 (37251 - 41602)	42655	-7.4
CKD - Dialysis (primary; >2 comorbidities)	TC	2020	493	19903258	21493061	40372 (37443 - 44633)	43596	-7.4
CKD - Dialysis (primary; >2 comorbidities)	TC	2021	77	3285085	3683045	42663 (37401 - 47535)	47832	-10.8
CKD - Dialysis (primary; >2 comorbidities)	TC	2022	82	3804212	3858362	46393 (41708 - 51127)	47053	-1.4
CKD - Dialysis (primary; >2 comorbidities)	TC	2023	73	3246102	3321083	44467 (39442 - 49750)	45494	-2.3
CKD - Dialysis (all)	TSC	2018	857	30102933	31928113	35126 (32797 - 37427)	37256	-5.7
CKD - Dialysis (all)	TSC	2019	819	29851487	31664804	36449 (34370 - 38406)	38663	-5.7
CKD - Dialysis (all)	TSC	2020	735	26549721	29168474	36122 (34163 - 37959)	39685	-9
CKD - Dialysis (all)	TSC	2021	197	7942613	8794824	40318 (36314 - 43907)	44644	-9.7
CKD - Dialysis (all)	TSC	2022	205	8796363	8806954	42909 (39265 - 46813)	42961	-0.1
CKD - Dialysis (all)	TSC	2023	186	7604954	8120158	40887 (37115 - 44363)	43657	-6.3
CKD - Dialysis (primary)	TSC	2018	836	29643984	31699988	35459 (33103 - 37766)	37919	-6.5
CKD - Dialysis (primary)	TSC	2019	792	29199805	31224586	36868 (34781 - 38937)	39425	-6.5
CKD - Dialysis (primary)	TSC	2020	716	26106832	28804727	36462 (34442 - 38349)	40230	-9.4
CKD - Dialysis (primary)	TSC	2021	193	7905381	8773564	40961 (36927 - 44660)	45459	-9.9
CKD - Dialysis (primary)	TSC	2022	200	8681293	8705881	43406 (39708 - 47319)	43529	-0.3
CKD - Dialysis (primary)	TSC	2023	176	7335734	7957878	41680 (37716 - 45266)	45215	-7.8
CKD - Dialysis (only)	TSC	2018	16	538311	589859	33644 (25077 - 40920)	36866	-8.7
CKD - Dialysis (only)	TSC	2019	20	728640	714715	36432 (30026 - 42153)	35736	1.9
CKD - Dialysis (only)	TSC	2020	18	652510	627224	36251 (30226 - 42787)	34846	4

Group	Outcome	Year	N	Predicted cost, total	Actual cost, total	Predicted cost, mean (variability interval)	Actual cost, mean	PE (%)
CKD - Dialysis (only)	TSC	2021	17	756701	860401	44512 (34386 - 51694)	50612	-12.1
CKD - Dialysis (only)	TSC	2022	15	675564	728436	45038 (36928 - 53040)	48562	-7.3
CKD - Dialysis (only)	TSC	2023	15	646515	705864	43101 (35466 - 49931)	47058	-8.4
CKD - Dialysis (primary; <=2 comorbidities)	TSC	2018	259	9179274	9979373	35441 (32343 - 38296)	38530	-8
CKD - Dialysis (primary; <=2 comorbidities)	TSC	2019	253	9250030	9982675	36561 (34073 - 39128)	39457	-7.3
CKD - Dialysis (primary; <=2 comorbidities)	TSC	2020	205	7601487	8140107	37080 (34387 - 39522)	39708	-6.6
CKD - Dialysis (primary; <=2 comorbidities)	TSC	2021	99	4078115	4522931	41193 (36528 - 45493)	45686	-9.8
CKD - Dialysis (primary; <=2 comorbidities)	TSC	2022	103	4445029	4472124	43156 (38745 - 47421)	43419	-0.6
CKD - Dialysis (primary; <=2 comorbidities)	TSC	2023	88	3689556	4140708	41927 (37568 - 46006)	47054	-10.9
CKD - Dialysis (primary; >2 comorbidities)	TSC	2018	561	19926398	21130757	35519 (32967 - 37773)	37666	-5.7
CKD - Dialysis (primary; >2 comorbidities)	TSC	2019	519	19221135	20527196	37035 (34978 - 39056)	39551	-6.4
CKD - Dialysis (primary; >2 comorbidities)	TSC	2020	493	17852836	20037395	36213 (34060 - 38323)	40644	-10.9
CKD - Dialysis (primary; >2 comorbidities)	TSC	2021	77	3070565	3390233	39877 (35017 - 44608)	44029	-9.4
CKD - Dialysis (primary; >2 comorbidities)	TSC	2022	82	3560700	3505322	43423 (38667 - 48535)	42748	1.6
CKD - Dialysis (primary; >2 comorbidities)	TSC	2023	73	2999663	3111306	41091 (36373 - 45689)	42621	-3.6
CKD - Dialysis (all)	SSC	2018	857	27907264	29619084	32564 (30486 - 34518)	34561	-5.8
CKD - Dialysis (all)	SSC	2019	819	27731569	29212573	33860 (31890 - 35510)	35669	-5.1
CKD - Dialysis (all)	SSC	2020	735	24515401	26945521	33354 (31434 - 35220)	36661	-9
CKD - Dialysis (all)	SSC	2021	197	7397405	8199912	37550 (33618 - 40772)	41624	-9.8
CKD - Dialysis (all)	SSC	2022	205	8096280	8051662	39494 (35307 - 43265)	39276	0.6
CKD - Dialysis (all)	SSC	2023	186	6993144	7425106	37598 (34222 - 40979)	39920	-5.8
CKD - Dialysis (primary)	SSC	2018	836	27575777	29476637	32985 (30787 - 34927)	35259	-6.4
CKD - Dialysis (primary)	SSC	2019	792	27196102	28904680	34339 (32430 - 35985)	36496	-5.9
CKD - Dialysis (primary)	SSC	2020	716	24150482	26718193	33730 (31717 - 35620)	37316	-9.6
CKD - Dialysis (primary)	SSC	2021	193	7370305	8189405	38188 (34154 - 41491)	42432	-10
CKD - Dialysis (primary)	SSC	2022	200	8002691	7973133	40013 (35847 - 43699)	39866	0.4
CKD - Dialysis (primary)	SSC	2023	176	6772941	7331392	38483 (34930 - 41919)	41656	-7.6
CKD - Dialysis (only)	SSC	2018	16	504701	562625	31544 (23106 - 38445)	35164	-10.3
CKD - Dialysis (only)	SSC	2019	20	679848	678270	33992 (28319 - 39101)	33913	0.2
CKD - Dialysis (only)	SSC	2020	18	602892	593355	33494 (27487 - 39238)	32964	1.6
CKD - Dialysis (only)	SSC	2021	17	708094	802743	41653 (32476 - 48482)	47220	-11.8
CKD - Dialysis (only)	SSC	2022	15	624002	683177	41600 (33461 - 48461)	45545	-8.7
CKD - Dialysis (only)	SSC	2023	15	590562	665408	39371 (31986 - 45659)	44361	-11.2
CKD - Dialysis (primary; <=2 comorbidities)	SSC	2018	259	8503738	9303169	32833 (30201 - 35499)	35920	-8.6
CKD - Dialysis (primary; <=2 comorbidities)	SSC	2019	253	8638220	9218640	34143 (31855 - 36362)	36437	-6.3
CKD - Dialysis (primary; <=2 comorbidities)	SSC	2020	205	7031734	7584802	34301 (31639 - 36844)	36999	-7.3
CKD - Dialysis (primary; <=2 comorbidities)	SSC	2021	99	3812485	4195034	38510 (33939 - 42167)	42374	-9.1
CKD - Dialysis (primary; <=2 comorbidities)	SSC	2022	103	4101267	4078800	39818 (35557 - 44174)	39600	0.6
CKD - Dialysis (primary; <=2 comorbidities)	SSC	2023	88	3421467	3821870	38880 (34876 - 42786)	43430	-10.5
CKD - Dialysis (primary; >2 comorbidities)	SSC	2018	561	18567338	19610842	33097 (30876 - 35099)	34957	-5.3
CKD - Dialysis (primary; >2 comorbidities)	SSC	2019	519	17878034	19007770	34447 (32421 - 36423)	36624	-5.9
CKD - Dialysis (primary; >2 comorbidities)	SSC	2020	493	16515856	18540036	33501 (31401 - 35458)	37607	-10.9
CKD - Dialysis (primary; >2 comorbidities)	SSC	2021	77	2849726	3191628	37009 (31844 - 41382)	41450	-10.7
CKD - Dialysis (primary; >2 comorbidities)	SSC	2022	82	3277422	3211157	39969 (34509 - 44135)	39160	2.1
CKD - Dialysis (primary; >2 comorbidities)	SSC	2023	73	2760911	2844113	37821 (33336 - 42037)	38960	-2.9
CKD - Dialysis (all)	SC	2018	857	30439151	31921801	35518 (33093 - 38208)	37248	-4.6

Group	Outcome	Year	N	Predicted cost, total	Actual cost, total	Predicted cost, mean (variability interval)	Actual cost, mean	PE (%)
CKD - Dialysis (all)	SC	2019	819	29664306	32218996	36220 (34090 - 38037)	39339	-7.9
CKD - Dialysis (all)	SC	2020	735	27189688	29149001	36993 (34756 - 39639)	39659	-6.7
CKD - Dialysis (all)	SC	2021	197	7917629	8802904	40191 (36566 - 44137)	44685	-10.1
CKD - Dialysis (all)	SC	2022	205	8704166	8830335	42459 (39265 - 46019)	43075	-1.4
CKD - Dialysis (all)	SC	2023	186	7588680	7908243	40799 (37514 - 44249)	42517	-4
CKD - Dialysis (primary)	SC	2018	836	30043292	31758057	35937 (33435 - 38727)	37988	-5.4
CKD - Dialysis (primary)	SC	2019	792	29054259	31893250	36685 (34526 - 38527)	40269	-8.9
CKD - Dialysis (primary)	SC	2020	716	26776045	28877073	37397 (35055 - 40148)	40331	-7.3
CKD - Dialysis (primary)	SC	2021	193	7882910	8788948	40844 (37168 - 44757)	45539	-10.3
CKD - Dialysis (primary)	SC	2022	200	8594210	8744158	42971 (39834 - 46551)	43721	-1.7
CKD - Dialysis (primary)	SC	2023	176	7314083	7803367	41557 (38044 - 45129)	44337	-6.3
CKD - Dialysis (only)	SC	2018	16	590711	562625	36919 (26955 - 55756)	35164	5
CKD - Dialysis (only)	SC	2019	20	715425	685158	35771 (29414 - 41756)	34258	4.4
CKD - Dialysis (only)	SC	2020	18	633974	673024	35221 (28425 - 41560)	37390	-5.8
CKD - Dialysis (only)	SC	2021	17	768405	813957	45200 (37366 - 53625)	47880	-5.6
CKD - Dialysis (only)	SC	2022	15	670787	723690	44719 (36423 - 51328)	48246	-7.3
CKD - Dialysis (only)	SC	2023	15	649883	669239	43326 (36332 - 49865)	44616	-2.9
CKD - Dialysis (primary; <=2 comorbidities)	SC	2018	259	9272394	10215191	35801 (32444 - 39697)	39441	-9.2
CKD - Dialysis (primary; <=2 comorbidities)	SC	2019	253	9199766	10589669	36363 (33741 - 38820)	41856	-13.1
CKD - Dialysis (primary; <=2 comorbidities)	SC	2020	205	7709745	8208347	37609 (34627 - 41038)	40041	-6.1
CKD - Dialysis (primary; <=2 comorbidities)	SC	2021	99	4051595	4490550	40925 (36285 - 44844)	45359	-9.8
CKD - Dialysis (primary; <=2 comorbidities)	SC	2022	103	4389759	4456271	42619 (38746 - 46648)	43265	-1.5
CKD - Dialysis (primary; <=2 comorbidities)	SC	2023	88	3650742	4080238	41486 (37433 - 45248)	46366	-10.5
CKD - Dialysis (primary; >2 comorbidities)	SC	2018	561	20180187	20980242	35972 (33540 - 38778)	37398	-3.8
CKD - Dialysis (primary; >2 comorbidities)	SC	2019	519	19139068	20618423	36877 (34661 - 38915)	39727	-7.2
CKD - Dialysis (primary; >2 comorbidities)	SC	2020	493	18432325	19995702	37388 (34862 - 40712)	40559	-7.8
CKD - Dialysis (primary; >2 comorbidities)	SC	2021	77	3062910	3484440	39778 (34763 - 44919)	45252	-12.1
CKD - Dialysis (primary; >2 comorbidities)	SC	2022	82	3533664	3564198	43093 (38372 - 47603)	43466	-0.9
CKD - Dialysis (primary; >2 comorbidities)	SC	2023	73	3013458	3053891	41280 (37027 - 45765)	41834	-1.3

Table S4. Predicted and actual costs, and prediction errors, of TC, TSC, SSC, and SC, for the Type 2 diabetes patients' groups. Cost predictions derived from updated algorithms, excluding historical cost information from the predictors' set.

Group	Outcome	Year	N	Predicted cost, total	Actual cost, total	Predicted cost, mean (variability interval)	Actual cost, mean	PE (%)
Type 2 diabetes (all)	TC	2018	52163	159698546	171086389	3062 (2970 - 3166)	3280	-6.7
Type 2 diabetes (all)	TC	2019	52099	157891401	170754765	3031 (2941 - 3130)	3278	-7.5
Type 2 diabetes (all)	TC	2020	49958	147188236	145043549	2946 (2856 - 3048)	2903	1.5
Type 2 diabetes (all)	TC	2021	51493	141150702	163844745	2741 (2637 - 2853)	3182	-13.9
Type 2 diabetes (all)	TC	2022	53125	171754187	187547933	3233 (3118 - 3366)	3530	-8.4
Type 2 diabetes (all)	TC	2023	54281	187551075	197902704	3455 (3337 - 3571)	3646	-5.2
Type 2 diabetes (primary)	TC	2018	28924	55768452	59313131	1928 (1850 - 2009)	2051	-6
Type 2 diabetes (primary)	TC	2019	29072	57119477	60060375	1965 (1889 - 2039)	2066	-4.9
Type 2 diabetes (primary)	TC	2020	28650	58650241	54418440	2047 (1957 - 2150)	1899	7.8
Type 2 diabetes (primary)	TC	2021	27947	52158168	57800526	1866 (1750 - 1975)	2068	-9.8
Type 2 diabetes (primary)	TC	2022	28783	62310596	67824080	2165 (2061 - 2279)	2356	-8.1
Type 2 diabetes (primary)	TC	2023	29020	67367916	69864265	2321 (2217 - 2431)	2407	-3.6
Type 2 diabetes (only)	TC	2018	7748	9251721	9903754	1194 (1109 - 1300)	1278	-6.6
Type 2 diabetes (only)	TC	2019	7717	9353984	10060740	1212 (1127 - 1307)	1304	-7
Type 2 diabetes (only)	TC	2020	7660	10034638	9379571	1310 (1193 - 1466)	1224	7
Type 2 diabetes (only)	TC	2021	8574	10324789	11280507	1204 (1028 - 1384)	1316	-8.5
Type 2 diabetes (only)	TC	2022	8813	12403008	13636213	1407 (1292 - 1542)	1547	-9
Type 2 diabetes (only)	TC	2023	8830	13294757	13424288	1506 (1333 - 1665)	1520	-1
Type 2 diabetes (primary; <=2 comorbidities)	TC	2018	18607	36583354	38608158	1966 (1871 - 2063)	2075	-5.2
Type 2 diabetes (primary; <=2 comorbidities)	TC	2019	18644	37286348	38444397	2000 (1912 - 2084)	2062	-3
Type 2 diabetes (primary; <=2 comorbidities)	TC	2020	18368	38519371	34470498	2097 (1998 - 2197)	1877	11.7
Type 2 diabetes (primary; <=2 comorbidities)	TC	2021	18704	38304737	42291911	2048 (1930 - 2160)	2261	-9.4
Type 2 diabetes (primary; <=2 comorbidities)	TC	2022	19305	45720590	49946960	2368 (2245 - 2491)	2587	-8.5
Type 2 diabetes (primary; <=2 comorbidities)	TC	2023	19464	49572727	51709396	2547 (2434 - 2651)	2657	-4.1
Type 2 diabetes (primary; >2 comorbidities)	TC	2018	2569	9933378	10801219	3867 (3598 - 4218)	4204	-8
Type 2 diabetes (primary; >2 comorbidities)	TC	2019	2711	10479144	11555238	3865 (3621 - 4172)	4262	-9.3
Type 2 diabetes (primary; >2 comorbidities)	TC	2020	2622	10096232	10568372	3851 (3583 - 4173)	4031	-4.5
Type 2 diabetes (primary; >2 comorbidities)	TC	2021	669	3528642	4228108	5275 (4629 - 6053)	6320	-16.5
Type 2 diabetes (primary; >2 comorbidities)	TC	2022	665	4186999	4240907	6296 (5585 - 7207)	6377	-1.3
Type 2 diabetes (primary; >2 comorbidities)	TC	2023	726	4500432	4730581	6199 (5495 - 7076)	6516	-4.9
Type 2 diabetes (all)	TSC	2018	52163	130370350	138811087	2499 (2426 - 2580)	2661	-6.1
Type 2 diabetes (all)	TSC	2019	52099	130787724	139595614	2510 (2425 - 2589)	2679	-6.3
Type 2 diabetes (all)	TSC	2020	49958	119915015	116595066	2400 (2330 - 2476)	2334	2.8

Group	Outcome	Year	N	Predicted cost, total	Actual cost, total	Predicted cost, mean (variability interval)	Actual cost, mean	PE (%)
Type 2 diabetes (all)	TSC	2021	51493	107564438	132607266	2089 (2018 - 2172)	2575	-18.9
Type 2 diabetes (all)	TSC	2022	53125	142623371	153123753	2685 (2587 - 2796)	2882	-6.9
Type 2 diabetes (all)	TSC	2023	54281	156668262	161905048	2886 (2790 - 2985)	2983	-3.2
Type 2 diabetes (primary)	TSC	2018	28924	45807997	48357605	1584 (1526 - 1652)	1672	-5.3
Type 2 diabetes (primary)	TSC	2019	29072	47394312	49271824	1630 (1564 - 1696)	1695	-3.8
Type 2 diabetes (primary)	TSC	2020	28650	47897802	42939439	1672 (1607 - 1737)	1499	11.5
Type 2 diabetes (primary)	TSC	2021	27947	38389741	47261323	1374 (1302 - 1447)	1691	-18.8
Type 2 diabetes (primary)	TSC	2022	28783	51834068	55729556	1801 (1719 - 1896)	1936	-7
Type 2 diabetes (primary)	TSC	2023	29020	56908178	57653754	1961 (1871 - 2056)	1987	-1.3
Type 2 diabetes (only)	TSC	2018	7748	7651604	7891312	988 (927 - 1066)	1018	-3
Type 2 diabetes (only)	TSC	2019	7717	7818359	7999715	1013 (942 - 1095)	1037	-2.3
Type 2 diabetes (only)	TSC	2020	7660	8188159	7115124	1069 (990 - 1163)	929	15.1
Type 2 diabetes (only)	TSC	2021	8574	7312723	9184625	853 (753 - 959)	1071	-20.4
Type 2 diabetes (only)	TSC	2022	8813	10345196	11303849	1174 (1055 - 1306)	1283	-8.5
Type 2 diabetes (only)	TSC	2023	8830	11318715	11241715	1282 (1142 - 1411)	1273	0.7
Type 2 diabetes (primary; <=2 comorbidities)	TSC	2018	18607	30435561	31971656	1636 (1568 - 1712)	1718	-4.8
Type 2 diabetes (primary; <=2 comorbidities)	TSC	2019	18644	31284201	32360139	1678 (1601 - 1761)	1736	-3.3
Type 2 diabetes (primary; <=2 comorbidities)	TSC	2020	18368	31818085	27870874	1732 (1660 - 1805)	1517	14.2
Type 2 diabetes (primary; <=2 comorbidities)	TSC	2021	18704	28351955	34743374	1516 (1430 - 1609)	1858	-18.4
Type 2 diabetes (primary; <=2 comorbidities)	TSC	2022	19305	38039183	41042352	1970 (1879 - 2077)	2126	-7.3
Type 2 diabetes (primary; <=2 comorbidities)	TSC	2023	19464	41932662	42640733	2154 (2053 - 2258)	2191	-1.7
Type 2 diabetes (primary; >2 comorbidities)	TSC	2018	2569	7720832	8494637	3005 (2770 - 3287)	3307	-9.1
Type 2 diabetes (primary; >2 comorbidities)	TSC	2019	2711	8291752	8911970	3059 (2840 - 3308)	3287	-7
Type 2 diabetes (primary; >2 comorbidities)	TSC	2020	2622	7891557	7953441	3010 (2782 - 3283)	3033	-0.8
Type 2 diabetes (primary; >2 comorbidities)	TSC	2021	669	2725063	3333324	4073 (3559 - 4681)	4983	-18.2
Type 2 diabetes (primary; >2 comorbidities)	TSC	2022	665	3449689	3383355	5188 (4533 - 6160)	5088	2
Type 2 diabetes (primary; >2 comorbidities)	TSC	2023	726	3656801	3771307	5037 (4403 - 5813)	5195	-3
Type 2 diabetes (all)	SSC	2018	52163	74594091	75293136	1430 (1387 - 1476)	1443	-0.9
Type 2 diabetes (all)	SSC	2019	52099	75025614	75201182	1440 (1398 - 1485)	1443	-0.2
Type 2 diabetes (all)	SSC	2020	49958	69582045	56127874	1393 (1349 - 1441)	1124	24
Type 2 diabetes (all)	SSC	2021	51493	54591036	68016947	1060 (1019 - 1109)	1321	-19.7
Type 2 diabetes (all)	SSC	2022	53125	78551545	77968103	1479 (1432 - 1527)	1468	0.7
Type 2 diabetes (all)	SSC	2023	54281	83981622	81971497	1547 (1503 - 1596)	1510	2.5
Type 2 diabetes (primary)	SSC	2018	28924	26681131	26703321	922 (891 - 957)	923	-0.1
Type 2 diabetes (primary)	SSC	2019	29072	27346911	26628230	941 (908 - 978)	916	2.7
Type 2 diabetes (primary)	SSC	2020	28650	27614923	19625166	964 (923 - 1007)	685	40.7
Type 2 diabetes (primary)	SSC	2021	27947	19596465	24928720	701 (665 - 744)	892	-21.4
Type 2 diabetes (primary)	SSC	2022	28783	29642892	28894532	1030 (997 - 1064)	1004	2.6

Group	Outcome	Year	N	Predicted cost, total	Actual cost, total	Predicted cost, mean (variability interval)	Actual cost, mean	PE (%)
Type 2 diabetes (primary)	SSC	2023	29020	30705394	29515028	1058 (1019 - 1098)	1017	4
Type 2 diabetes (only)	SSC	2018	7748	4856121	4730343	627 (594 - 665)	611	2.7
Type 2 diabetes (only)	SSC	2019	7717	4903578	4754860	635 (598 - 677)	616	3.1
Type 2 diabetes (only)	SSC	2020	7660	5143459	3572084	671 (631 - 720)	466	44
Type 2 diabetes (only)	SSC	2021	8574	3930515	4861798	458 (422 - 497)	567	-19.2
Type 2 diabetes (only)	SSC	2022	8813	6189102	5752774	702 (667 - 739)	653	7.6
Type 2 diabetes (only)	SSC	2023	8830	6281568	5805163	711 (667 - 770)	657	8.2
Type 2 diabetes (primary; <=2 comorbidities)	SSC	2018	18607	17448625	17533603	938 (903 - 982)	942	-0.5
Type 2 diabetes (primary; <=2 comorbidities)	SSC	2019	18644	17804016	17487070	955 (919 - 992)	938	1.8
Type 2 diabetes (primary; <=2 comorbidities)	SSC	2020	18368	18121110	12535373	987 (946 - 1039)	682	44.6
Type 2 diabetes (primary; <=2 comorbidities)	SSC	2021	18704	14057210	18022172	752 (708 - 800)	964	-22
Type 2 diabetes (primary; <=2 comorbidities)	SSC	2022	19305	21276386	21047859	1102 (1061 - 1142)	1090	1.1
Type 2 diabetes (primary; <=2 comorbidities)	SSC	2023	19464	22140605	21421789	1138 (1093 - 1186)	1101	3.4
Type 2 diabetes (primary; >2 comorbidities)	SSC	2018	2569	4376386	4439374	1704 (1576 - 1859)	1728	-1.4
Type 2 diabetes (primary; >2 comorbidities)	SSC	2019	2711	4639317	4386300	1711 (1570 - 1861)	1618	5.8
Type 2 diabetes (primary; >2 comorbidities)	SSC	2020	2622	4350353	3517709	1659 (1519 - 1806)	1342	23.7
Type 2 diabetes (primary; >2 comorbidities)	SSC	2021	669	1608741	2044750	2405 (2031 - 2800)	3056	-21.3
Type 2 diabetes (primary; >2 comorbidities)	SSC	2022	665	2177404	2093900	3274 (2926 - 3655)	3149	4
Type 2 diabetes (primary; >2 comorbidities)	SSC	2023	726	2283221	2288076	3145 (2763 - 3581)	3152	-0.2
Type 2 diabetes (all)	SC	2018	52163	104631365	107568437	2006 (1945 - 2074)	2062	-2.7
Type 2 diabetes (all)	SC	2019	52099	102756792	106360332	1972 (1915 - 2039)	2042	-3.4
Type 2 diabetes (all)	SC	2020	49958	96717544	84576357	1936 (1869 - 2012)	1693	14.4
Type 2 diabetes (all)	SC	2021	51493	87696883	99254426	1703 (1620 - 1788)	1928	-11.6
Type 2 diabetes (all)	SC	2022	53125	108320331	112392283	2039 (1975 - 2104)	2116	-3.6
Type 2 diabetes (all)	SC	2023	54281	116143454	117969152	2140 (2067 - 2216)	2173	-1.5
Type 2 diabetes (primary)	SC	2018	28924	36698949	37658847	1269 (1222 - 1326)	1302	-2.5
Type 2 diabetes (primary)	SC	2019	29072	37066177	37416781	1275 (1229 - 1326)	1287	-0.9
Type 2 diabetes (primary)	SC	2020	28650	38203626	31104168	1333 (1271 - 1410)	1086	22.8
Type 2 diabetes (primary)	SC	2021	27947	32625122	35467923	1167 (1093 - 1252)	1269	-8
Type 2 diabetes (primary)	SC	2022	28783	40196301	40989056	1397 (1342 - 1458)	1424	-1.9
Type 2 diabetes (primary)	SC	2023	29020	41526172	41725538	1431 (1367 - 1494)	1438	-0.5
Type 2 diabetes (only)	SC	2018	7748	6414078	6742785	828 (781 - 884)	870	-4.9
Type 2 diabetes (only)	SC	2019	7717	6410398	6815885	831 (783 - 887)	883	-5.9
Type 2 diabetes (only)	SC	2020	7660	6916059	5836531	903 (821 - 1023)	762	18.5
Type 2 diabetes (only)	SC	2021	8574	6759890	6957680	788 (702 - 890)	811	-2.8
Type 2 diabetes (only)	SC	2022	8813	8125771	8085138	922 (864 - 988)	917	0.5
Type 2 diabetes (only)	SC	2023	8830	8244351	7987736	934 (863 - 1034)	905	3.2
Type 2 diabetes (primary; <=2 comorbidities)	SC	2018	18607	23669275	24170106	1272 (1215 - 1337)	1299	-2.1

Group	Outcome	Year	N	Predicted cost, total	Actual cost, total	Predicted cost, mean (variability interval)	Actual cost, mean	PE (%)
Type 2 diabetes (primary; <=2 comorbidities)	SC	2019	18644	23799188	23571328	1277 (1221 - 1335)	1264	1
Type 2 diabetes (primary; <=2 comorbidities)	SC	2020	18368	24751571	19134996	1348 (1277 - 1433)	1042	29.4
Type 2 diabetes (primary; <=2 comorbidities)	SC	2021	18704	23465640	25570708	1255 (1174 - 1349)	1367	-8.2
Type 2 diabetes (primary; <=2 comorbidities)	SC	2022	19305	29133444	29952467	1509 (1448 - 1581)	1552	-2.7
Type 2 diabetes (primary; <=2 comorbidities)	SC	2023	19464	30132018	30490452	1548 (1478 - 1623)	1567	-1.2
Type 2 diabetes (primary; >2 comorbidities)	SC	2018	2569	6615596	6745957	2575 (2379 - 2810)	2626	-1.9
Type 2 diabetes (primary; >2 comorbidities)	SC	2019	2711	6856591	7029568	2529 (2311 - 2777)	2593	-2.5
Type 2 diabetes (primary; >2 comorbidities)	SC	2020	2622	6535995	6132641	2493 (2285 - 2741)	2339	6.6
Type 2 diabetes (primary; >2 comorbidities)	SC	2021	669	2399592	2939534	3587 (3081 - 4107)	4394	-18.4
Type 2 diabetes (primary; >2 comorbidities)	SC	2022	665	2937087	2951452	4417 (3905 - 5004)	4438	-0.5
Type 2 diabetes (primary; >2 comorbidities)	SC	2023	726	3149802	3247350	4339 (3826 - 4858)	4473	-3

Table S5. Predicted and actual costs, and prediction errors, of TC, TSC, SSC, and SC, for the Heart failure patients' groups. Cost predictions derived from updated algorithms, excluding historical cost information from the predictors' set.

Group	Outcome	Year	N	Predicted cost, total	Actual cost, total	Predicted cost, mean (variability interval)	Actual cost, mean	PE (%)
Heart failure (all)	TC	2018	15867	72196185	77713559	4550 (4379 - 4750)	4898	-7.1
Heart failure (all)	TC	2019	15067	66757230	74657953	4431 (4262 - 4613)	4955	-10.6
Heart failure (all)	TC	2020	13831	61279496	63031940	4431 (4271 - 4616)	4557	-2.8
Heart failure (all)	TC	2021	14139	57863597	71996390	4092 (3931 - 4286)	5092	-19.6
Heart failure (all)	TC	2022	15214	71185076	82691593	4679 (4500 - 4904)	5435	-13.9
Heart failure (all)	TC	2023	14970	76580851	83201288	5116 (4895 - 5367)	5558	-8
Heart failure (primary)	TC	2018	11478	39470856	43092649	3439 (3299 - 3583)	3754	-8.4
Heart failure (primary)	TC	2019	10961	36476648	40346949	3328 (3191 - 3473)	3681	-9.6
Heart failure (primary)	TC	2020	10208	33888397	33108266	3320 (3185 - 3460)	3243	2.4
Heart failure (primary)	TC	2021	9955	29380485	35740869	2951 (2802 - 3109)	3590	-17.8
Heart failure (primary)	TC	2022	10938	36444683	43962359	3332 (3186 - 3498)	4019	-17.1
Heart failure (primary)	TC	2023	10624	40024787	44403164	3767 (3580 - 3965)	4180	-9.9
Heart failure (only)	TC	2018	109	161492	209556	1482 (1097 - 2350)	1923	-22.9
Heart failure (only)	TC	2019	113	161742	154185	1431 (1086 - 2312)	1364	4.9
Heart failure (only)	TC	2020	106	163440	143905	1542 (1117 - 3136)	1358	13.6
Heart failure (only)	TC	2021	4050	9535627	11684501	2354 (2211 - 2529)	2885	-18.4
Heart failure (only)	TC	2022	4628	12366261	15212784	2672 (2525 - 2846)	3287	-18.7
Heart failure (only)	TC	2023	4268	12537336	13459129	2938 (2758 - 3139)	3153	-6.8
Heart failure (primary; <=2 comorbidities)	TC	2018	2940	7447945	8127457	2533 (2380 - 2720)	2764	-8.4
Heart failure (primary; <=2 comorbidities)	TC	2019	2667	6570167	6922796	2464 (2313 - 2681)	2596	-5.1
Heart failure (primary; <=2 comorbidities)	TC	2020	2615	6428247	5811221	2458 (2295 - 2635)	2222	10.6
Heart failure (primary; <=2 comorbidities)	TC	2021	5439	17707910	21581714	3256 (3054 - 3453)	3968	-17.9
Heart failure (primary; <=2 comorbidities)	TC	2022	5795	21377699	25743051	3689 (3501 - 3897)	4442	-17
Heart failure (primary; <=2 comorbidities)	TC	2023	5857	24448514	28157501	4174 (3943 - 4444)	4807	-13.2
Heart failure (primary; >2 comorbidities)	TC	2018	8429	31861420	34755636	3780 (3617 - 3960)	4123	-8.3
Heart failure (primary; >2 comorbidities)	TC	2019	8181	29744740	33269967	3636 (3478 - 3818)	4067	-10.6
Heart failure (primary; >2 comorbidities)	TC	2020	7487	27296710	27153140	3646 (3480 - 3829)	3627	0.5
Heart failure (primary; >2 comorbidities)	TC	2021	466	2136948	2474653	4586 (3851 - 5449)	5310	-13.6
Heart failure (primary; >2 comorbidities)	TC	2022	515	2700723	3006524	5244 (4498 - 6271)	5838	-10.2
Heart failure (primary; >2 comorbidities)	TC	2023	499	3038937	2786534	6090 (5075 - 7312)	5584	9.1
Heart failure (all)	TSC	2018	15867	54093707	57111338	3409 (3285 - 3565)	3599	-5.3
Heart failure (all)	TSC	2019	15067	51391442	56340973	3411 (3276 - 3562)	3739	-8.8
Heart failure (all)	TSC	2020	13831	46742692	48107163	3380 (3254 - 3522)	3478	-2.8
Heart failure (all)	TSC	2021	14139	42714607	52032856	3021 (2893 - 3161)	3680	-17.9
Heart failure (all)	TSC	2022	15214	54958033	60862352	3612 (3445 - 3810)	4000	-9.7
Heart failure (all)	TSC	2023	14970	59168402	62631708	3952 (3765 - 4164)	4184	-5.5
Heart failure (primary)	TSC	2018	11478	28529158	30520386	2486 (2387 - 2604)	2659	-6.5
Heart failure (primary)	TSC	2019	10961	27070918	29390134	2470 (2369 - 2592)	2681	-7.9
Heart failure (primary)	TSC	2020	10208	25061757	24465505	2455 (2349 - 2567)	2397	2.4
Heart failure (primary)	TSC	2021	9955	20312844	24723314	2040 (1935 - 2150)	2484	-17.8
Heart failure (primary)	TSC	2022	10938	26574385	31151897	2430 (2314 - 2576)	2848	-14.7
Heart failure (primary)	TSC	2023	10624	29528019	32111149	2779 (2620 - 2963)	3023	-8
Heart failure (only)	TSC	2018	109	125812	112146	1154 (830 - 1854)	1029	12.2
Heart failure (only)	TSC	2019	113	129868	101857	1149 (828 - 2599)	901	27.5
Heart failure (only)	TSC	2020	106	140274	94263	1323 (823 - 3283)	889	48.8

Group	Outcome	Year	N	Predicted cost, total	Actual cost, total	Predicted cost, mean (variability interval)	Actual cost, mean	PE (%)
Heart failure (only)	TSC	2021	4050	6589656	8221126	1627 (1531 - 1750)	2030	-19.8
Heart failure (only)	TSC	2022	4628	8962592	10792350	1937 (1821 - 2069)	2332	-17
Heart failure (only)	TSC	2023	4268	9239716	9766276	2165 (2022 - 2332)	2288	-5.4
Heart failure (primary; <=2 comorbidities)	TSC	2018	2940	5496113	5658639	1869 (1759 - 2019)	1925	-2.9
Heart failure (primary; <=2 comorbidities)	TSC	2019	2667	4926563	5017149	1847 (1731 - 2025)	1881	-1.8
Heart failure (primary; <=2 comorbidities)	TSC	2020	2615	4832871	4323900	1848 (1742 - 1994)	1653	11.8
Heart failure (primary; <=2 comorbidities)	TSC	2021	5439	12253833	14743787	2253 (2114 - 2398)	2711	-16.9
Heart failure (primary; <=2 comorbidities)	TSC	2022	5795	15574962	18193228	2688 (2536 - 2892)	3139	-14.4
Heart failure (primary; <=2 comorbidities)	TSC	2023	5857	18010521	20301080	3075 (2875 - 3309)	3466	-11.3
Heart failure (primary; >2 comorbidities)	TSC	2018	8429	22907233	24749600	2718 (2599 - 2863)	2936	-7.4
Heart failure (primary; >2 comorbidities)	TSC	2019	8181	22014487	24271128	2691 (2563 - 2834)	2967	-9.3
Heart failure (primary; >2 comorbidities)	TSC	2020	7487	20088611	20047343	2683 (2557 - 2821)	2678	0.2
Heart failure (primary; >2 comorbidities)	TSC	2021	466	1469355	1758400	3153 (2696 - 3791)	3773	-16.4
Heart failure (primary; >2 comorbidities)	TSC	2022	515	2036831	2166318	3955 (3327 - 4799)	4206	-6
Heart failure (primary; >2 comorbidities)	TSC	2023	499	2277782	2043793	4565 (3721 - 5696)	4096	11.4
Heart failure (all)	SSC	2018	15867	32890807	34681823	2073 (1976 - 2174)	2186	-5.2
Heart failure (all)	SSC	2019	15067	31506083	33270526	2091 (1999 - 2189)	2208	-5.3
Heart failure (all)	SSC	2020	13831	28699339	25972273	2075 (1982 - 2169)	1878	10.5
Heart failure (all)	SSC	2021	14139	23738596	30456230	1679 (1589 - 1787)	2154	-22.1
Heart failure (all)	SSC	2022	15214	32689557	35874168	2149 (2051 - 2251)	2358	-8.9
Heart failure (all)	SSC	2023	14970	34883974	36130155	2330 (2233 - 2440)	2414	-3.4
Heart failure (primary)	SSC	2018	11478	16292934	17549829	1419 (1342 - 1495)	1529	-7.2
Heart failure (primary)	SSC	2019	10961	15386414	16364418	1404 (1327 - 1480)	1493	-6
Heart failure (primary)	SSC	2020	10208	14098855	11663208	1381 (1313 - 1459)	1143	20.9
Heart failure (primary)	SSC	2021	9955	9901892	12902462	995 (928 - 1078)	1296	-23.3
Heart failure (primary)	SSC	2022	10938	14544586	16998560	1330 (1257 - 1403)	1554	-14.4
Heart failure (primary)	SSC	2023	10624	16162649	17521661	1521 (1433 - 1608)	1649	-7.8
Heart failure (only)	SSC	2018	109	84586	92328	776 (510 - 1349)	847	-8.4
Heart failure (only)	SSC	2019	113	79306	77090	702 (501 - 1131)	682	2.9
Heart failure (only)	SSC	2020	106	73492	68103	693 (507 - 1052)	642	7.9
Heart failure (only)	SSC	2021	4050	3324355	4817627	821 (751 - 908)	1190	-31
Heart failure (only)	SSC	2022	4628	5100767	6350563	1102 (1033 - 1173)	1372	-19.7
Heart failure (only)	SSC	2023	4268	5261368	5622965	1233 (1141 - 1329)	1317	-6.4
Heart failure (primary; <=2 comorbidities)	SSC	2018	2940	3174281	3398388	1080 (998 - 1181)	1156	-6.6
Heart failure (primary; <=2 comorbidities)	SSC	2019	2667	2825110	3024747	1059 (973 - 1146)	1134	-6.6
Heart failure (primary; <=2 comorbidities)	SSC	2020	2615	2783855	2220015	1065 (982 - 1172)	849	25.4
Heart failure (primary; <=2 comorbidities)	SSC	2021	5439	5852247	7320332	1076 (992 - 1190)	1346	-20.1
Heart failure (primary; <=2 comorbidities)	SSC	2022	5795	8401310	9479668	1450 (1354 - 1545)	1636	-11.4
Heart failure (primary; <=2 comorbidities)	SSC	2023	5857	9723234	10929153	1660 (1556 - 1777)	1866	-11
Heart failure (primary; >2 comorbidities)	SSC	2018	8429	13034067	14059112	1546 (1449 - 1641)	1668	-7.3
Heart failure (primary; >2 comorbidities)	SSC	2019	8181	12481998	13262581	1526 (1441 - 1616)	1621	-5.9
Heart failure (primary; >2 comorbidities)	SSC	2020	7487	11241508	9375091	1501 (1418 - 1597)	1252	19.9
Heart failure (primary; >2 comorbidities)	SSC	2021	466	725291	764503	1556 (1198 - 2023)	1641	-5.1
Heart failure (primary; >2 comorbidities)	SSC	2022	515	1042509	1168330	2024 (1724 - 2392)	2269	-10.8
Heart failure (primary; >2 comorbidities)	SSC	2023	499	1178046	969542	2361 (1946 - 2833)	1943	21.5
Heart failure (all)	SC	2018	15867	51019336	55284044	3215 (3090 - 3366)	3484	-7.7

Group	Outcome	Year	N	Predicted cost, total	Actual cost, total	Predicted cost, mean (variability interval)	Actual cost, mean	PE (%)
Heart failure (all)	SC	2019	15067	46847912	51587506	3109 (2972 - 3241)	3424	-9.2
Heart failure (all)	SC	2020	13831	42993218	40897051	3108 (2974 - 3263)	2957	5.1
Heart failure (all)	SC	2021	14139	38472092	50419765	2721 (2570 - 2875)	3566	-23.7
Heart failure (all)	SC	2022	15214	48657935	57703409	3198 (3070 - 3333)	3793	-15.7
Heart failure (all)	SC	2023	14970	52172483	56699735	3485 (3328 - 3649)	3788	-8
Heart failure (primary)	SC	2018	11478	27066455	30122092	2358 (2243 - 2481)	2624	-10.1
Heart failure (primary)	SC	2019	10961	24654692	27321232	2249 (2137 - 2358)	2493	-9.8
Heart failure (primary)	SC	2020	10208	22695430	20305969	2223 (2104 - 2345)	1989	11.8
Heart failure (primary)	SC	2021	9955	18578187	23920017	1866 (1746 - 2004)	2403	-22.3
Heart failure (primary)	SC	2022	10938	24056458	29809023	2199 (2088 - 2313)	2725	-19.3
Heart failure (primary)	SC	2023	10624	26359860	29813676	2481 (2339 - 2625)	2806	-11.6
Heart failure (only)	SC	2018	109	122925	189738	1128 (778 - 1767)	1741	-35.2
Heart failure (only)	SC	2019	113	116018	129418	1027 (734 - 1579)	1145	-10.4
Heart failure (only)	SC	2020	106	107711	117745	1016 (758 - 1462)	1111	-8.5
Heart failure (only)	SC	2021	4050	6151126	8281002	1519 (1404 - 1663)	2045	-25.7
Heart failure (only)	SC	2022	4628	8375995	10770996	1810 (1682 - 1938)	2327	-22.2
Heart failure (only)	SC	2023	4268	8431312	9315818	1975 (1841 - 2132)	2183	-9.5
Heart failure (primary; <=2 comorbidities)	SC	2018	2940	5064283	5867206	1723 (1591 - 1871)	1996	-13.7
Heart failure (primary; <=2 comorbidities)	SC	2019	2667	4457466	4930395	1671 (1544 - 1814)	1849	-9.6
Heart failure (primary; <=2 comorbidities)	SC	2020	2615	4346839	3707336	1662 (1527 - 1788)	1418	17.2
Heart failure (primary; <=2 comorbidities)	SC	2021	5439	11066138	14158259	2035 (1866 - 2233)	2603	-21.8
Heart failure (primary; <=2 comorbidities)	SC	2022	5795	13980909	17029491	2413 (2269 - 2569)	2939	-17.9
Heart failure (primary; <=2 comorbidities)	SC	2023	5857	15983001	18785574	2729 (2556 - 2913)	3207	-14.9
Heart failure (primary; >2 comorbidities)	SC	2018	8429	21879247	24065148	2596 (2460 - 2751)	2855	-9.1
Heart failure (primary; >2 comorbidities)	SC	2019	8181	20081209	22261420	2455 (2314 - 2584)	2721	-9.8
Heart failure (primary; >2 comorbidities)	SC	2020	7487	18240880	16480888	2436 (2289 - 2588)	2201	10.7
Heart failure (primary; >2 comorbidities)	SC	2021	466	1360923	1480756	2920 (2333 - 3699)	3178	-8.1
Heart failure (primary; >2 comorbidities)	SC	2022	515	1699554	2008536	3300 (2804 - 3865)	3900	-15.4
Heart failure (primary; >2 comorbidities)	SC	2023	499	1945547	1712283	3899 (3299 - 4594)	3431	13.6

Table S6. Predicted and actual costs, and prediction errors, of TC, TSC, SSC, and SC, for the Parkinson's disease/Parkinsonian syndromes patients' groups. Cost predictions derived from updated algorithms, excluding historical cost information from the predictors' set.

Group	Outcome	Year	N	Predicted cost, total	Actual cost, total	Predicted cost, mean (variability interval)	Actual cost, mean	PE (%)
Parkinson (all)	TC	2018	3960	14496580	17386398	3661 (3309 - 4031)	4391	-16.6
Parkinson (all)	TC	2019	3829	13654357	17491095	3566 (3248 - 3917)	4568	-21.9
Parkinson (all)	TC	2020	3556	12391950	14494573	3485 (3144 - 3911)	4076	-14.5
Parkinson (all)	TC	2021	3923	11328383	16145272	2888 (2580 - 3196)	4116	-29.8
Parkinson (all)	TC	2022	3969	13407684	17756949	3378 (3030 - 3914)	4474	-24.5
Parkinson (all)	TC	2023	3887	14646858	18260157	3768 (3349 - 4434)	4698	-19.8
Parkinson (primary condition)	TC	2018	2188	6135838	7931924	2804 (2464 - 3170)	3625	-22.6
Parkinson (primary condition)	TC	2019	2194	6113940	8851356	2787 (2448 - 3187)	4034	-30.9
Parkinson (primary condition)	TC	2020	2084	5903093	7237220	2833 (2471 - 3235)	3473	-18.4
Parkinson (primary condition)	TC	2021	2457	5716145	8467500	2326 (1966 - 2693)	3446	-32.5
Parkinson (primary condition)	TC	2022	2517	6819327	9676797	2709 (2344 - 3236)	3845	-29.5
Parkinson (primary condition)	TC	2023	2424	7222687	9501033	2980 (2553 - 3624)	3920	-24
Parkinson (only)	TC	2018	887	2075780	2780993	2340 (1868 - 2905)	3135	-25.4
Parkinson (only)	TC	2019	884	2074196	3239183	2346 (1900 - 2805)	3664	-36
Parkinson (only)	TC	2020	844	1978312	2722801	2344 (1913 - 2884)	3226	-27.3
Parkinson (only)	TC	2021	922	1828647	2967768	1983 (1583 - 2430)	3219	-38.4
Parkinson (only)	TC	2022	932	2120308	3310733	2275 (1892 - 2776)	3552	-36
Parkinson (only)	TC	2023	854	2055596	3075595	2407 (1908 - 3020)	3601	-33.2
Parkinson (primary; <=2 comobidities)	TC	2018	1128	3391301	4406521	3006 (2636 - 3401)	3906	-23
Parkinson (primary; <=2 comobidities)	TC	2019	1142	3401058	4893754	2978 (2586 - 3414)	4285	-30.5
Parkinson (primary; <=2 comobidities)	TC	2020	1074	3326866	3853058	3098 (2644 - 3607)	3588	-13.7
Parkinson (primary; <=2 comobidities)	TC	2021	1392	3418887	5011495	2456 (2058 - 2895)	3600	-31.8
Parkinson (primary; <=2 comobidities)	TC	2022	1419	4037845	5491593	2846 (2449 - 3506)	3870	-26.5
Parkinson (primary; <=2 comobidities)	TC	2023	1412	4423964	5625194	3133 (2650 - 3921)	3984	-21.4
Parkinson (primary; >2 comobidities)	TC	2018	173	668757	744410	3866 (3069 - 4871)	4303	-10.2
Parkinson (primary; >2 comobidities)	TC	2019	168	638686	718419	3802 (2949 - 4966)	4276	-11.1
Parkinson (primary; >2 comobidities)	TC	2020	166	597915	661362	3602 (2821 - 4644)	3984	-9.6
Parkinson (primary; >2 comobidities)	TC	2021	143	468612	488236	3277 (2406 - 4533)	3414	-4
Parkinson (primary; >2 comobidities)	TC	2022	166	661174	874471	3983 (2964 - 6362)	5268	-24.4
Parkinson (primary; >2 comobidities)	TC	2023	158	743127	800244	4703 (3605 - 6308)	5065	-7.1
Parkinson (all)	TSC	2018	3960	11731385	14665609	2962 (2677 - 3263)	3703	-20
Parkinson (all)	TSC	2019	3829	11306089	14839374	2953 (2689 - 3272)	3876	-23.8
Parkinson (all)	TSC	2020	3556	10326795	12301563	2904 (2624 - 3240)	3459	-16.1
Parkinson (all)	TSC	2021	3923	8959140	13242495	2284 (2031 - 2542)	3376	-32.3
Parkinson (all)	TSC	2022	3969	11032820	14846479	2780 (2446 - 3352)	3741	-25.7
Parkinson (all)	TSC	2023	3887	12363287	15075434	3181 (2762 - 3844)	3878	-18
Parkinson (primary condition)	TSC	2018	2188	5011375	6923175	2290 (1981 - 2623)	3164	-27.6
Parkinson (primary condition)	TSC	2019	2194	5144349	7691858	2345 (2041 - 2699)	3506	-33.1
Parkinson (primary condition)	TSC	2020	2084	5034948	6456435	2416 (2133 - 2741)	3098	-22
Parkinson (primary condition)	TSC	2021	2457	4616598	7038464	1879 (1606 - 2167)	2865	-34.4

Group	Outcome	Year	N	Predicted cost, total	Actual cost, total	Predicted cost, mean (variability interval)	Actual cost, mean	PE (%)
Parkinson (primary condition)	TSC	2022	2517	5638234	8338204	2240 (1939 - 2818)	3313	-32.4
Parkinson (primary condition)	TSC	2023	2424	6201610	7927276	2558 (2174 - 3103)	3270	-21.8
Parkinson (only)	TSC	2018	887	1664250	2464471	1876 (1533 - 2258)	2778	-32.5
Parkinson (only)	TSC	2019	884	1757936	2851979	1989 (1602 - 2428)	3226	-38.4
Parkinson (only)	TSC	2020	844	1723414	2469550	2042 (1678 - 2483)	2926	-30.2
Parkinson (only)	TSC	2021	922	1501819	2512548	1629 (1294 - 2000)	2725	-40.2
Parkinson (only)	TSC	2022	932	1751113	2874320	1879 (1546 - 2358)	3084	-39.1
Parkinson (only)	TSC	2023	854	1798419	2637430	2106 (1689 - 2559)	3088	-31.8
Parkinson (primary; <=2 comobidities)	TSC	2018	1128	2786550	3856680	2470 (2153 - 2841)	3419	-27.7
Parkinson (primary; <=2 comobidities)	TSC	2019	1142	2855758	4182697	2501 (2164 - 2921)	3663	-31.7
Parkinson (primary; <=2 comobidities)	TSC	2020	1074	2818976	3400016	2625 (2293 - 3029)	3166	-17.1
Parkinson (primary; <=2 comobidities)	TSC	2021	1392	2754049	4088679	1978 (1702 - 2328)	2937	-32.6
Parkinson (primary; <=2 comobidities)	TSC	2022	1419	3329285	4744431	2346 (2017 - 3050)	3344	-29.8
Parkinson (primary; <=2 comobidities)	TSC	2023	1412	3790981	4633352	2685 (2243 - 3435)	3281	-18.2
Parkinson (primary; >2 comobidities)	TSC	2018	173	560575	602024	3240 (2600 - 4180)	3480	-6.9
Parkinson (primary; >2 comobidities)	TSC	2019	168	530655	657182	3159 (2520 - 4034)	3912	-19.3
Parkinson (primary; >2 comobidities)	TSC	2020	166	492559	586869	2967 (2280 - 3896)	3535	-16.1
Parkinson (primary; >2 comobidities)	TSC	2021	143	360730	437236	2523 (1860 - 3551)	3058	-17.5
Parkinson (primary; >2 comobidities)	TSC	2022	166	557836	719453	3360 (2533 - 5031)	4334	-22.5
Parkinson (primary; >2 comobidities)	TSC	2023	158	612211	656493	3875 (2990 - 5112)	4155	-6.7
Parkinson (all)	SSC	2018	3960	7704655	8206421	1946 (1744 - 2166)	2072	-6.1
Parkinson (all)	SSC	2019	3829	7235663	8210579	1890 (1690 - 2113)	2144	-11.9
Parkinson (all)	SSC	2020	3556	6471848	6077514	1820 (1606 - 2059)	1709	6.5
Parkinson (all)	SSC	2021	3923	4751137	6546559	1211 (1058 - 1391)	1669	-27.4
Parkinson (all)	SSC	2022	3969	6254063	8177209	1576 (1423 - 1780)	2060	-23.5
Parkinson (all)	SSC	2023	3887	7432922	8326508	1912 (1693 - 2144)	2142	-10.7
Parkinson (primary condition)	SSC	2018	2188	3213214	3638205	1469 (1283 - 1684)	1663	-11.7
Parkinson (primary condition)	SSC	2019	2194	3221784	4116318	1468 (1287 - 1669)	1876	-21.7
Parkinson (primary condition)	SSC	2020	2084	3078536	3020782	1477 (1269 - 1698)	1450	1.9
Parkinson (primary condition)	SSC	2021	2457	2360574	3494464	961 (796 - 1148)	1422	-32.4
Parkinson (primary condition)	SSC	2022	2517	3153139	4608655	1253 (1106 - 1435)	1831	-31.6
Parkinson (primary condition)	SSC	2023	2424	3672548	4221957	1515 (1310 - 1755)	1742	-13
Parkinson (only)	SSC	2018	887	1053011	1102226	1187 (964 - 1432)	1243	-4.5
Parkinson (only)	SSC	2019	884	1099548	1392362	1244 (1015 - 1520)	1575	-21
Parkinson (only)	SSC	2020	844	1047168	1044549	1241 (1016 - 1517)	1238	0.3
Parkinson (only)	SSC	2021	922	697608	1201115	757 (581 - 980)	1303	-41.9
Parkinson (only)	SSC	2022	932	958985	1480631	1029 (864 - 1229)	1589	-35.2
Parkinson (only)	SSC	2023	854	1065888	1352919	1248 (1000 - 1558)	1584	-21.2
Parkinson (primary; <=2 comobidities)	SSC	2018	1128	1813862	2227034	1608 (1403 - 1891)	1974	-18.6
Parkinson (primary; <=2 comobidities)	SSC	2019	1142	1804498	2445540	1580 (1334 - 1826)	2141	-26.2
Parkinson (primary; <=2 comobidities)	SSC	2020	1074	1749382	1690713	1629 (1375 - 1908)	1574	3.5
Parkinson (primary; <=2 comobidities)	SSC	2021	1392	1461781	2127606	1050 (844 - 1296)	1528	-31.3
Parkinson (primary; <=2 comobidities)	SSC	2022	1419	1900402	2719342	1339 (1173 - 1578)	1916	-30.1
Parkinson (primary; <=2 comobidities)	SSC	2023	1412	2232424	2512243	1581 (1359 - 1838)	1779	-11.1
Parkinson (primary; >2 comobidities)	SSC	2018	173	346340	308945	2002 (1503 - 2619)	1786	12.1

Group	Outcome	Year	N	Predicted cost, total	Actual cost, total	Predicted cost, mean (variability interval)	Actual cost, mean	PE (%)
Parkinson (primary; >2 comobidities)	SSC	2019	168	317737	278416	1891 (1375 - 2644)	1657	14.1
Parkinson (primary; >2 comobidities)	SSC	2020	166	281986	285520	1699 (1264 - 2311)	1720	-1.2
Parkinson (primary; >2 comobidities)	SSC	2021	143	201185	165742	1407 (916 - 2185)	1159	21.4
Parkinson (primary; >2 comobidities)	SSC	2022	166	293752	408682	1770 (1343 - 2410)	2462	-28.1
Parkinson (primary; >2 comobidities)	SSC	2023	158	374235	356795	2369 (1665 - 3259)	2258	4.9
Parkinson (all)	SC	2018	3960	10384642	10927210	2622 (2361 - 2887)	2759	-5
Parkinson (all)	SC	2019	3829	9641698	10862299	2518 (2286 - 2763)	2837	-11.2
Parkinson (all)	SC	2020	3556	8697100	8270524	2446 (2194 - 2752)	2326	5.2
Parkinson (all)	SC	2021	3923	7301588	9449336	1861 (1649 - 2158)	2409	-22.7
Parkinson (all)	SC	2022	3969	8829055	11087679	2225 (2009 - 2468)	2794	-20.4
Parkinson (all)	SC	2023	3887	9882993	11511230	2543 (2257 - 2854)	2961	-14.1
Parkinson (primary condition)	SC	2018	2188	4287144	4646953	1959 (1677 - 2251)	2124	-7.7
Parkinson (primary condition)	SC	2019	2194	4203106	5275815	1916 (1661 - 2186)	2405	-20.3
Parkinson (primary condition)	SC	2020	2084	4023584	3801568	1931 (1677 - 2239)	1824	5.8
Parkinson (primary condition)	SC	2021	2457	3630784	4923500	1478 (1237 - 1759)	2004	-26.3
Parkinson (primary condition)	SC	2022	2517	4439140	5947248	1764 (1532 - 2032)	2363	-25.4
Parkinson (primary condition)	SC	2023	2424	4784798	5795715	1974 (1724 - 2262)	2391	-17.4
Parkinson (only)	SC	2018	887	1441998	1418748	1626 (1289 - 2107)	1599	1.6
Parkinson (only)	SC	2019	884	1408138	1779566	1593 (1319 - 1966)	2013	-20.9
Parkinson (only)	SC	2020	844	1346182	1297800	1595 (1318 - 1994)	1538	3.7
Parkinson (only)	SC	2021	922	1083143	1656335	1175 (938 - 1501)	1796	-34.6
Parkinson (only)	SC	2022	932	1341431	1917044	1439 (1191 - 1769)	2057	-30
Parkinson (only)	SC	2023	854	1366605	1791084	1600 (1285 - 2010)	2097	-23.7
Parkinson (primary; <=2 comobidities)	SC	2018	1128	2393864	2776875	2122 (1820 - 2455)	2462	-13.8
Parkinson (primary; <=2 comobidities)	SC	2019	1142	2359400	3156596	2066 (1758 - 2379)	2764	-25.3
Parkinson (primary; <=2 comobidities)	SC	2020	1074	2294367	2143755	2136 (1793 - 2585)	1996	7
Parkinson (primary; <=2 comobidities)	SC	2021	1392	2243553	3050422	1612 (1303 - 1925)	2191	-26.5
Parkinson (primary; <=2 comobidities)	SC	2022	1419	2674691	3466504	1885 (1620 - 2208)	2443	-22.8
Parkinson (primary; <=2 comobidities)	SC	2023	1412	2914768	3504085	2064 (1759 - 2395)	2482	-16.8
Parkinson (primary; >2 comobidities)	SC	2018	173	451283	451331	2609 (1984 - 3387)	2609	0
Parkinson (primary; >2 comobidities)	SC	2019	168	435569	339653	2593 (1972 - 3484)	2022	28.2
Parkinson (primary; >2 comobidities)	SC	2020	166	383034	360013	2307 (1721 - 3089)	2169	6.4
Parkinson (primary; >2 comobidities)	SC	2021	143	304087	216742	2126 (1422 - 3356)	1516	40.3
Parkinson (primary; >2 comobidities)	SC	2022	166	423018	563700	2548 (1917 - 3544)	3396	-25
Parkinson (primary; >2 comobidities)	SC	2023	158	503425	500546	3186 (2318 - 4485)	3168	0.6

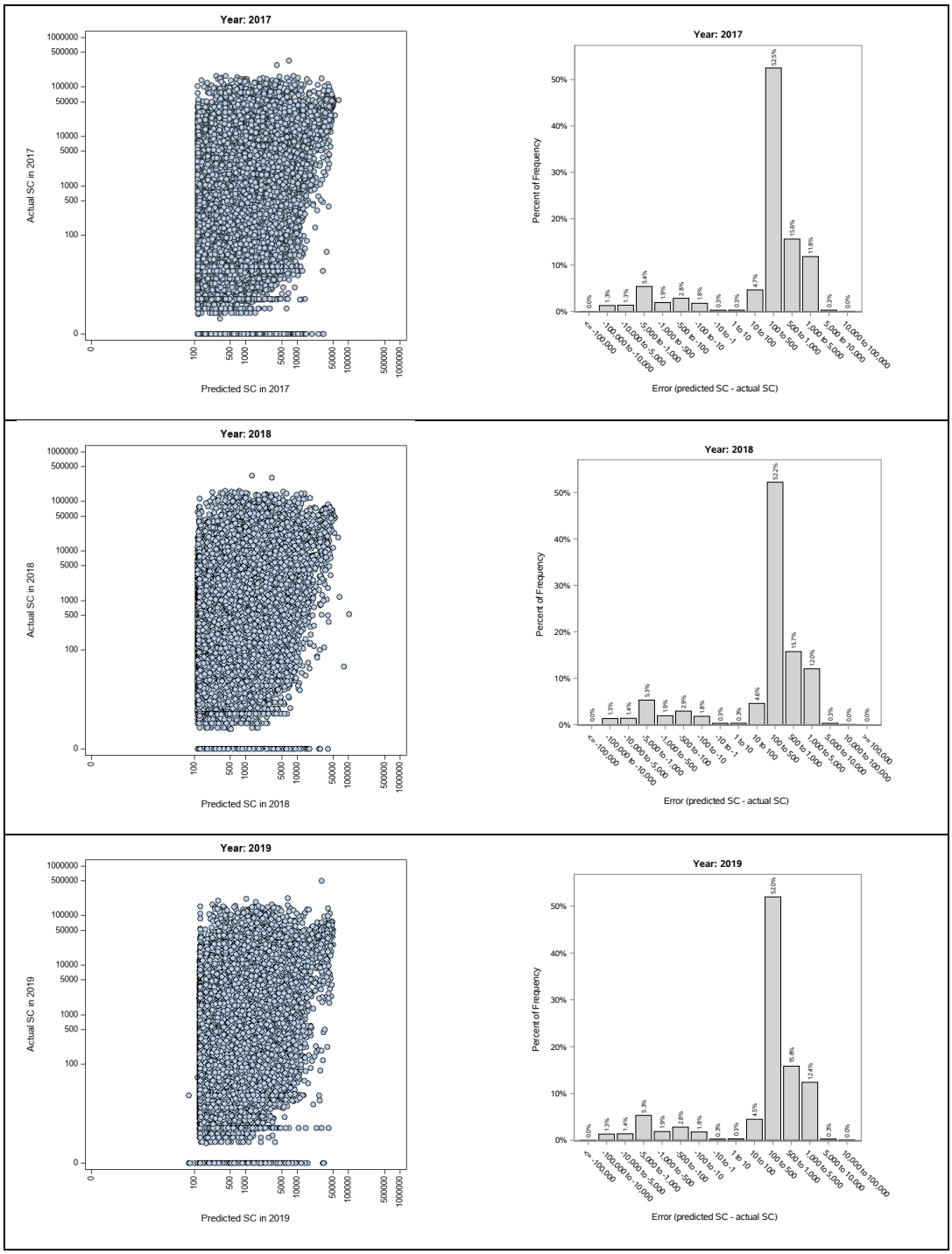
Table S7. Predicted and actual costs. and prediction errors, of TC, TSC, SSC, and SC, for the Active neoplasia patients' groups. Cost predictions derived from updated algorithms, excluding historical cost information from the predictors' set.

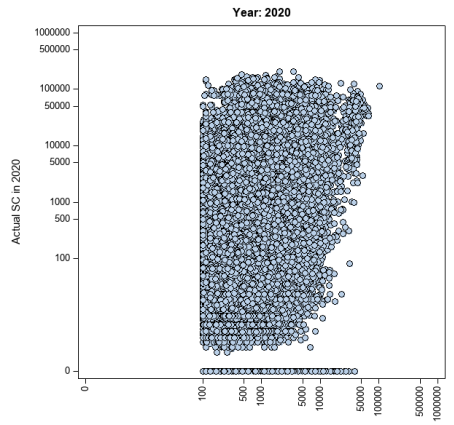
Group	Outcome	Year	N	Predicted cost. total	Actual cost. total	Predicted cost, mean (variability interval)	Actual cost, mean	PE (%)
Neoplasia (all)	TC	2018	17312	104617796	123458073	6043 (5764 - 6347)	7131	-15.3
Neoplasia (all)	TC	2019	17125	103667856	128521592	6054 (5751 - 6341)	7505	-19.3
Neoplasia (all)	TC	2020	15616	96305542	110855798	6167 (5874 - 6488)	7099	-13.1
Neoplasia (all)	TC	2021	20589	108429763	141026776	5266 (5003 - 5500)	6850	-23.1
Neoplasia (all)	TC	2022	21878	142358300	154648410	6507 (6199 - 6815)	7069	-7.9
Neoplasia (all)	TC	2023	24214	144825938	164000849	5981 (5701 - 6267)	6773	-11.7
Neoplasia (primary)	TC	2018	16838	95922799	113715539	5697 (5426 - 5990)	6754	-15.6
Neoplasia (primary)	TC	2019	16783	97243977	121019468	5794 (5504 - 6085)	7211	-19.6
Neoplasia (primary)	TC	2020	15371	91420032	104850745	5948 (5653 - 6261)	6821	-12.8
Neoplasia (primary)	TC	2021	19777	96223433	126346613	4865 (4631 - 5089)	6389	-23.8
Neoplasia (primary)	TC	2022	21019	127268581	139659496	6055 (5767 - 6343)	6644	-8.9
Neoplasia (primary)	TC	2023	23282	129140392	146561993	5547 (5285 - 5808)	6295	-11.9
Neoplasia (only)	TC	2018	5057	25436062	33741850	5030 (4637 - 5407)	6672	-24.6
Neoplasia (only)	TC	2019	5142	26712485	37616104	5195 (4829 - 5579)	7315	-29
Neoplasia (only)	TC	2020	5119	27162682	34413287	5306 (4934 - 5708)	6723	-21.1
Neoplasia (only)	TC	2021	6214	24911912	34706733	4009 (3734 - 4313)	5585	-28.2
Neoplasia (only)	TC	2022	6474	32651300	37405588	5043 (4712 - 5426)	5778	-12.7
Neoplasia (only)	TC	2023	7043	31535093	36661285	4478 (4162 - 4782)	5205	-14
Neoplasia (primary; <=2 comorbidities)	TC	2018	7838	43747036	51718891	5581 (5278 - 5904)	6598	-15.4
Neoplasia (primary; <=2 comorbidities)	TC	2019	7750	43964038	53078981	5673 (5365 - 6008)	6849	-17.2
Neoplasia (primary; <=2 comorbidities)	TC	2020	6847	40632923	45513422	5934 (5602 - 6271)	6647	-10.7
Neoplasia (primary; <=2 comorbidities)	TC	2021	11702	58312315	76952267	4983 (4723 - 5235)	6576	-24.2
Neoplasia (primary; <=2 comorbidities)	TC	2022	12528	77679723	85205953	6200 (5887 - 6536)	6801	-8.8
Neoplasia (primary; <=2 comorbidities)	TC	2023	13908	78707778	90645273	5659 (5383 - 5947)	6517	-13.2
Neoplasia (primary; >2 comorbidities)	TC	2018	3943	26739701	28254797	6782 (6357 - 7293)	7166	-5.4
Neoplasia (primary; >2 comorbidities)	TC	2019	3891	26567454	30324383	6828 (6361 - 7306)	7793	-12.4
Neoplasia (primary; >2 comorbidities)	TC	2020	3405	23624428	24924037	6938 (6494 - 7431)	7320	-5.2
Neoplasia (primary; >2 comorbidities)	TC	2021	1861	12999206	14687614	6985 (6423 - 7618)	7892	-11.5
Neoplasia (primary; >2 comorbidities)	TC	2022	2017	16937559	17047955	8397 (7810 - 9113)	8452	-0.6
Neoplasia (primary; >2 comorbidities)	TC	2023	2331	18897521	19255435	8107 (7522 - 8752)	8261	-1.9
Neoplasia (all)	TSC	2018	17312	93402552	112082408	5395 (5145 - 5667)	6474	-16.7
Neoplasia (all)	TSC	2019	17125	94198957	117531887	5501 (5209 - 5770)	6863	-19.9
Neoplasia (all)	TSC	2020	15616	87283724	101912545	5589 (5294 - 5893)	6526	-14.4
Neoplasia (all)	TSC	2021	20589	96868339	127885766	4705 (4473 - 4940)	6211	-24.3
Neoplasia (all)	TSC	2022	21878	129554692	139656225	5922 (5579 - 6228)	6383	-7.2
Neoplasia (all)	TSC	2023	24214	130239693	148610751	5379 (5122 - 5642)	6137	-12.4
Neoplasia (primary)	TSC	2018	16838	85288842	102981034	5065 (4827 - 5324)	6116	-17.2
Neoplasia (primary)	TSC	2019	16783	88247049	110600891	5258 (4971 - 5516)	6590	-20.2
Neoplasia (primary)	TSC	2020	15371	82855827	96185643	5390 (5095 - 5676)	6258	-13.9
Neoplasia (primary)	TSC	2021	19777	85590159	114126318	4328 (4130 - 4554)	5771	-25
Neoplasia (primary)	TSC	2022	21019	115412840	125722406	5491 (5188 - 5798)	5981	-8.2
Neoplasia (primary)	TSC	2023	23282	115673256	132667462	4968 (4729 - 5203)	5698	-12.8
Neoplasia (only)	TSC	2018	5057	23264698	31725004	4600 (4261 - 4937)	6273	-26.7
Neoplasia (only)	TSC	2019	5142	24945189	35497764	4851 (4522 - 5200)	6903	-29.7
Neoplasia (only)	TSC	2020	5119	25385191	32392672	4959 (4594 - 5323)	6328	-21.6

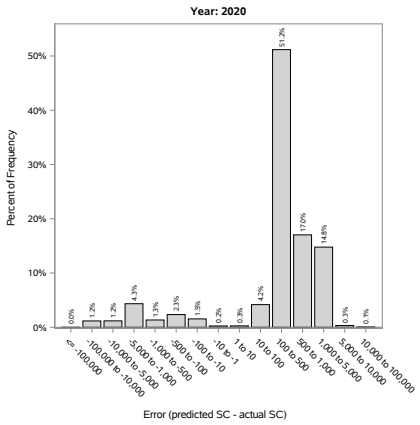
Group	Outcome	Year	N	Predicted cost. total	Actual cost. total	Predicted cost, mean (variability interval)	Actual cost, mean	PE (%)
Neoplasia (only)	TSC	2021	6214	22976465	32800748	3698 (3437 - 3974)	5279	-30
Neoplasia (only)	TSC	2022	6474	30456258	34948978	4704 (4350 - 5059)	5398	-12.9
Neoplasia (only)	TSC	2023	7043	29274126	34255823	4156 (3854 - 4437)	4864	-14.5
Neoplasia (primary; <=2 comorbidities)	TSC	2018	7838	39196206	47056938	5001 (4743 - 5289)	6004	-16.7
Neoplasia (primary; <=2 comorbidities)	TSC	2019	7750	40113756	48984262	5176 (4850 - 5469)	6321	-18.1
Neoplasia (primary; <=2 comorbidities)	TSC	2020	6847	36960530	41887447	5398 (5070 - 5700)	6118	-11.8
Neoplasia (primary; <=2 comorbidities)	TSC	2021	11702	51453959	68738466	4397 (4170 - 4620)	5874	-25.1
Neoplasia (primary; <=2 comorbidities)	TSC	2022	12528	70190809	76286655	5603 (5249 - 5931)	6089	-8
Neoplasia (primary; <=2 comorbidities)	TSC	2023	13908	70174940	81762400	5046 (4796 - 5290)	5879	-14.2
Neoplasia (primary; >2 comorbidities)	TSC	2018	3943	22827938	24199092	5789 (5412 - 6264)	6137	-5.7
Neoplasia (primary; >2 comorbidities)	TSC	2019	3891	23188103	26118865	5959 (5560 - 6381)	6713	-11.2
Neoplasia (primary; >2 comorbidities)	TSC	2020	3405	20510105	21905523	6024 (5595 - 6571)	6433	-6.4
Neoplasia (primary; >2 comorbidities)	TSC	2021	1861	11159734	12587104	5997 (5444 - 6545)	6764	-11.3
Neoplasia (primary; >2 comorbidities)	TSC	2022	2017	14765773	14486773	7321 (6712 - 8019)	7182	1.9
Neoplasia (primary; >2 comorbidities)	TSC	2023	2331	16224189	16649239	6960 (6466 - 7481)	7143	-2.6
Neoplasia (all)	SSC	2018	17312	47292932	53509752	2732 (2568 - 2905)	3091	-11.6
Neoplasia (all)	SSC	2019	17125	46022704	54026695	2687 (2524 - 2857)	3155	-14.8
Neoplasia (all)	SSC	2020	15616	41530972	46095774	2660 (2514 - 2835)	2952	-9.9
Neoplasia (all)	SSC	2021	20589	41842649	58121671	2032 (1935 - 2153)	2823	-28
Neoplasia (all)	SSC	2022	21878	61945998	65063189	2831 (2685 - 3002)	2974	-4.8
Neoplasia (all)	SSC	2023	24214	61308606	67813259	2660 (2615 - 2966)	2801	-9.6
Neoplasia (primary)	SSC	2018	16838	43142504	49553338	2562 (2409 - 2743)	2943	-12.9
Neoplasia (primary)	SSC	2019	16783	42730618	50684844	2546 (2390 - 2705)	3020	-15.7
Neoplasia (primary)	SSC	2020	15371	39080262	43262124	2542 (2402 - 2703)	2815	-9.7
Neoplasia (primary)	SSC	2021	19777	37951409	53597126	1919 (1822 - 2029)	2710	-29.2
Neoplasia (primary)	SSC	2022	21019	56680486	60390419	2697 (2561 - 2863)	2873	-6.1
Neoplasia (primary)	SSC	2023	23282	56080453	62525566	2542 (2501 - 2844)	2686	-10.3
Neoplasia (only)	SSC	2018	5057	11878209	15880504	2349 (2161 - 2562)	3140	-25.2
Neoplasia (only)	SSC	2019	5142	11841539	16723455	2303 (2136 - 2498)	3252	-29.2
Neoplasia (only)	SSC	2020	5119	11698603	14831704	2285 (2088 - 2498)	2897	-21.1
Neoplasia (only)	SSC	2021	6214	10063083	16380076	1619 (1502 - 1752)	2636	-38.6
Neoplasia (only)	SSC	2022	6474	14655325	17216745	2264 (2089 - 2465)	2659	-14.9
Neoplasia (only)	SSC	2023	7043	14152245	16741068	2285 (2215 - 2664)	2377	-15.5
Neoplasia (primary; <=2 comorbidities)	SSC	2018	7838	19432132	21620820	2479 (2322 - 2653)	2758	-10.1
Neoplasia (primary; <=2 comorbidities)	SSC	2019	7750	19091017	21926931	2463 (2298 - 2650)	2829	-12.9
Neoplasia (primary; <=2 comorbidities)	SSC	2020	6847	17209545	18592766	2513 (2359 - 2689)	2715	-7.4
Neoplasia (primary; <=2 comorbidities)	SSC	2021	11702	22837570	31339566	1952 (1834 - 2074)	2678	-27.1
Neoplasia (primary; <=2 comorbidities)	SSC	2022	12528	34448608	36325983	2750 (2595 - 2926)	2900	-5.2
Neoplasia (primary; <=2 comorbidities)	SSC	2023	13908	33970174	37851704	2513 (2430 - 2813)	2722	-10.3
Neoplasia (primary; >2 comorbidities)	SSC	2018	3943	11832163	12052014	3001 (2758 - 3316)	3057	-1.8
Neoplasia (primary; >2 comorbidities)	SSC	2019	3891	11798063	12034458	3032 (2791 - 3313)	3093	-2
Neoplasia (primary; >2 comorbidities)	SSC	2020	3405	10172113	9837654	2987 (2741 - 3274)	2889	3.4
Neoplasia (primary; >2 comorbidities)	SSC	2021	1861	5050756	5877483	2714 (2459 - 2996)	3158	-14.1
Neoplasia (primary; >2 comorbidities)	SSC	2022	2017	7576553	6847691	3756 (3399 - 4128)	3395	10.6
Neoplasia (primary; >2 comorbidities)	SSC	2023	2331	7958033	7932793	2987 (2888 - 3382)	3403	0.3
Neoplasia (all)	SC	2018	17312	58958989	64885417	3406 (3223 - 3607)	3748	-9.1

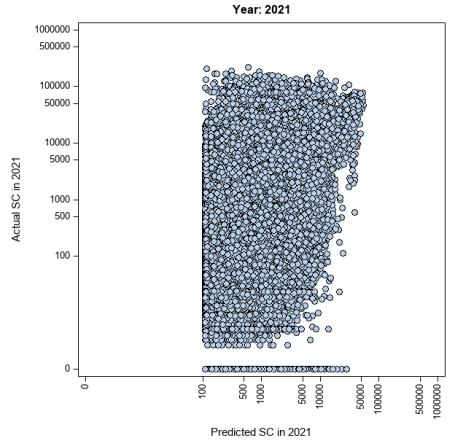
Group	Outcome	Year	N	Predicted cost. total	Actual cost. total	Predicted cost, mean (variability interval)	Actual cost, mean	PE (%)
Neoplasia (all)	SC	2019	17125	55871811	65016399	3263 (3044 - 3454)	3797	-14.1
Neoplasia (all)	SC	2020	15616	50672162	55039027	3245 (3070 - 3417)	3525	-7.9
Neoplasia (all)	SC	2021	20589	54069560	71262681	2626 (2489 - 2770)	3461	-24.1
Neoplasia (all)	SC	2022	21878	75108112	80055374	3433 (3245 - 3645)	3659	-6.2
Neoplasia (all)	SC	2023	24214	76228665	83203357	3148 (3003 - 3317)	3436	-8.4
Neoplasia (primary)	SC	2018	16838	54155493	60287842	3216 (3041 - 3415)	3580	-10.2
Neoplasia (primary)	SC	2019	16783	52104709	61103420	3105 (2907 - 3293)	3641	-14.7
Neoplasia (primary)	SC	2020	15371	47766057	51927227	3108 (2940 - 3277)	3378	-8
Neoplasia (primary)	SC	2021	19777	49061452	65817421	2481 (2363 - 2604)	3328	-25.5
Neoplasia (primary)	SC	2022	21019	68804980	74327509	3273 (3092 - 3472)	3536	-7.4
Neoplasia (primary)	SC	2023	23282	69723645	76420097	2995 (2858 - 3155)	3282	-8.8
Neoplasia (only)	SC	2018	5057	14024564	17897350	2773 (2536 - 3033)	3539	-21.6
Neoplasia (only)	SC	2019	5142	13706358	18841795	2666 (2442 - 2914)	3664	-27.3
Neoplasia (only)	SC	2020	5119	13463992	16852319	2630 (2425 - 2839)	3292	-20.1
Neoplasia (only)	SC	2021	6214	12171491	18286060	1959 (1817 - 2115)	2943	-33.4
Neoplasia (only)	SC	2022	6474	16964438	19673354	2620 (2411 - 2829)	3039	-13.8
Neoplasia (only)	SC	2023	7043	16435544	19146531	2334 (2171 - 2514)	2719	-14.2
Neoplasia (primary; <=2 comorbidities)	SC	2018	7838	24284898	26282773	3098 (2901 - 3326)	3353	-7.6
Neoplasia (primary; <=2 comorbidities)	SC	2019	7750	23115086	26021649	2983 (2759 - 3215)	3358	-11.2
Neoplasia (primary; <=2 comorbidities)	SC	2020	6847	20927789	22218741	3056 (2874 - 3265)	3245	-5.8
Neoplasia (primary; <=2 comorbidities)	SC	2021	11702	29883652	39553367	2554 (2416 - 2707)	3380	-24.4
Neoplasia (primary; <=2 comorbidities)	SC	2022	12528	42161646	45245281	3365 (3158 - 3588)	3612	-6.8
Neoplasia (primary; <=2 comorbidities)	SC	2023	13908	42614845	46734578	3064 (2908 - 3219)	3360	-8.8
Neoplasia (primary; >2 comorbidities)	SC	2018	3943	15846030	16107719	4019 (3751 - 4333)	4085	-1.6
Neoplasia (primary; >2 comorbidities)	SC	2019	3891	15283265	16239976	3928 (3625 - 4275)	4174	-5.9
Neoplasia (primary; >2 comorbidities)	SC	2020	3405	13374276	12856168	3928 (3636 - 4259)	3776	4
Neoplasia (primary; >2 comorbidities)	SC	2021	1861	7006310	7977994	3765 (3393 - 4176)	4287	-12.2
Neoplasia (primary; >2 comorbidities)	SC	2022	2017	9678896	9408873	4799 (4382 - 5319)	4665	2.9
Neoplasia (primary; >2 comorbidities)	SC	2023	2331	10673255	10538989	4579 (4216 - 4965)	4521	1.3

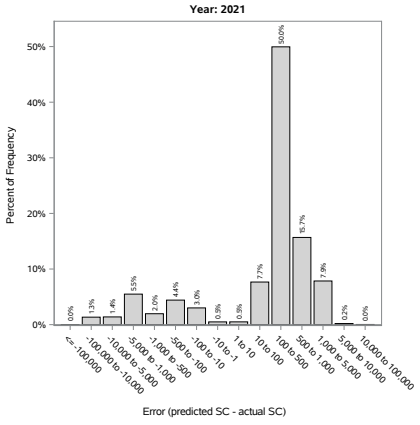
Figure S1. Scatter plots of actual SC vs. predicted SC (left-hand panels) and distributions of absolute errors (right-hand panels), in the whole population. Cost predictions derived from updated algorithms, excluding historical cost information from the predictors' set.

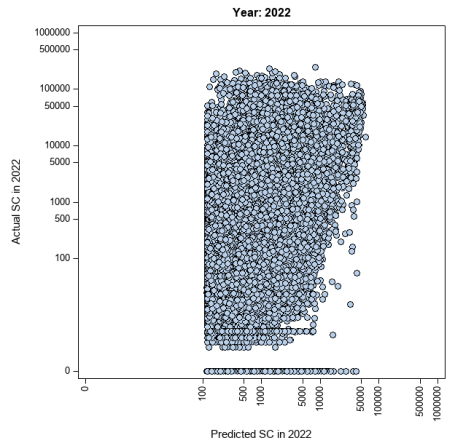


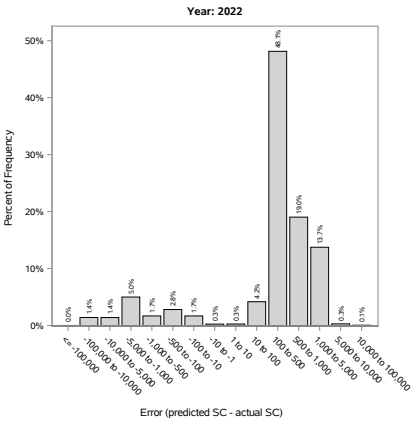

Year: 2020
Actual SC in 2020
Predicted SC in 2020

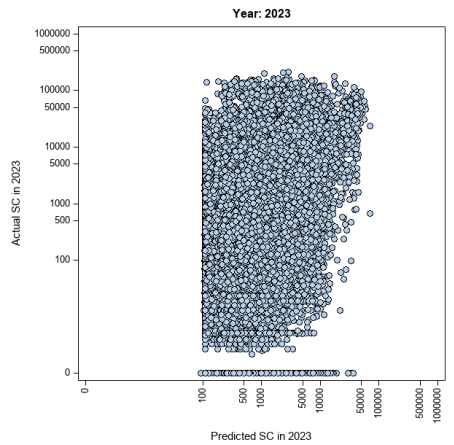

Year: 2020
Percent of Frequency
Error (predicted SC - actual SC)


Year: 2021
Actual SC in 2021
Predicted SC in 2021


Year: 2021
Percent of Frequency
Error (predicted SC - actual SC)


Year: 2022
Actual SC in 2022
Predicted SC in 2022


Year: 2022
Percent of Frequency
Error (predicted SC - actual SC)


Year: 2023
Actual SC in 2023
Predicted SC in 2023

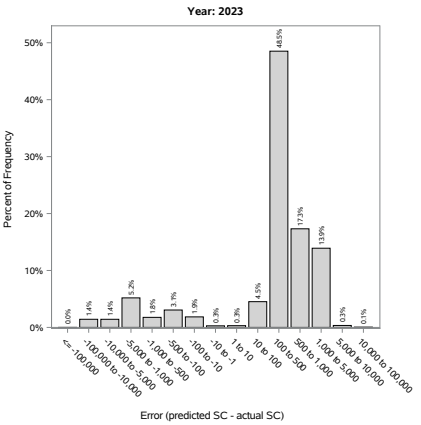
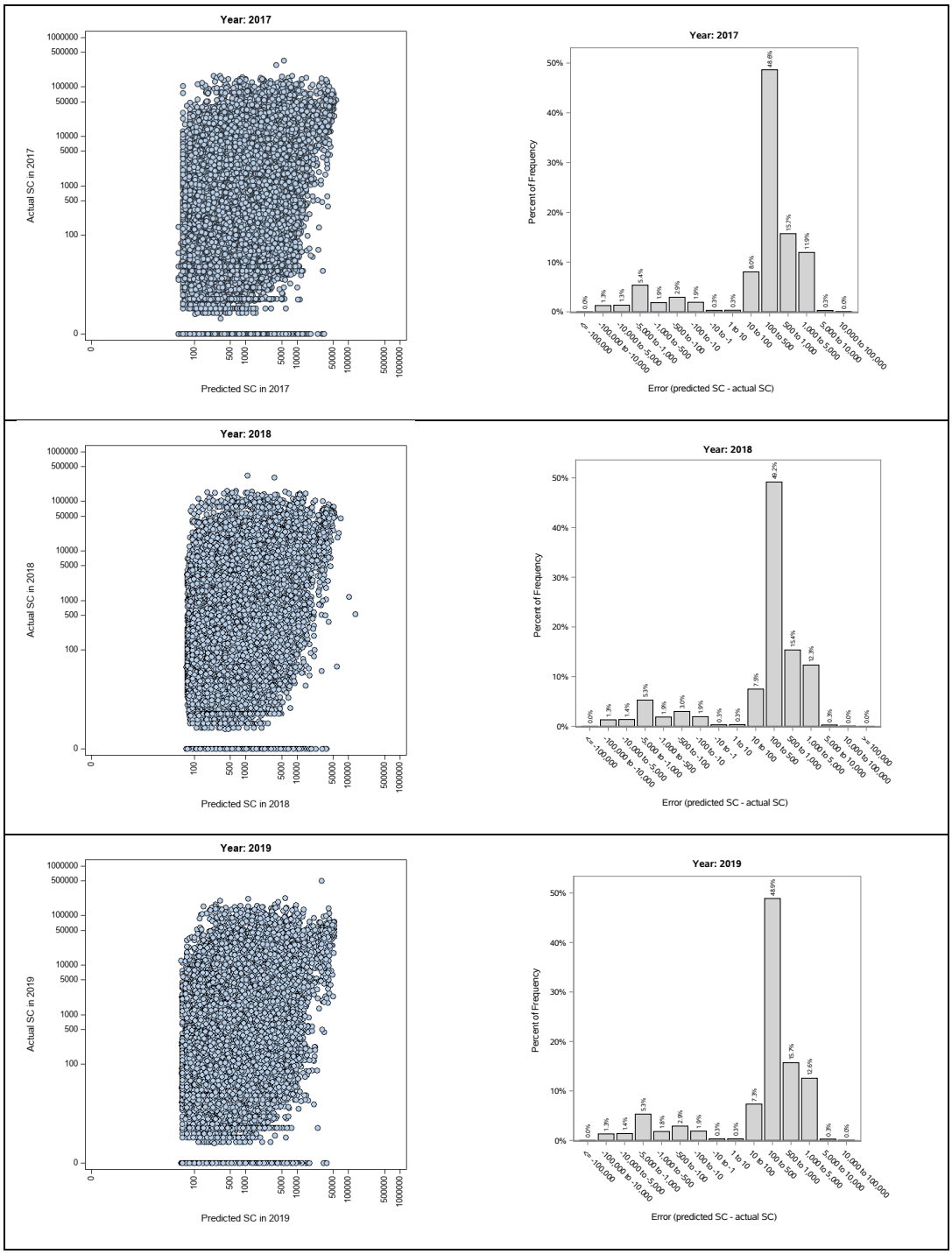
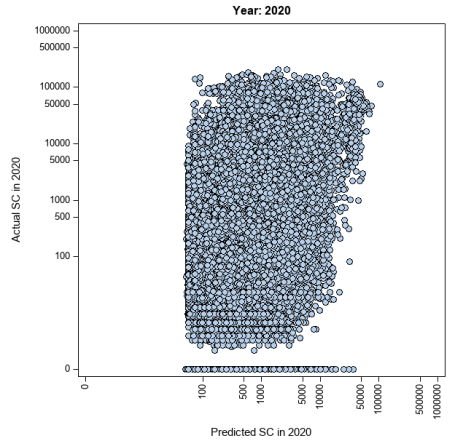
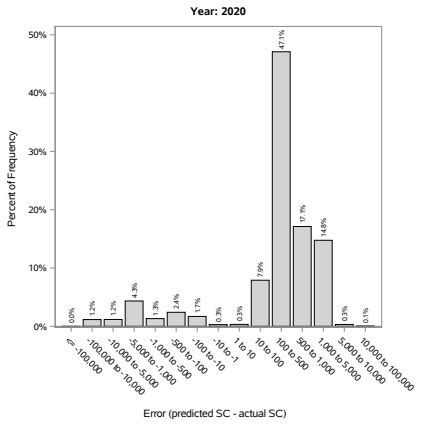
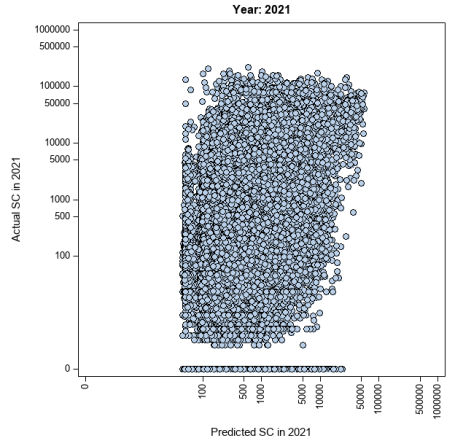

Year: 2023
Percent of Frequency
Error (predicted SC - actual SC)

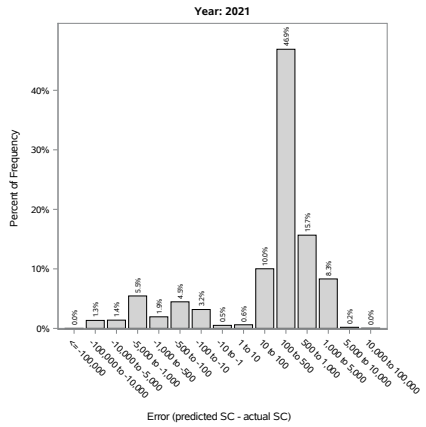
Figure S2. Scatter plots of actual SC vs. predicted SC (left-hand panels) and distributions of absolute errors (right-hand panels), in the whole population. Cost predictions derived from updated algorithms, including historical cost information in the predictors' set.

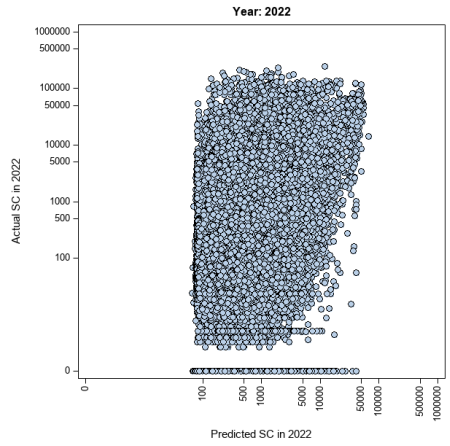


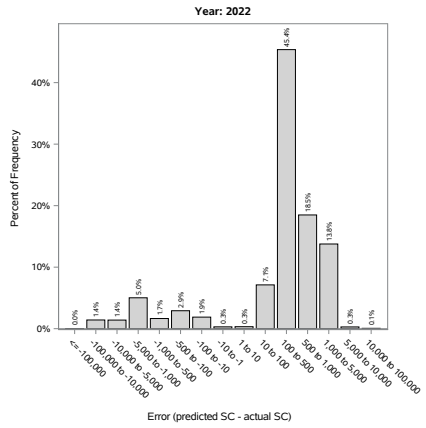

Year: 2020
Actual SC in 2020
Predicted SC in 2020

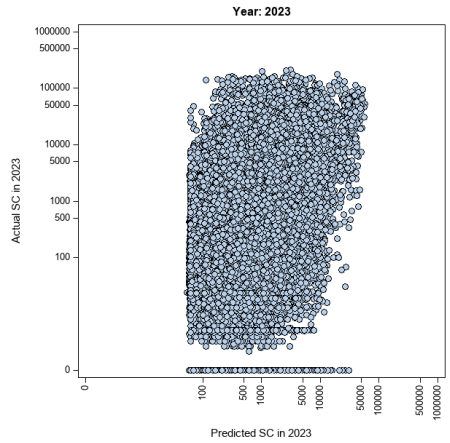

Year: 2020
Percent of Frequency
Error (predicted SC - actual SC)


Year: 2021
Actual SC in 2021
Predicted SC in 2021


Year: 2021
Percent of Frequency
Error (predicted SC - actual SC)


Year: 2022
Actual SC in 2022
Predicted SC in 2022


Year: 2022
Percent of Frequency
Error (predicted SC - actual SC)


Year: 2023
Actual SC in 2023
Predicted SC in 2023

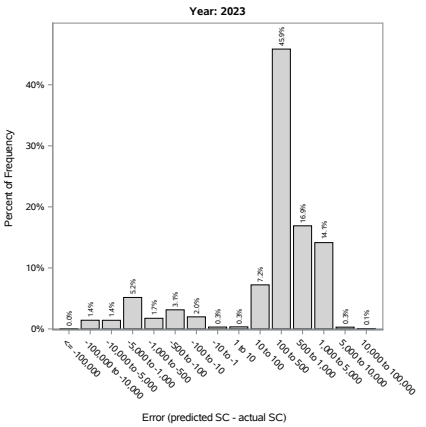

Year: 2023
Percent of Frequency
Error (predicted SC - actual SC)

Figure S3. Plot of variable importance scores for all training sets, considering TC as the outcome and excluding historical cost data from the set of predictors.

Figure S4. Plot of variable importance scores for all training sets, considering TC as the outcome and including historical cost data from the set of predictors.

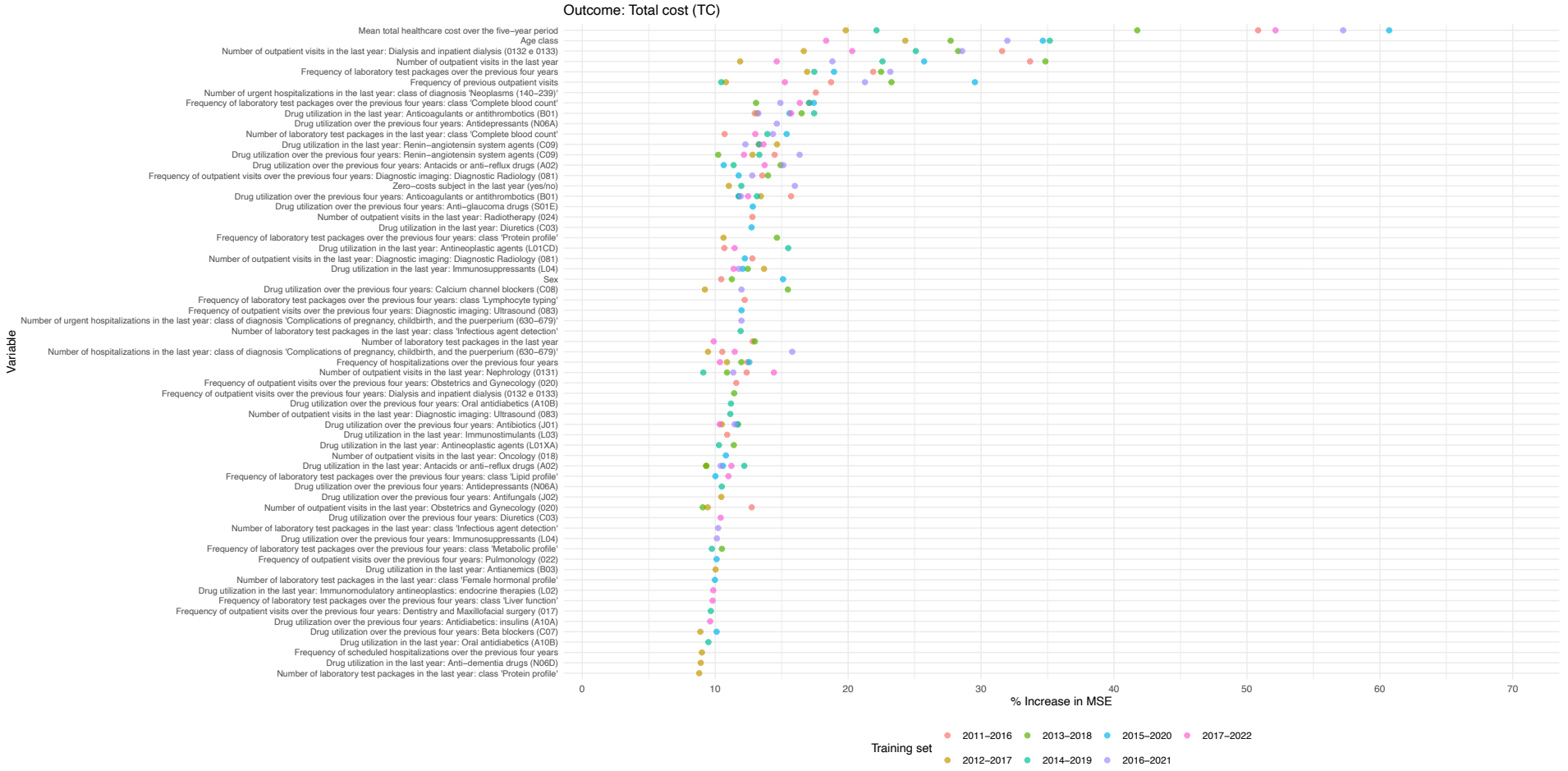


Figure S5. Plot of variable importance scores for all training sets, considering TSC as the outcome and excluding historical cost data from the set of predictors.

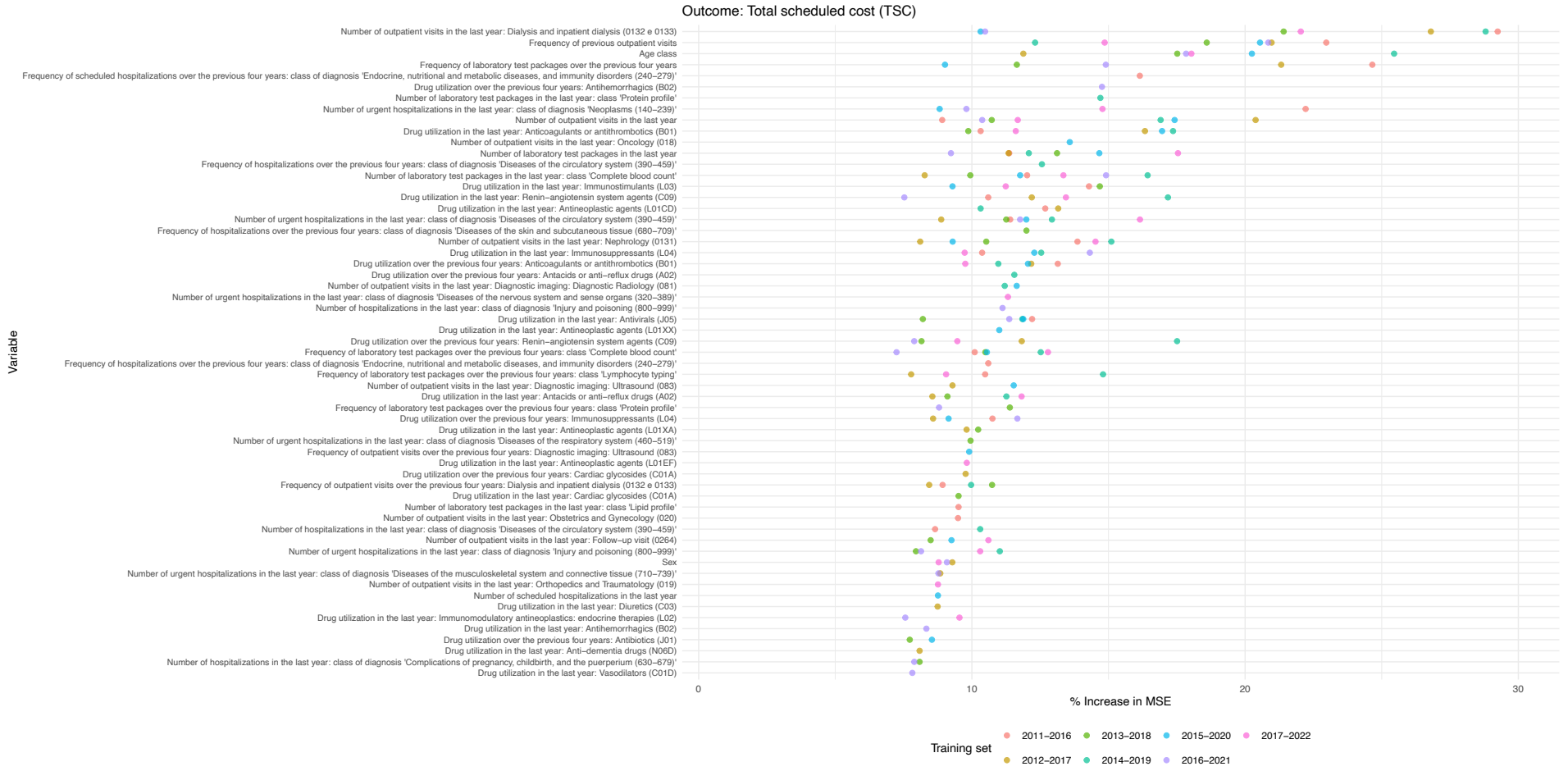


Figure S6. Plot of variable importance scores for all training sets, considering TSC as the outcome and including historical cost data from the set of predictors.

Figure S7. Plot of variable importance scores for all training sets, considering SC as the outcome and excluding historical cost data from the set of predictors.

Figure S8. Plot of variable importance scores for all training sets, considering SC as the outcome and including historical cost data from the set of predictors.

Figure S9. Plot of variable importance scores for all training sets, considering SSC as the outcome and excluding historical cost data from the set of predictors.

Figure S10. Plot of variable importance scores for all training sets, considering SSC as the outcome and including historical cost data from the set of predictors.

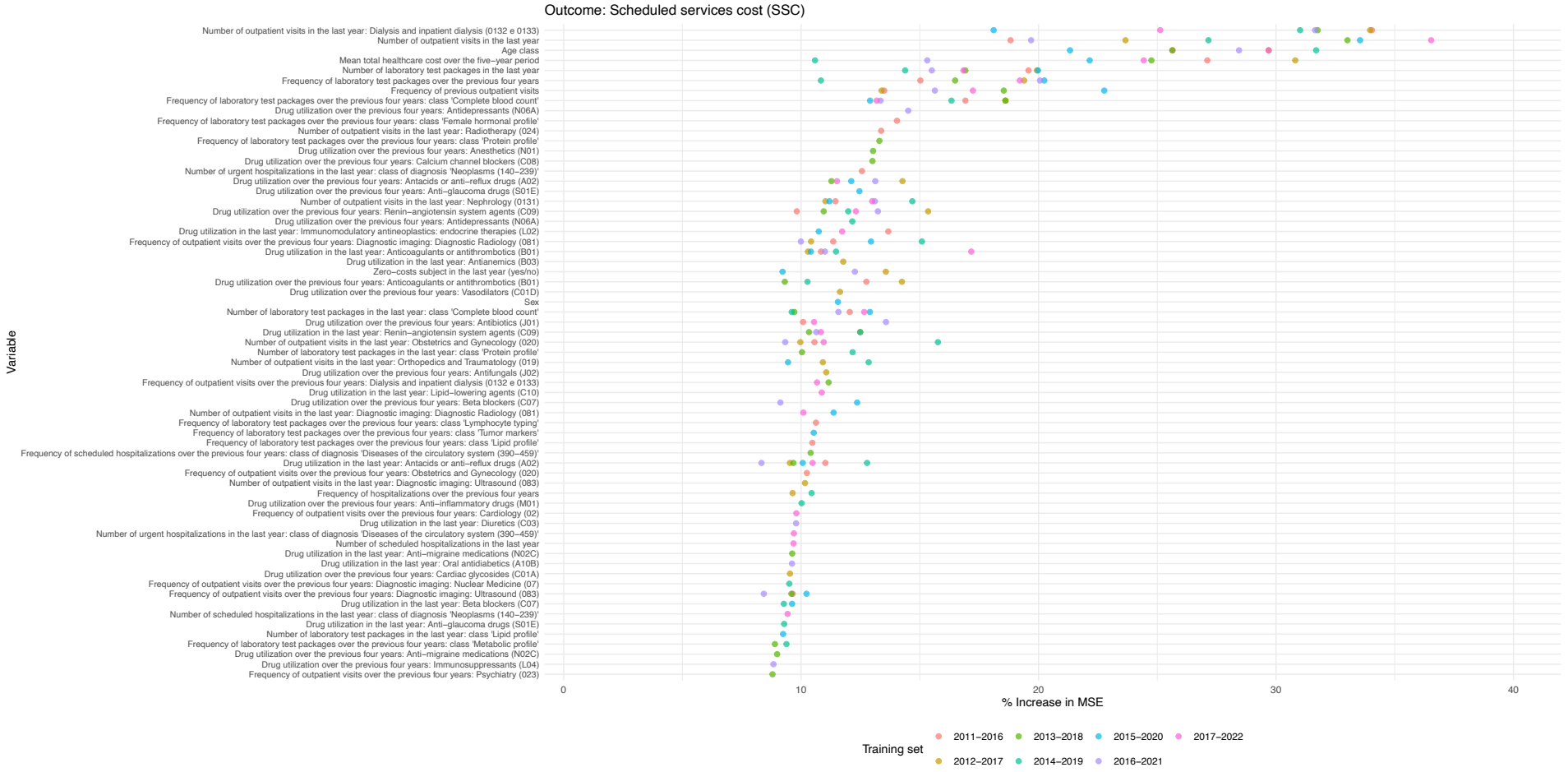


Figure S11. Prediction errors (PE) of TC, TSC, SSC, and SC, for the whole population. Cost predictions derived from non-updated algorithms, excluding historical cost information from the predictors' set.

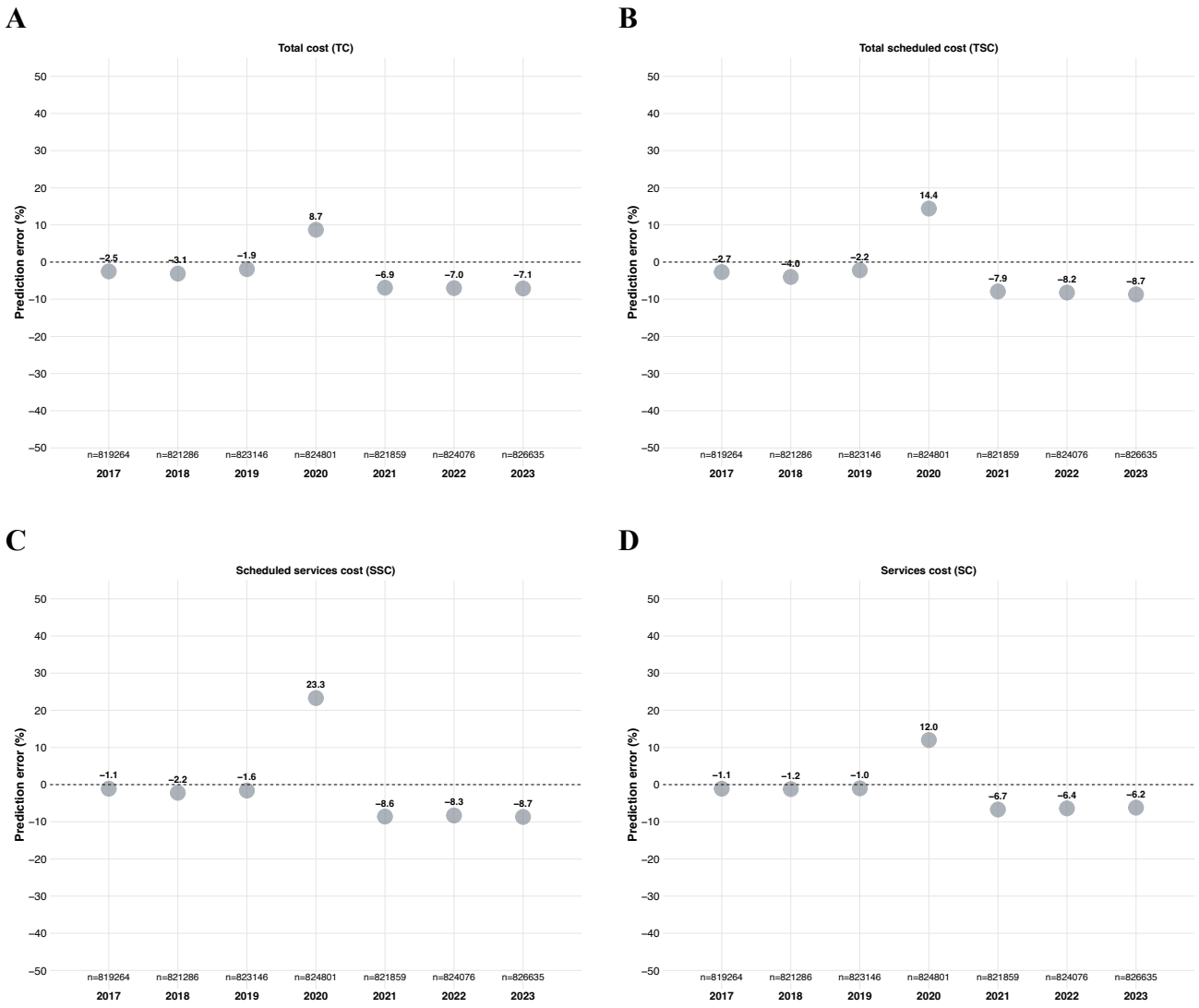


Figure S12. Prediction errors (PE) of TC, TSC, SSC, and SC, for the whole population. Cost predictions derived from non-updated algorithms, including historical cost information in the predictors' set.

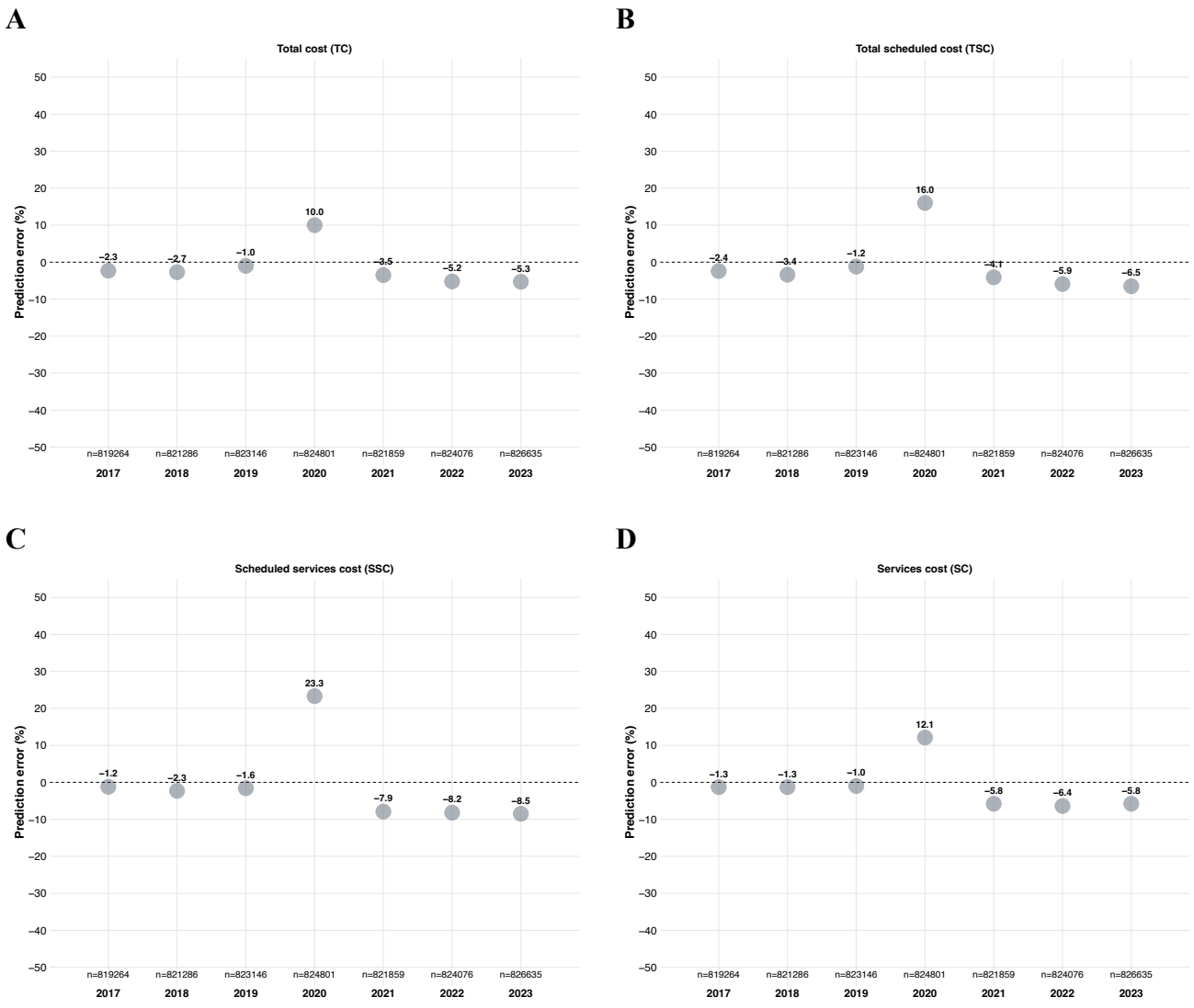
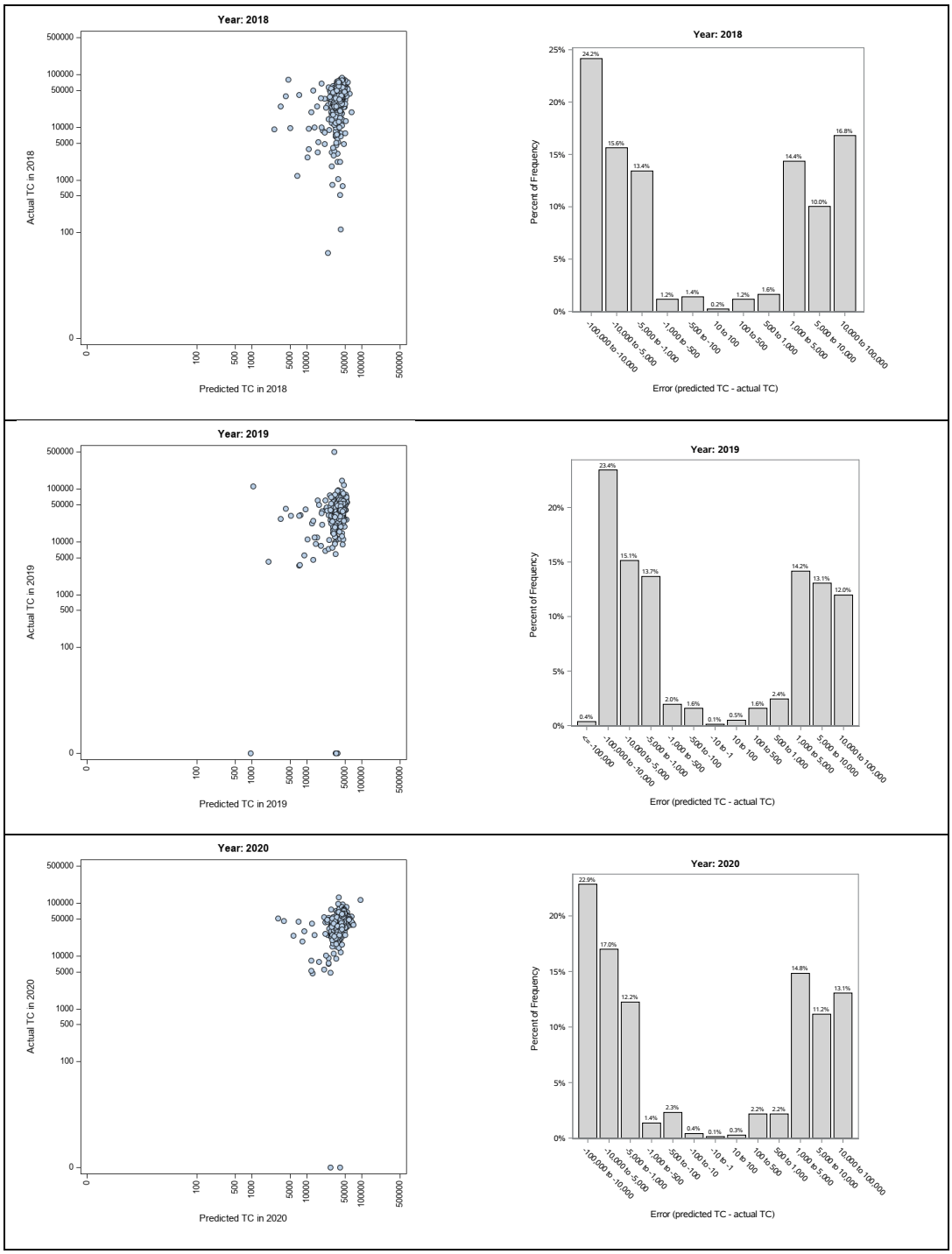


Figure S13. Scatter plots of actual TC vs. predicted TC (left-hand panels) and distributions of absolute errors (right-hand panels), in the CKD - Dialysis "all" patients' group. Cost predictions derived from updated algorithms, excluding historical cost information from the predictors' set.



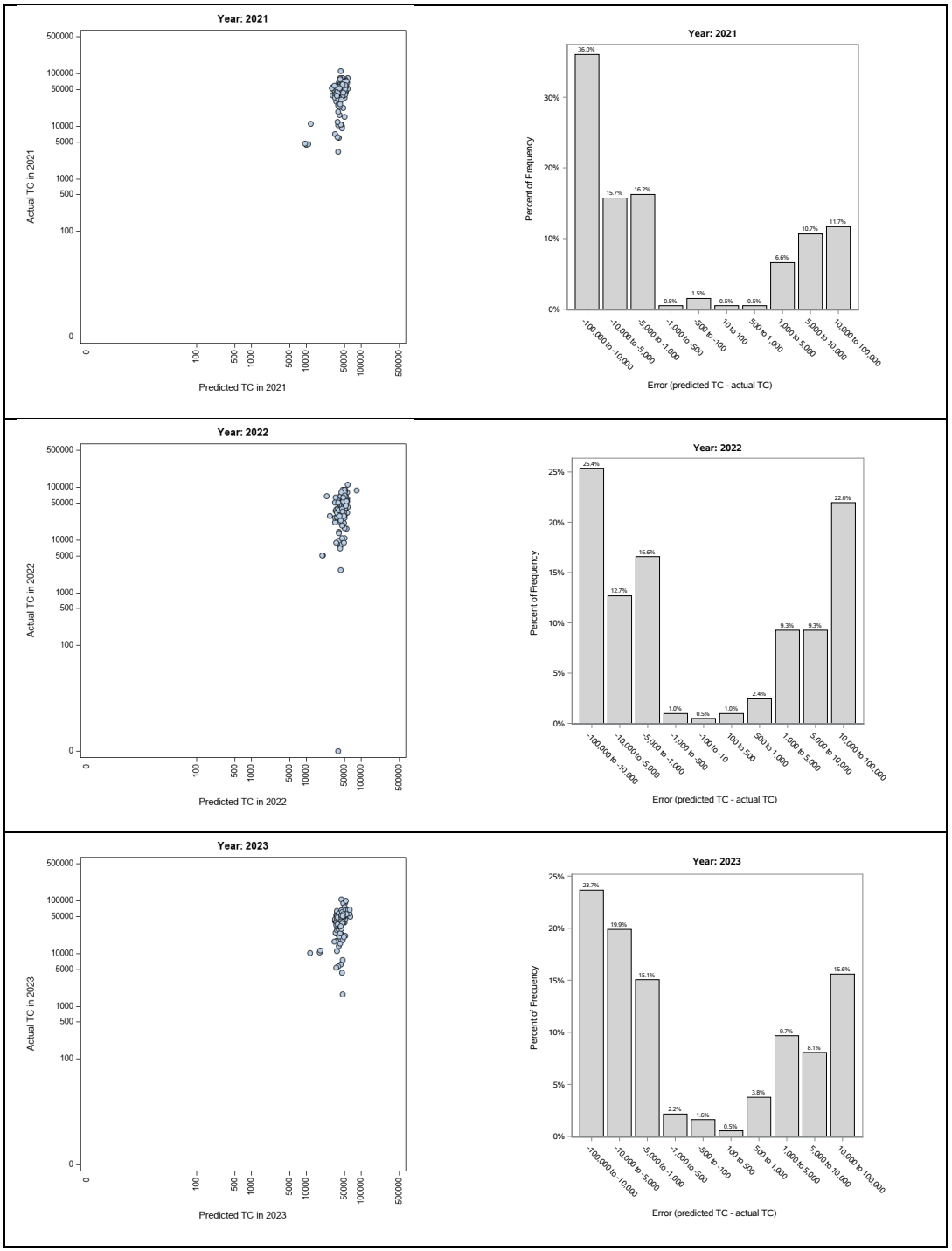

Year: 2021
Actual TC in 2021
Predicted TC in 2021
Year: 2021
Percent of Frequency
Error (predicted TC - actual TC)
-100,000 to -10,000
-10,000 to -5,000
-5,000 to -1,000
-1,000 to -500
-500 to -100
10 to 100
500 to 1,000
1,000 to 5,000
5,000 to 10,000
10,000 to 100,000
Year: 2022
Actual TC in 2022
Predicted TC in 2022
Year: 2022
Percent of Frequency
Error (predicted TC - actual TC)
-100,000 to -10,000
-10,000 to -5,000
-5,000 to -1,000
-1,000 to -500
-100 to -10
100 to 500
500 to 1,000
1,000 to 5,000
5,000 to 10,000
10,000 to 100,000
Year: 2023
Actual TC in 2023
Predicted TC in 2023
Year: 2023
Percent of Frequency
Error (predicted TC - actual TC)
-100,000 to -10,000
-10,000 to -5,000
-5,000 to -1,000
-1,000 to -500
-500 to -100
100 to 500
500 to 1,000
1,000 to 5,000
5,000 to 10,000
10,000 to 100,000

Figure S14. Prediction errors (PE) of TC for the CKD - Dialysis patients' groups. Cost predictions derived from updated algorithms, including historical cost information in the predictors' set.

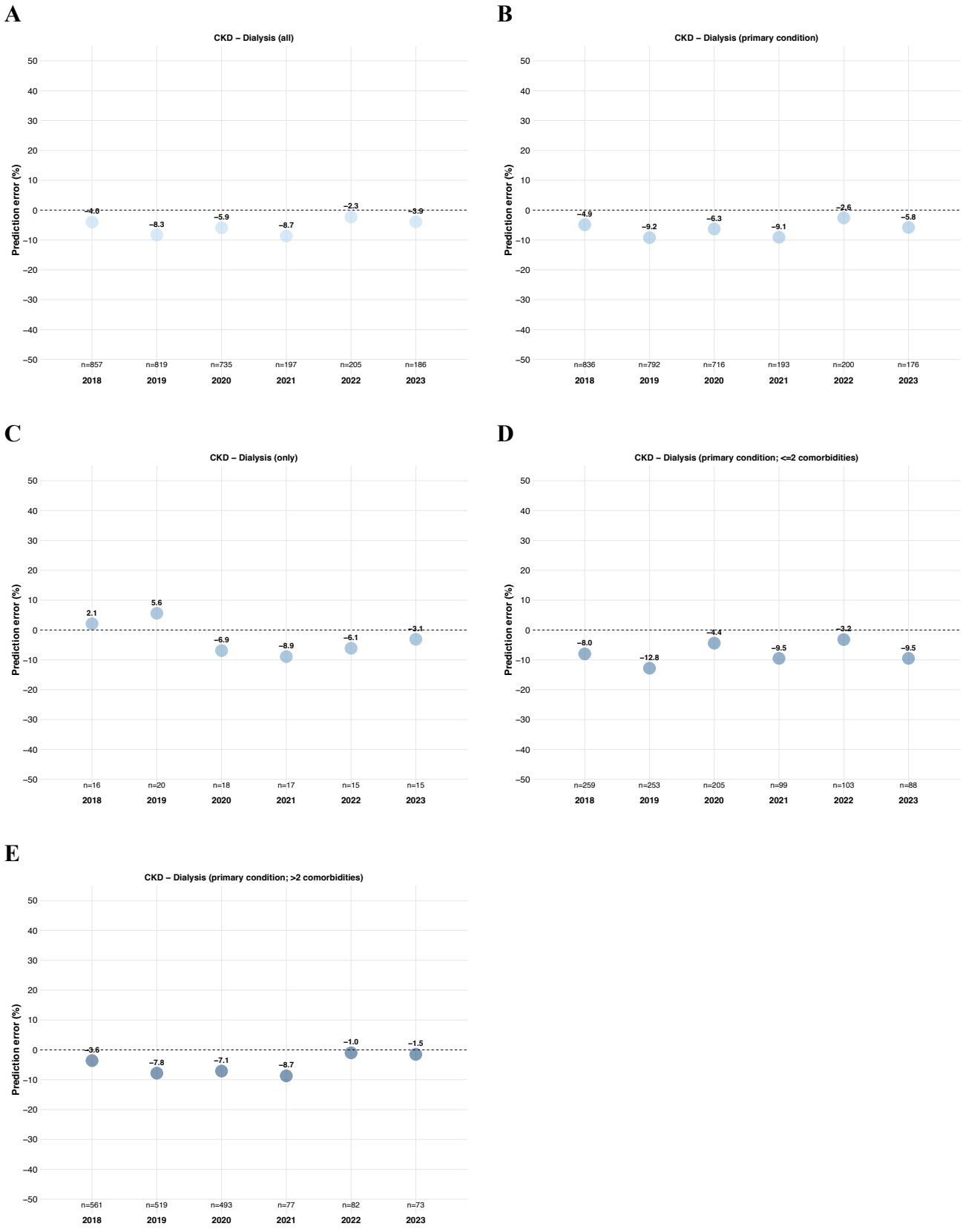


Figure S15. Annual predicted and actual mean TC for CKD - Dialysis patients' groups. Cost predictions derived from updated algorithms, including historical cost information in the predictors' set.

Figure S16. Prediction errors (PE) of TSC for the CKD - Dialysis patients' groups. Cost predictions derived from updated algorithms, including historical cost information in the predictors' set.

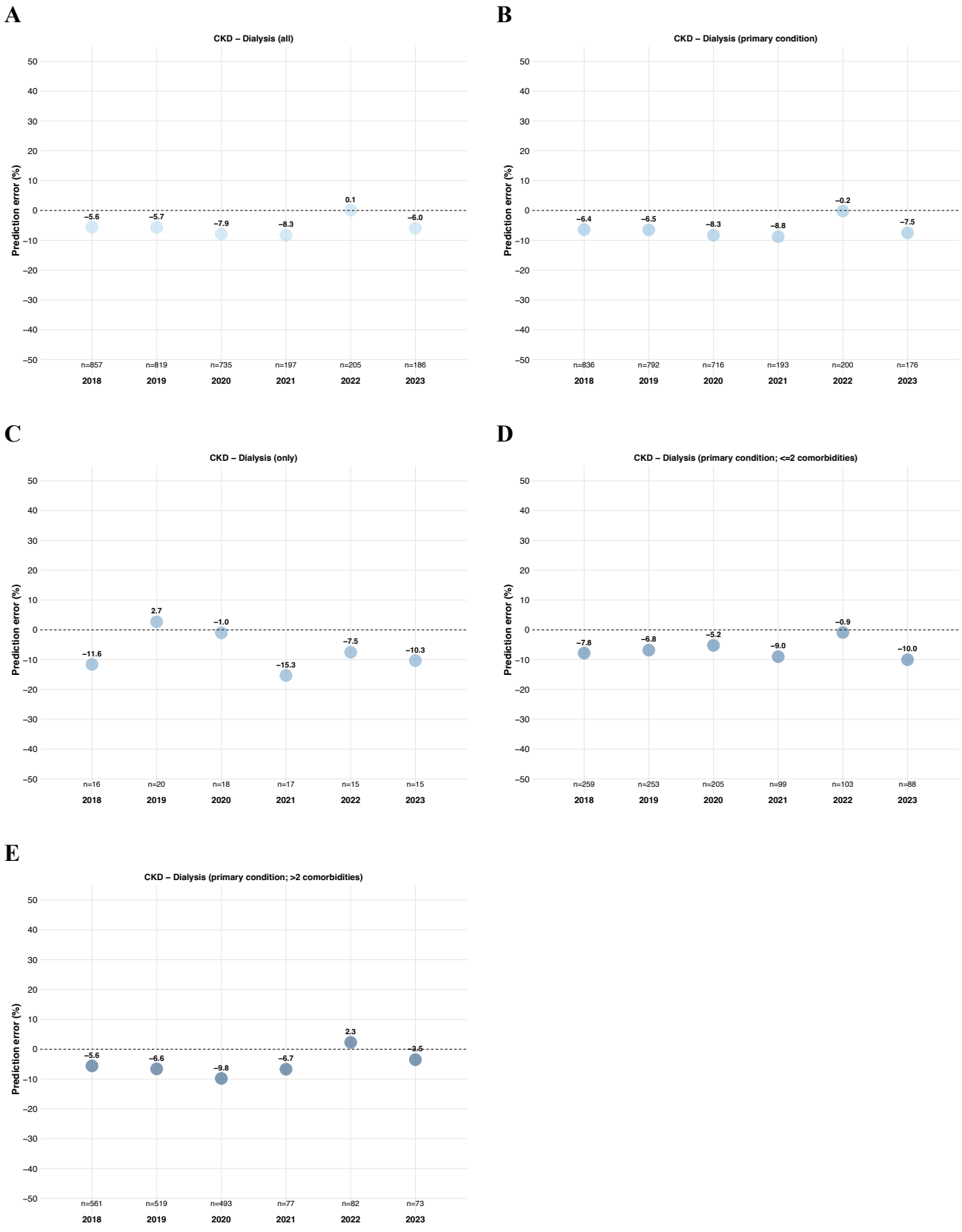


Figure S17. Annual predicted and actual mean TSC for CKD - Dialysis patients' groups. Cost predictions derived from updated algorithms, including historical cost information in the predictors' set.

Figure S18. Prediction errors (PE) of SSC for the CKD - Dialysis patients' groups. Cost predictions derived from updated algorithms, including historical cost information in the predictors' set.

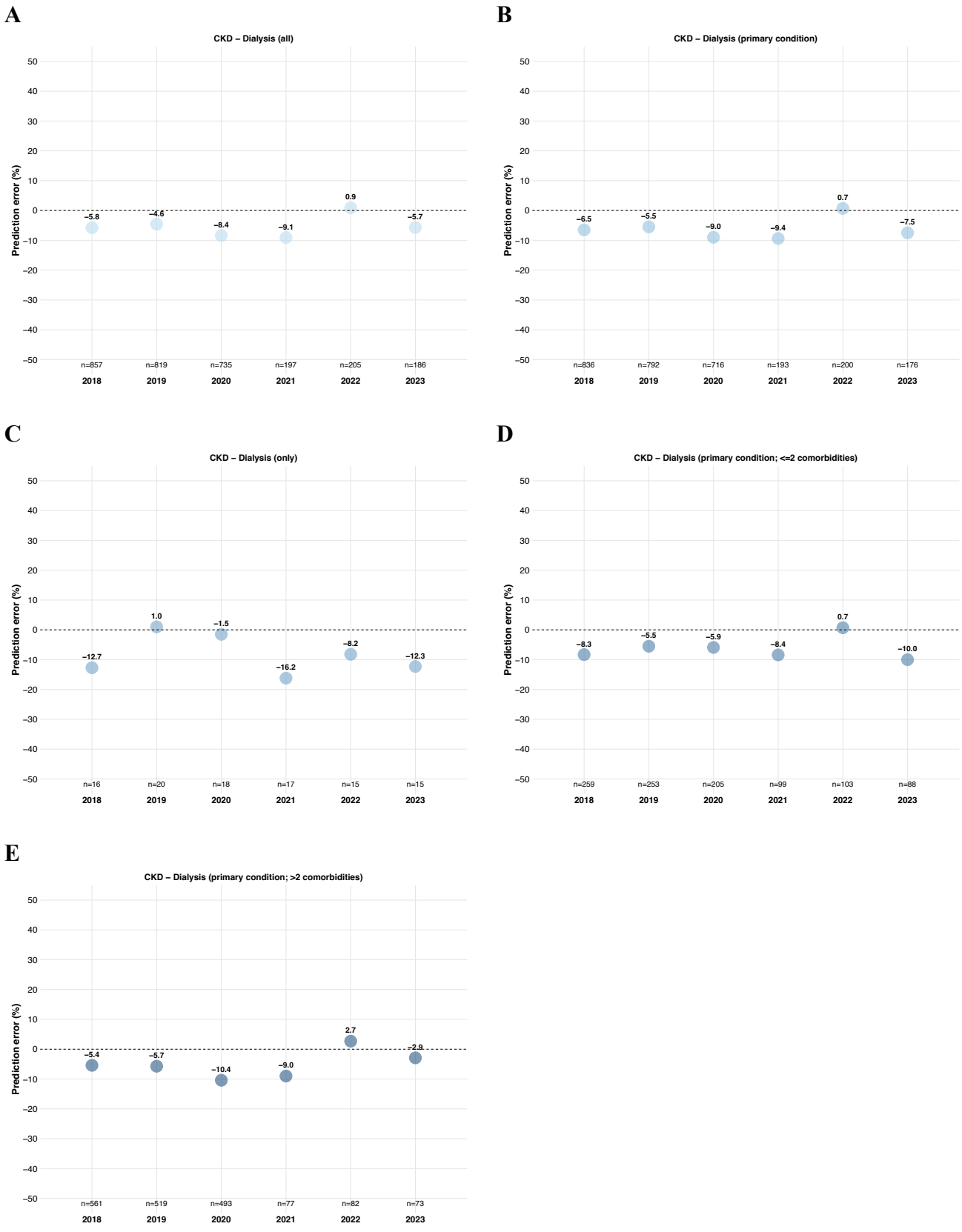


Figure S19. Annual predicted and actual mean SSC for CKD - Dialysis patients' groups. Cost predictions derived from updated algorithms, including historical cost information in the predictors' set.

Figure S20. Prediction errors (PE) of SC for the CKD - Dialysis patients' groups. Cost predictions derived from updated algorithms, including historical cost information in the predictors' set.

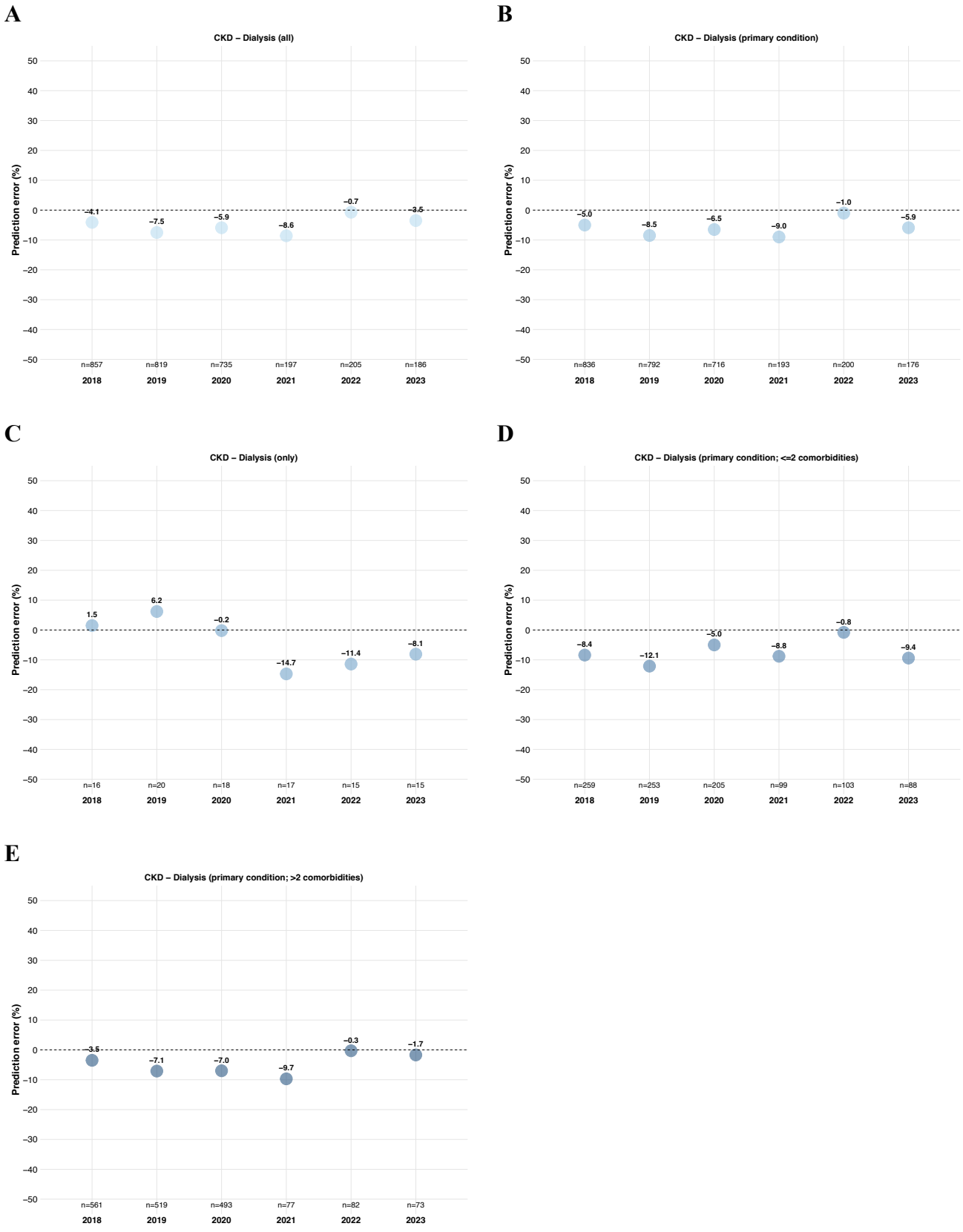


Figure S21. Annual predicted and actual mean SC for CKD - Dialysis patients' groups. Cost predictions derived from updated algorithms, including historical cost information in the predictors' set.

Figure S22. Prediction errors (PE) of TC, TSC, SSC, and SC, for the CKD - Dialysis "all" patients' group. Cost predictions derived from non-updated algorithms, excluding historical cost information from the predictors' set.

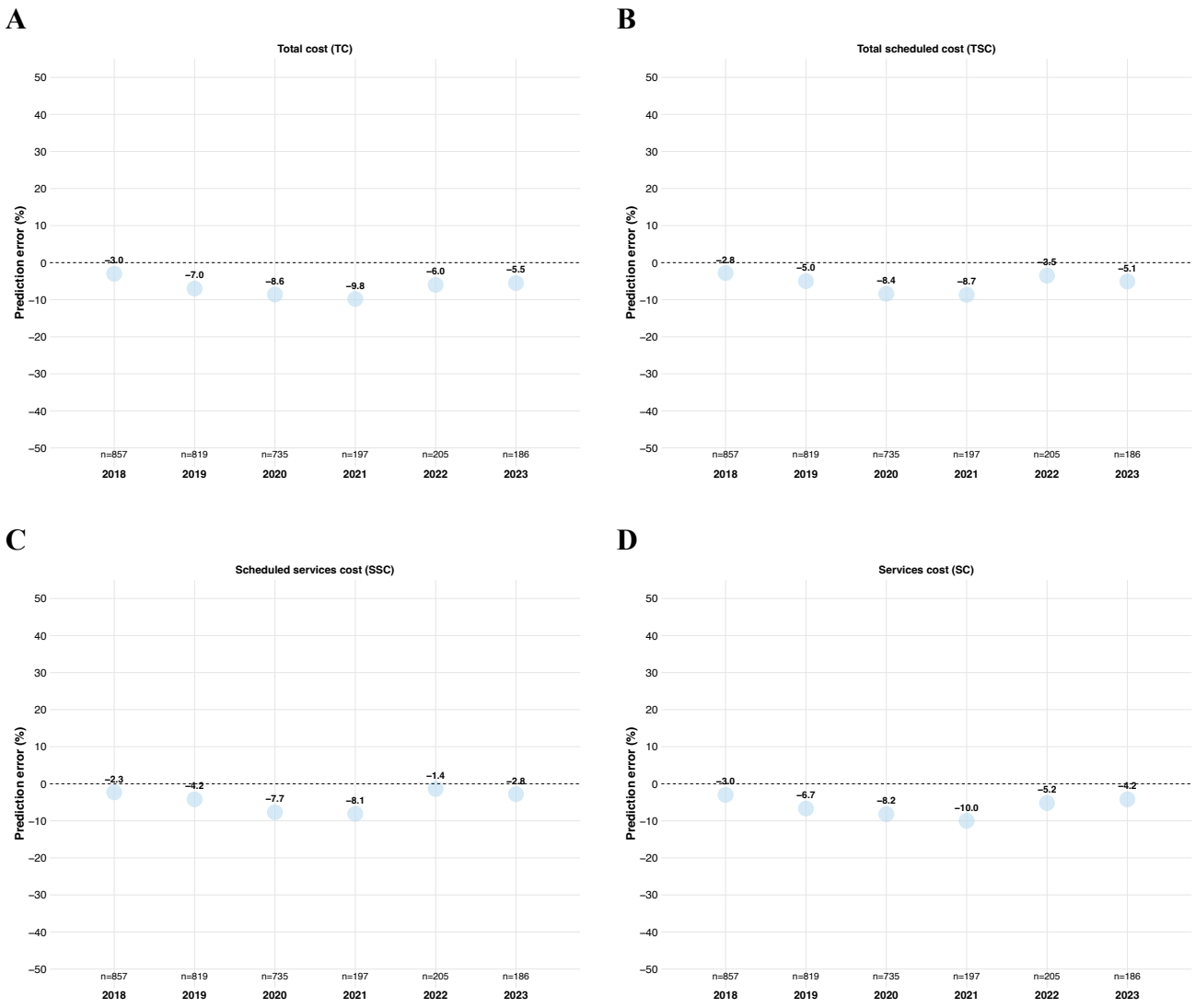


Figure S23. Prediction errors (PE) of TC, TSC, SSC, and SC, for the CKD - Dialysis "all" patients' group. Cost predictions derived from non-updated algorithms, including historical cost information in the predictors' set.

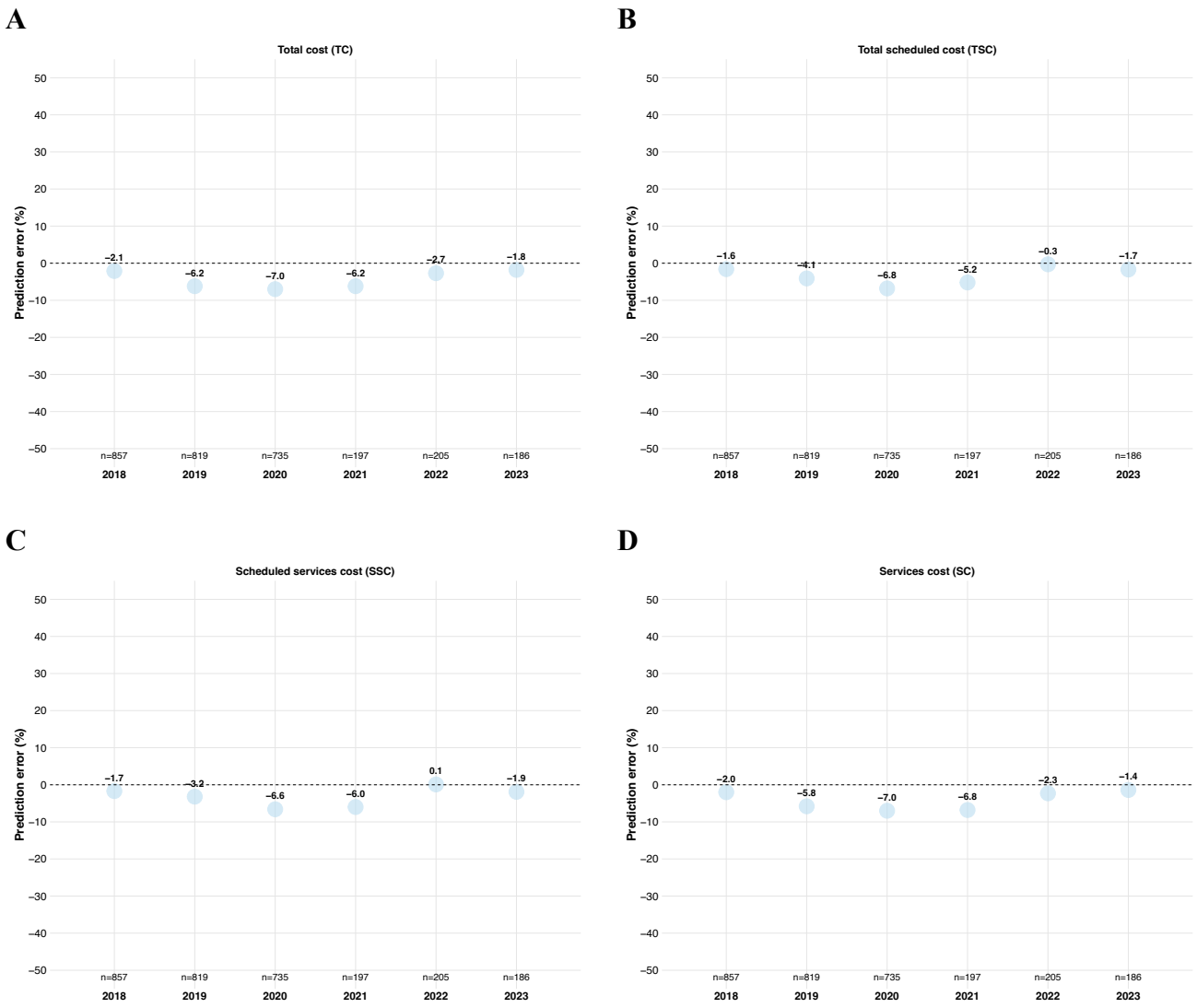
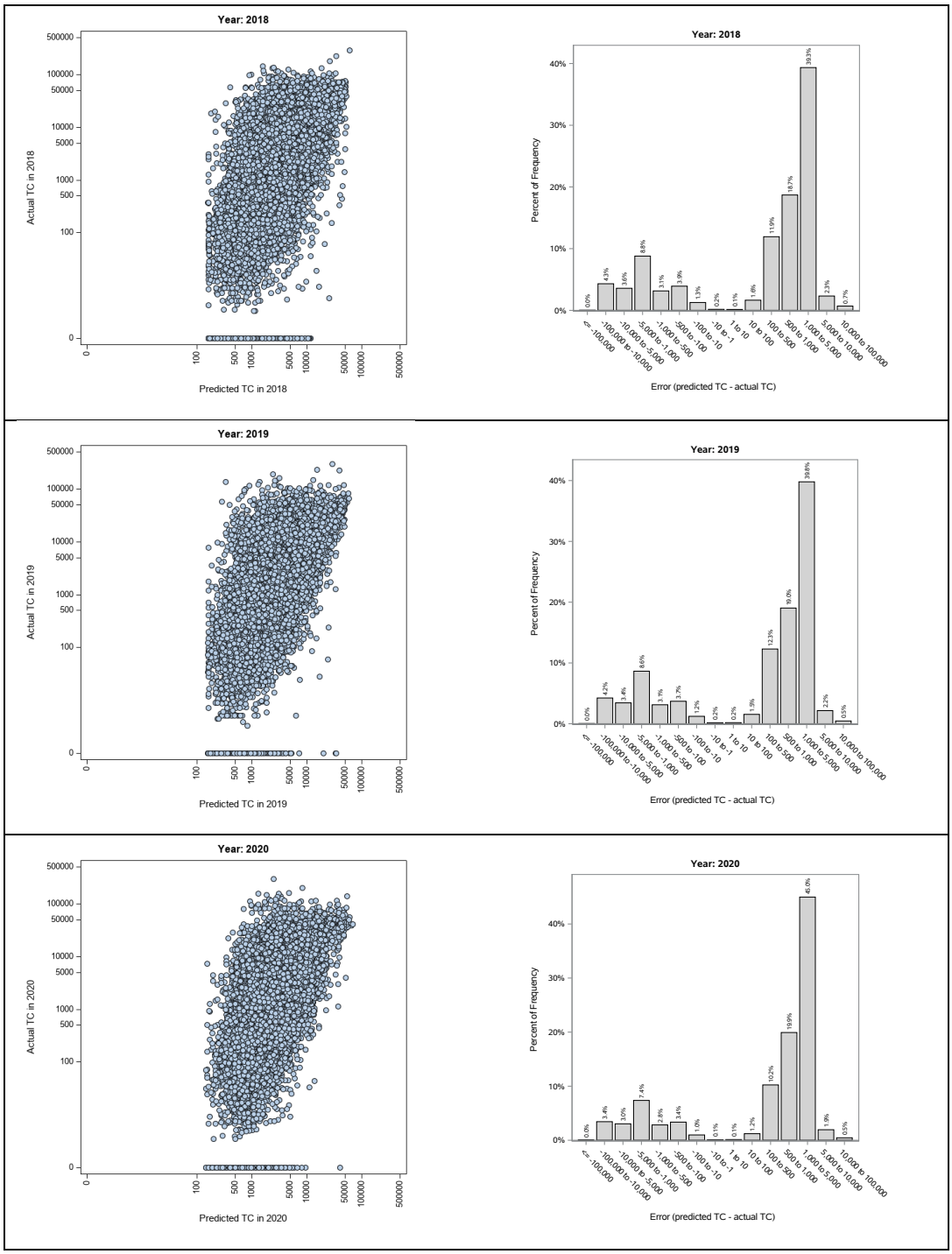


Figure S24. Scatter plots of actual TC vs. predicted TC (left-hand panels) and distributions of absolute errors (right-hand panels), in the Type 2 diabetes "all" patients' group. Cost predictions derived from updated algorithms, excluding historical cost information from the predictors' set.



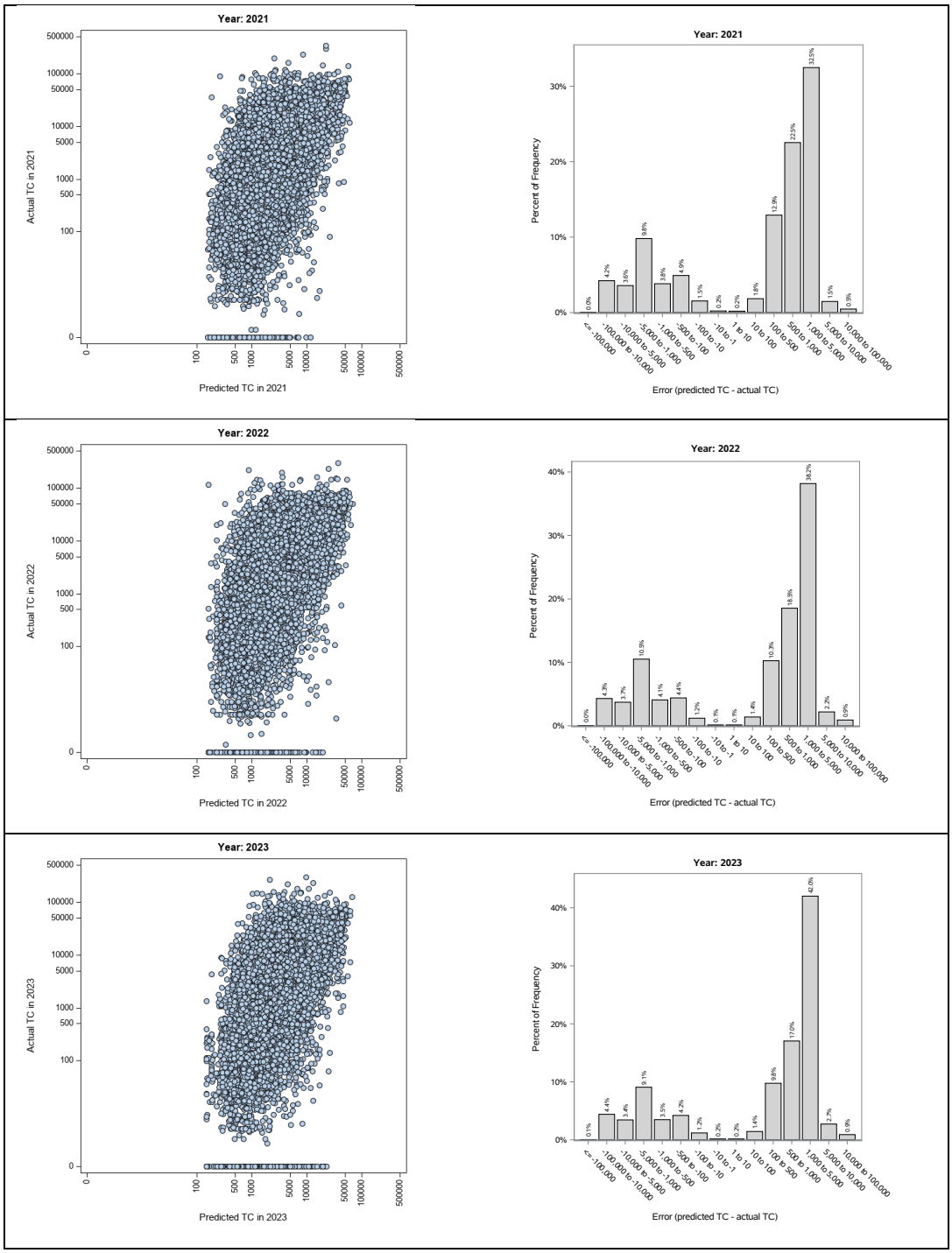

Year: 2021
Actual TC in 2021
Predicted TC in 2021
Year: 2021
Percent of Frequency
Error (predicted TC - actual TC)
Year: 2022
Actual TC in 2022
Predicted TC in 2022
Year: 2022
Percent of Frequency
Error (predicted TC - actual TC)
Year: 2023
Actual TC in 2023
Predicted TC in 2023
Year: 2023
Percent of Frequency
Error (predicted TC - actual TC)

Figure S25. Prediction errors (PE) of TC for the Type 2 diabetes patients' groups. Cost predictions derived from updated algorithms, including historical cost information in the predictors' set.

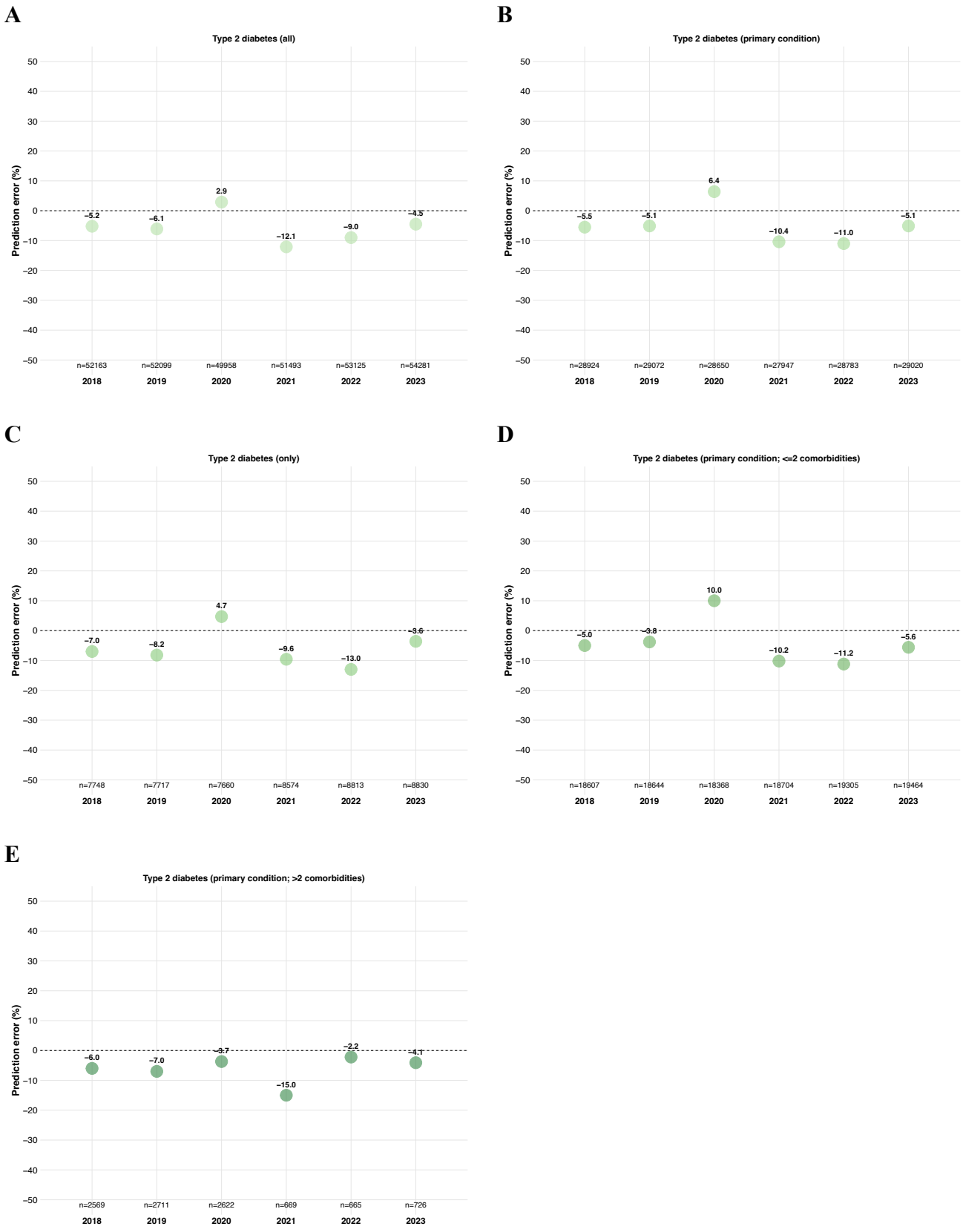


Figure S26. Annual predicted and actual mean TC for Type 2 diabetes patients' groups. Cost predictions derived from updated algorithms, including historical cost information in the predictors' set.

Figure S27. Prediction errors (PE) of TSC for the Type 2 diabetes patients' groups. Cost predictions derived from updated algorithms, including historical cost information in the predictors' set.

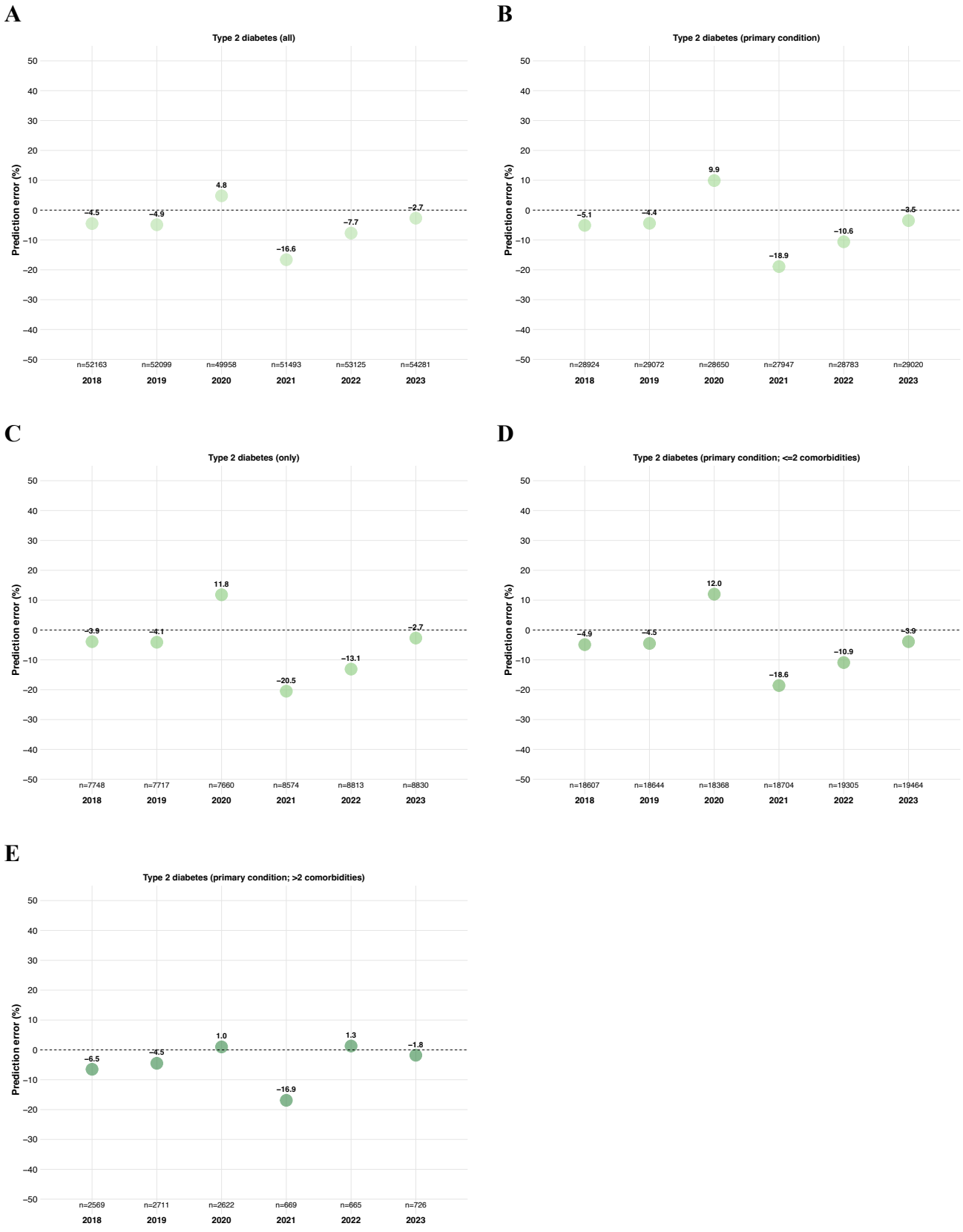


Figure S28. Annual predicted and actual mean TSC for Type 2 diabetes patients' groups. Cost predictions derived from updated algorithms, including historical cost information in the predictors' set.

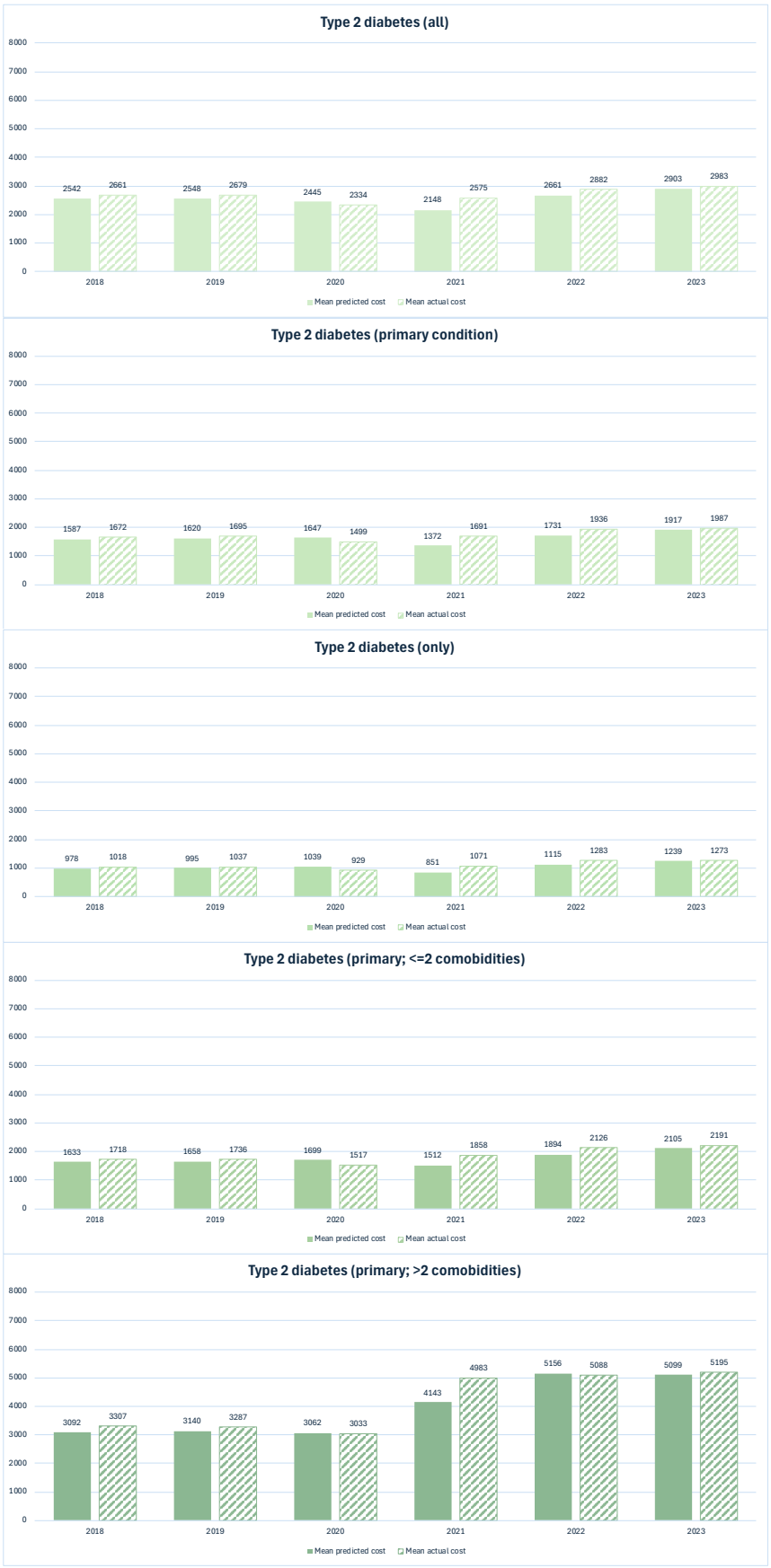


Figure S29. Prediction errors (PE) of SSC for the Type 2 diabetes patients' groups. Cost predictions derived from updated algorithms, including historical cost information in the predictors' set.

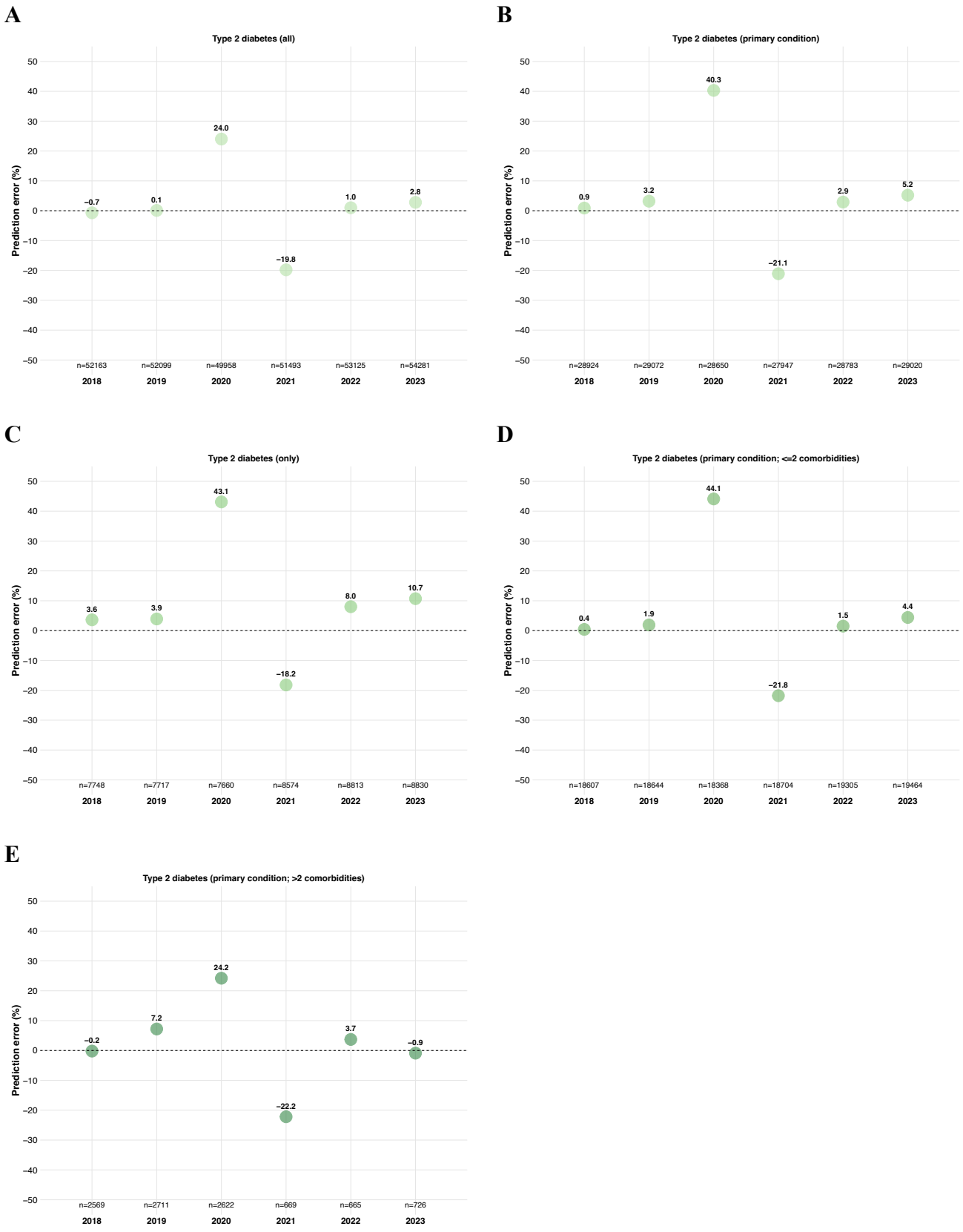


Figure S30. Annual predicted and actual mean SSC for Type 2 diabetes patients' groups. Cost predictions derived from updated algorithms, including historical cost information in the predictors' set.

Figure S31. Prediction errors (PE) of SC for the Type 2 diabetes patients' groups. Cost predictions derived from updated algorithms, including historical cost information in the predictors' set.

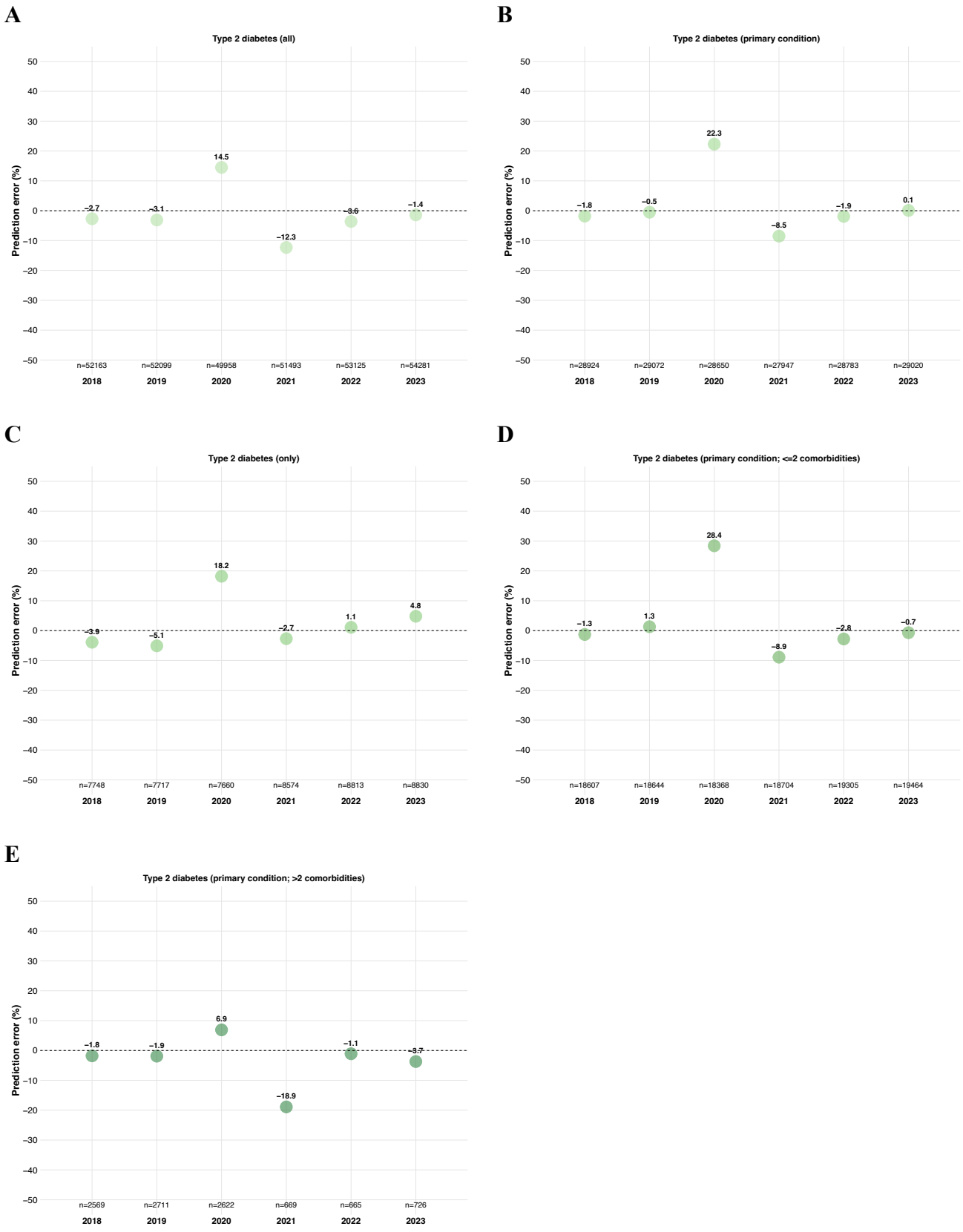


Figure S32. Annual predicted and actual mean SC for Type 2 diabetes patients' groups. Cost predictions derived from updated algorithms, including historical cost information in the predictors' set.

Figure S33. Prediction errors (PE) of TC, TSC, SSC, and SC, for the Type 2 diabetes "all" patients' group. Cost predictions derived from non-updated algorithms, excluding historical cost information from the predictors' set.

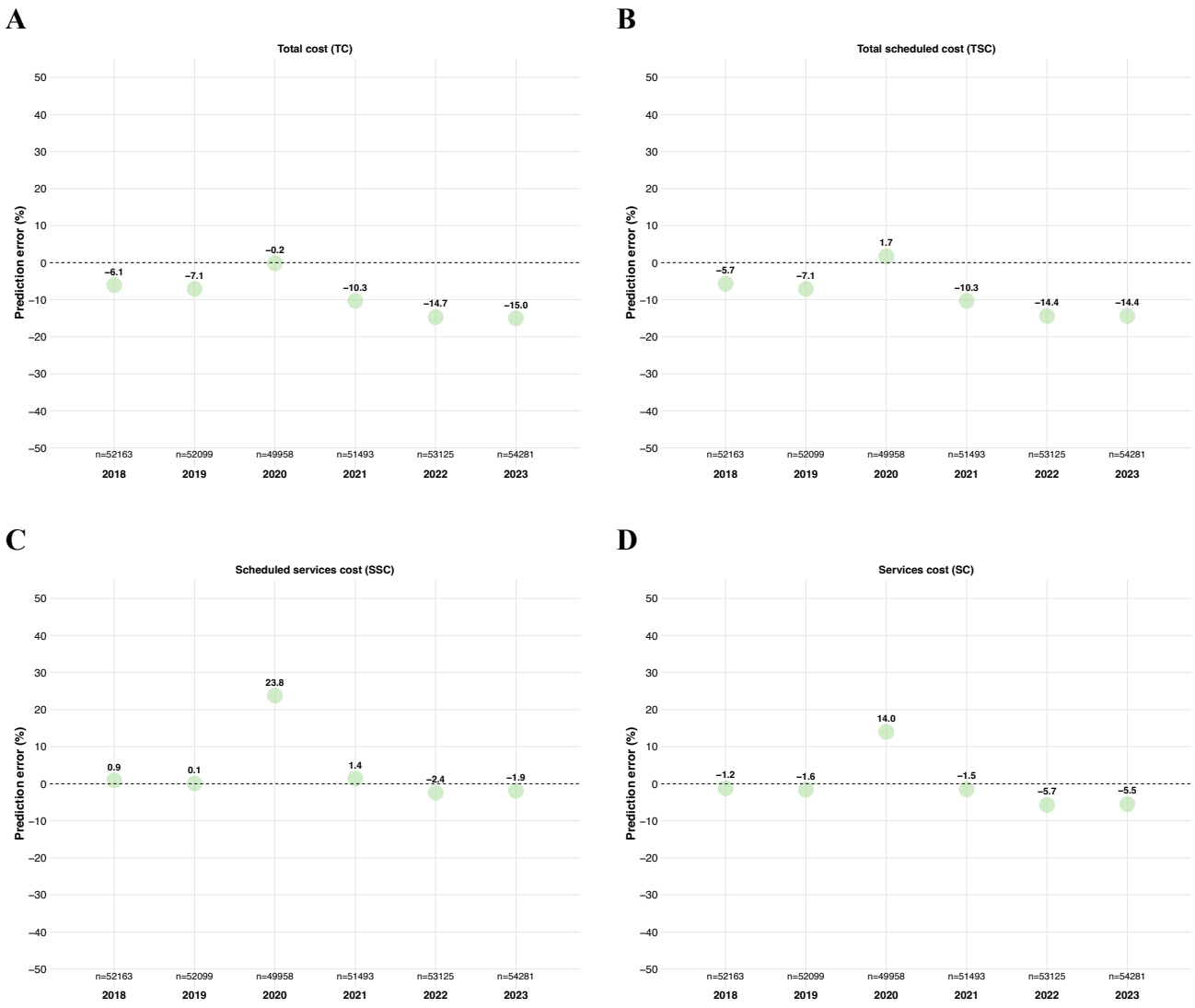


Figure S34. Prediction errors (PE) of TC, TSC, SSC, and SC, for the Type 2 diabetes "all" patients' group. Cost predictions derived from non-updated algorithms, including historical cost information in the predictors' set.

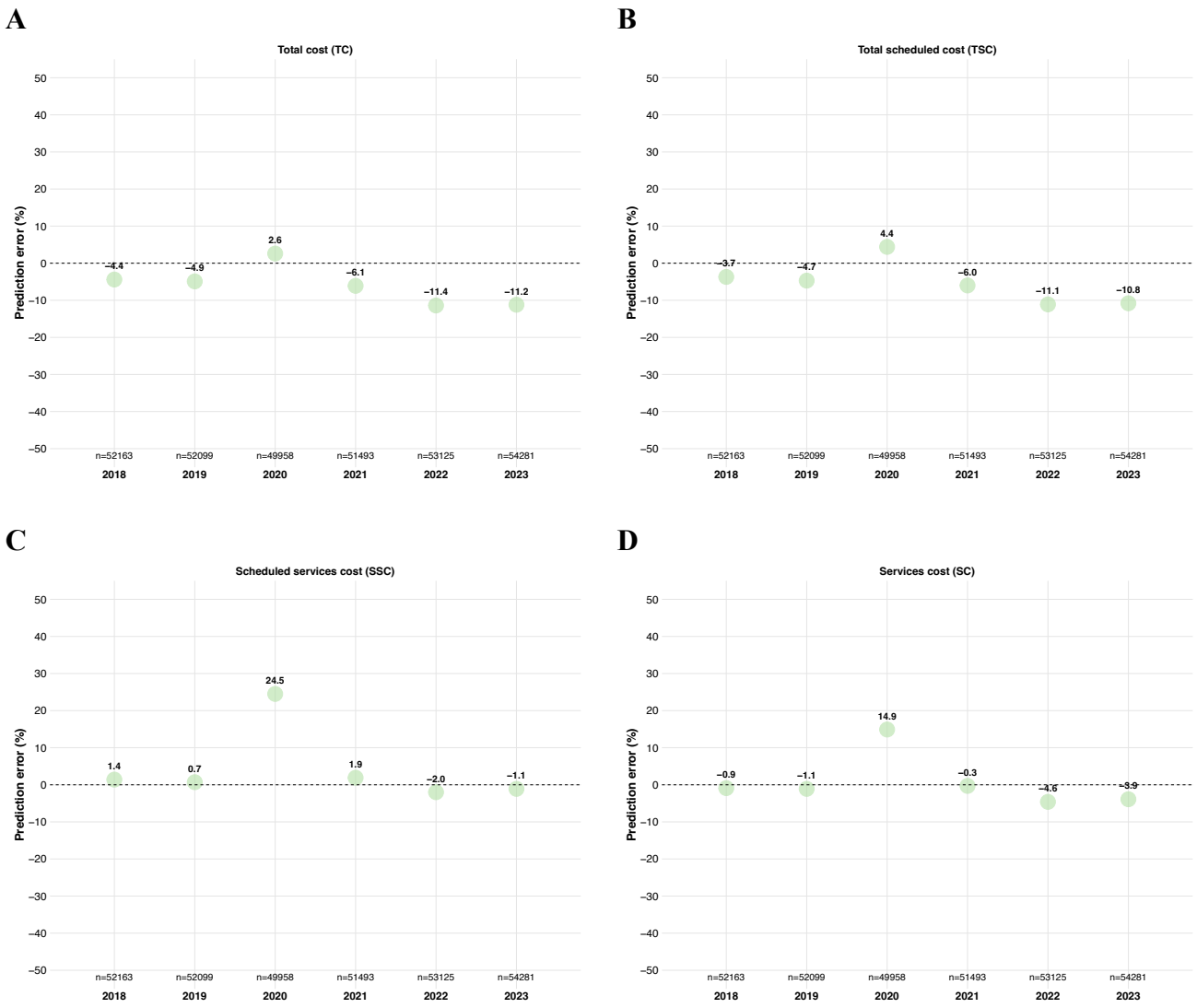
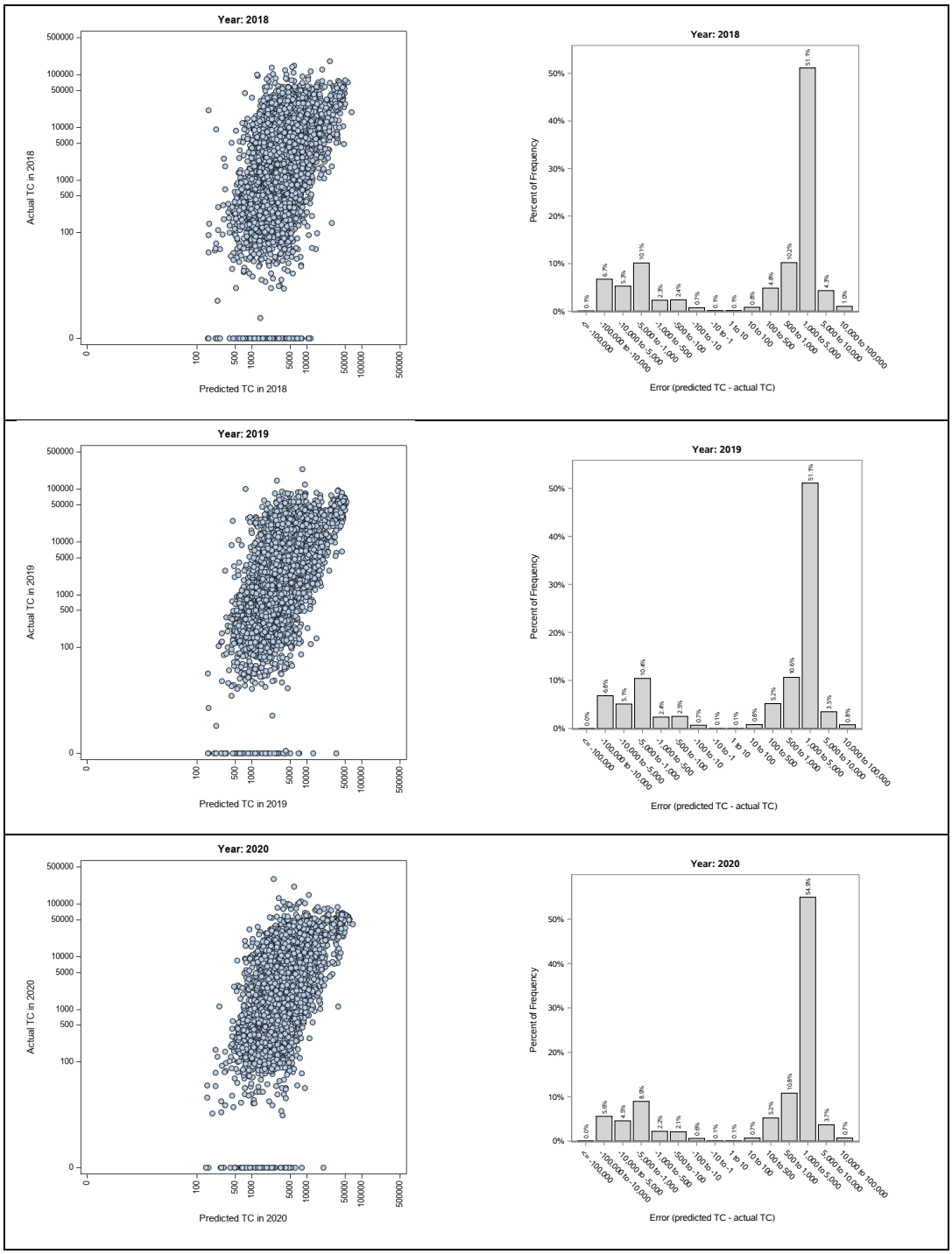


Figure S35. Scatter plots of actual TC vs. predicted TC (left-hand panels) and distributions of absolute errors (right-hand panels), in the Heart failure "all" patients' group. Cost predictions derived from updated algorithms, excluding historical cost information from the predictors' set.



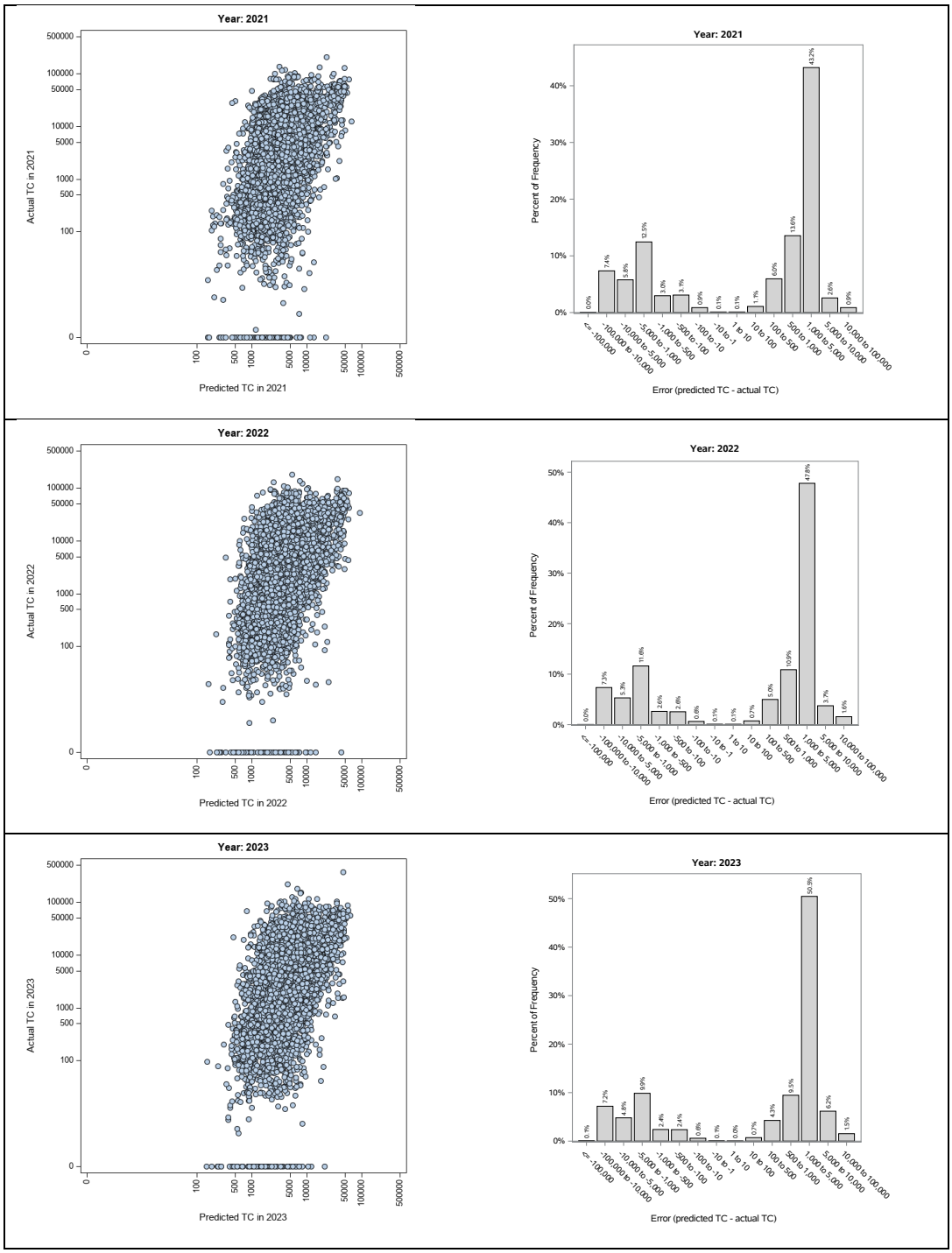

Year: 2021
Actual TC in 2021
Predicted TC in 2021
Year: 2021
Percent of Frequency
Error (predicted TC - actual TC)
Year: 2022
Actual TC in 2022
Predicted TC in 2022
Year: 2022
Percent of Frequency
Error (predicted TC - actual TC)
Year: 2023
Actual TC in 2023
Predicted TC in 2023
Year: 2023
Percent of Frequency
Error (predicted TC - actual TC)

Figure S36. Prediction errors (PE) of TC for the Heart failure patients' groups. Cost predictions derived from updated algorithms, including historical cost information in the predictors' set.

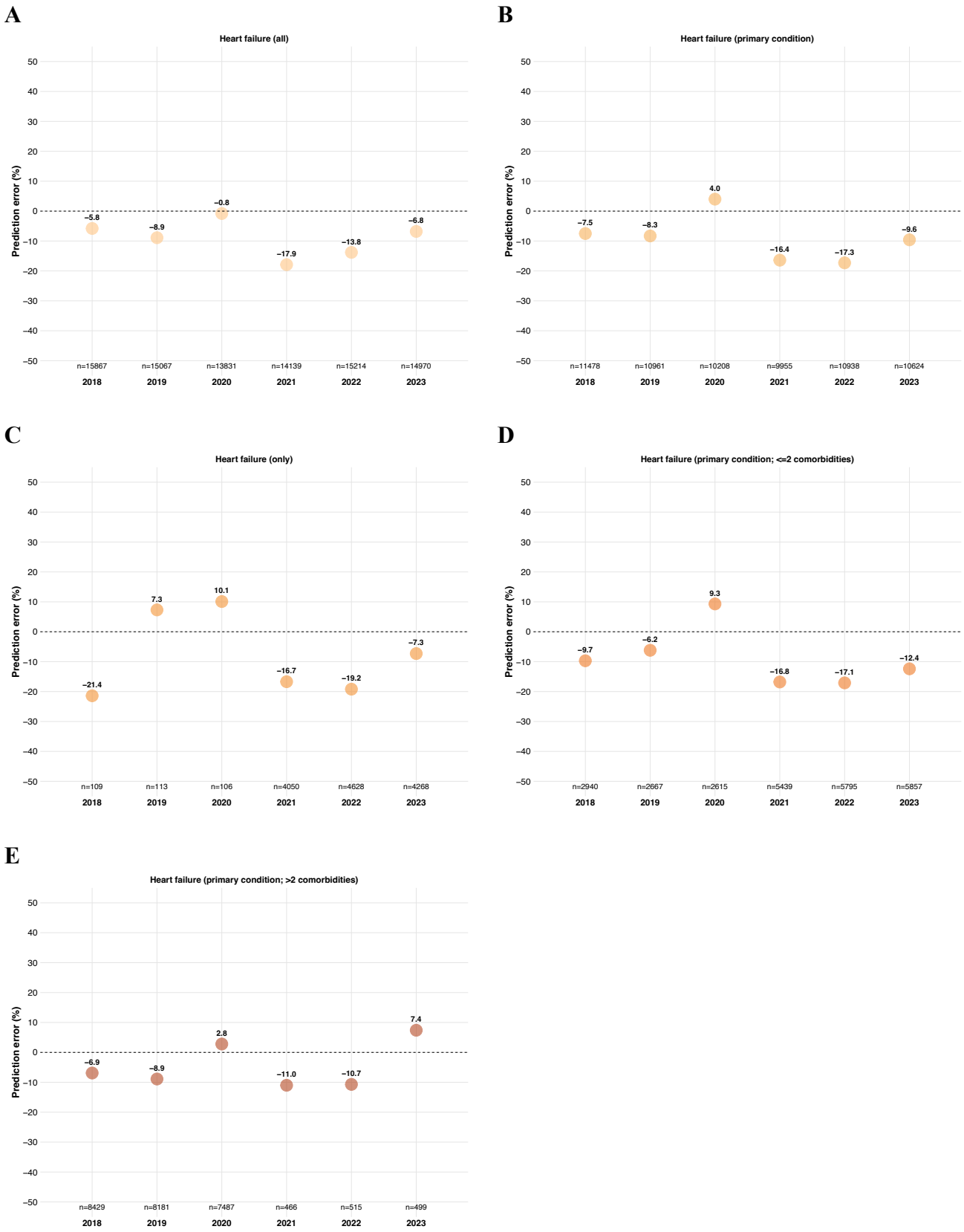


Figure S37. Annual predicted and actual mean TC for Heart failure patients' groups. Cost predictions derived from updated algorithms, including historical cost information in the predictors' set.

Figure S38. Prediction errors (PE) of TSC for the Heart failure patients' groups. Cost predictions derived from updated algorithms, including historical cost information in the predictors' set.

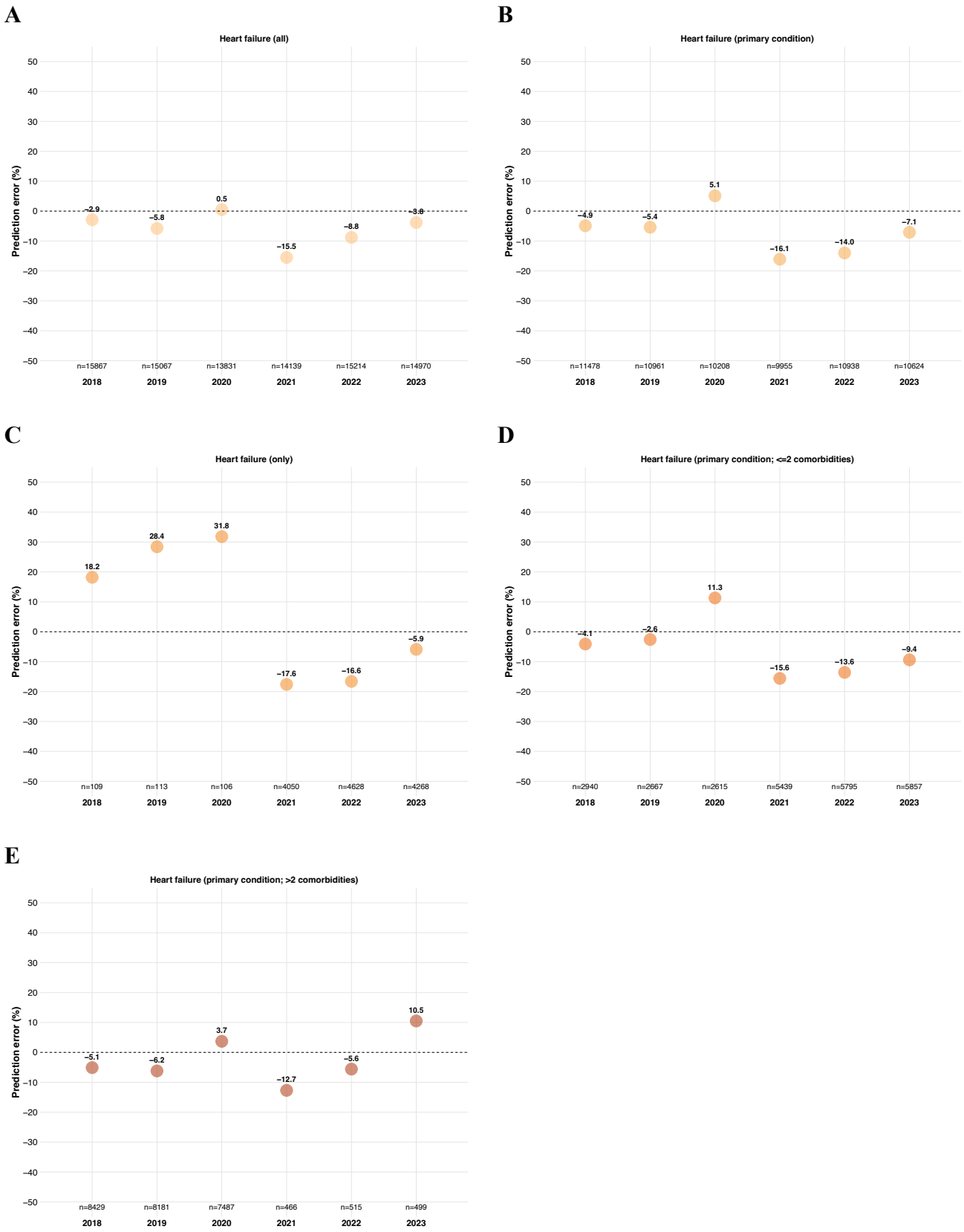


Figure S39. Annual predicted and actual mean TSC for Heart failure patients' groups. Cost predictions derived from updated algorithms, including historical cost information in the predictors' set.

Figure S40. Prediction errors (PE) of SSC for the Heart failure patients' groups. Cost predictions derived from updated algorithms, including historical cost information in the predictors' set.

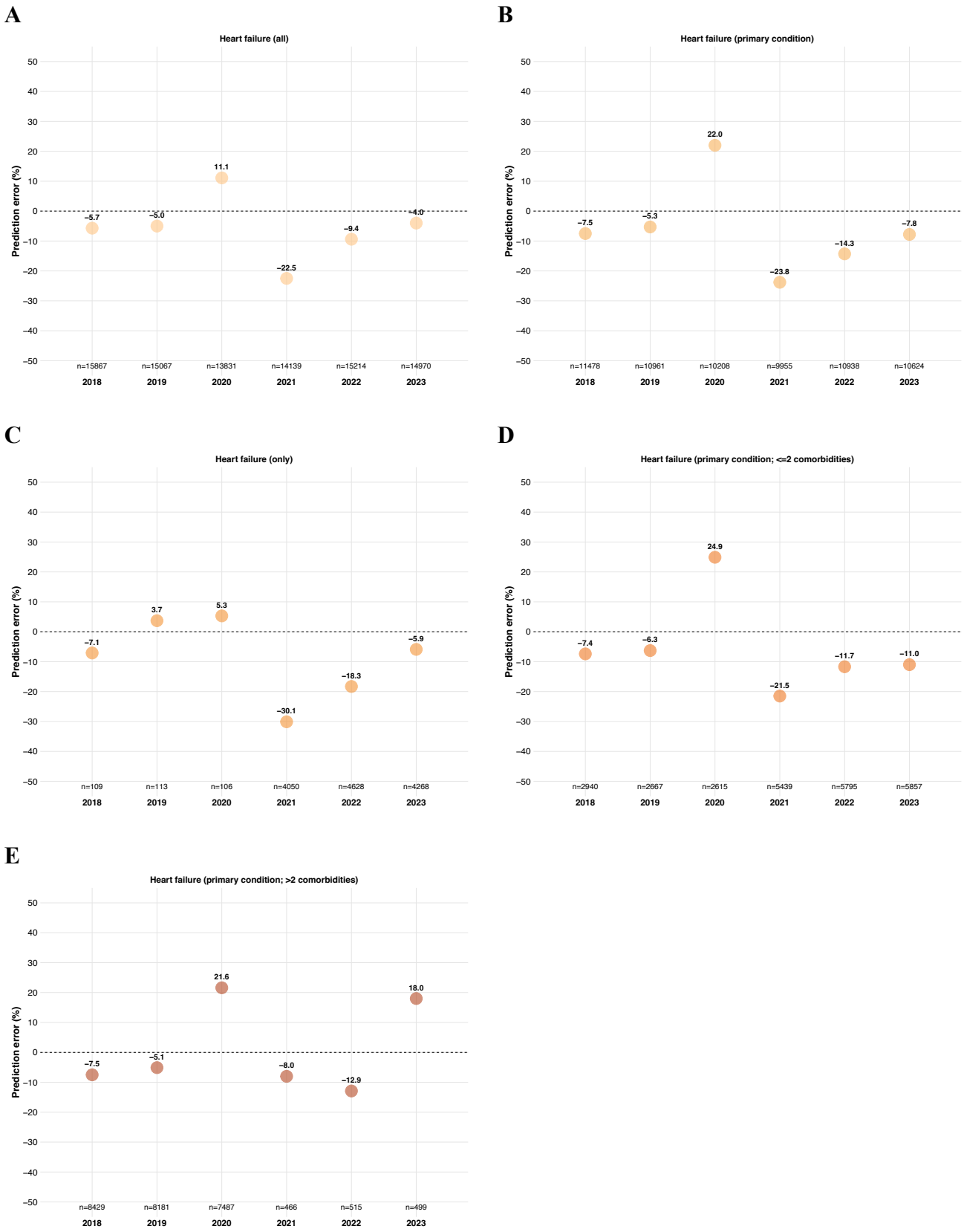


Figure S41. Annual predicted and actual mean SSC for Heart failure patients' groups. Cost predictions derived from updated algorithms, including historical cost information in the predictors' set.

Figure S42. Prediction errors (PE) of SC for the Heart failure patients' groups. Cost predictions derived from updated algorithms, including historical cost information in the predictors' set.

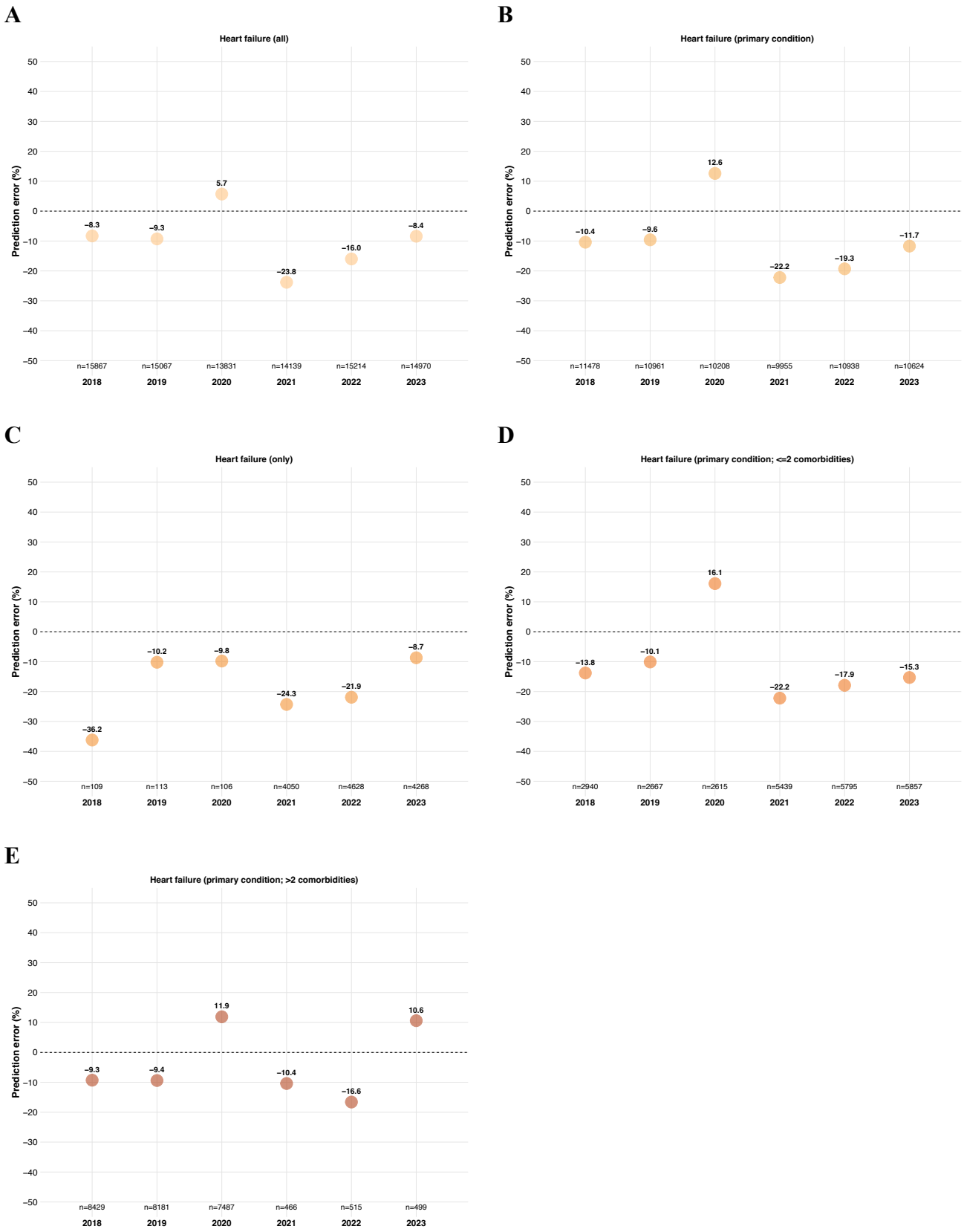


Figure S43. Annual predicted and actual mean SC for Heart failure patients' groups. Cost predictions derived from updated algorithms, including historical cost information in the predictors' set.

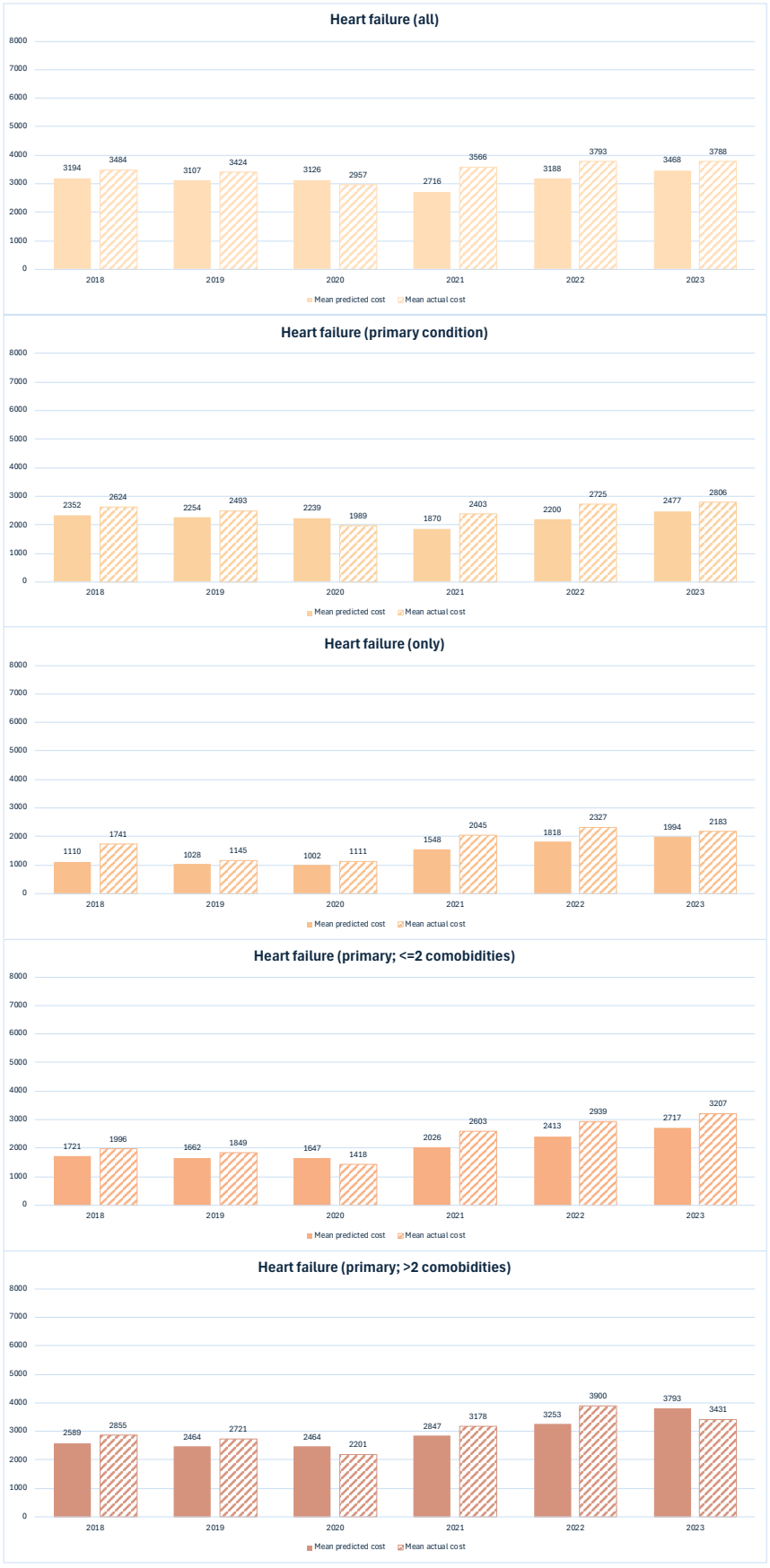


Figure S44. Prediction errors (PE) of TC, TSC, SSC, and SC, for the Heart failure "all" patients' group. Cost predictions derived from non-updated algorithms, excluding historical cost information from the predictors' set.

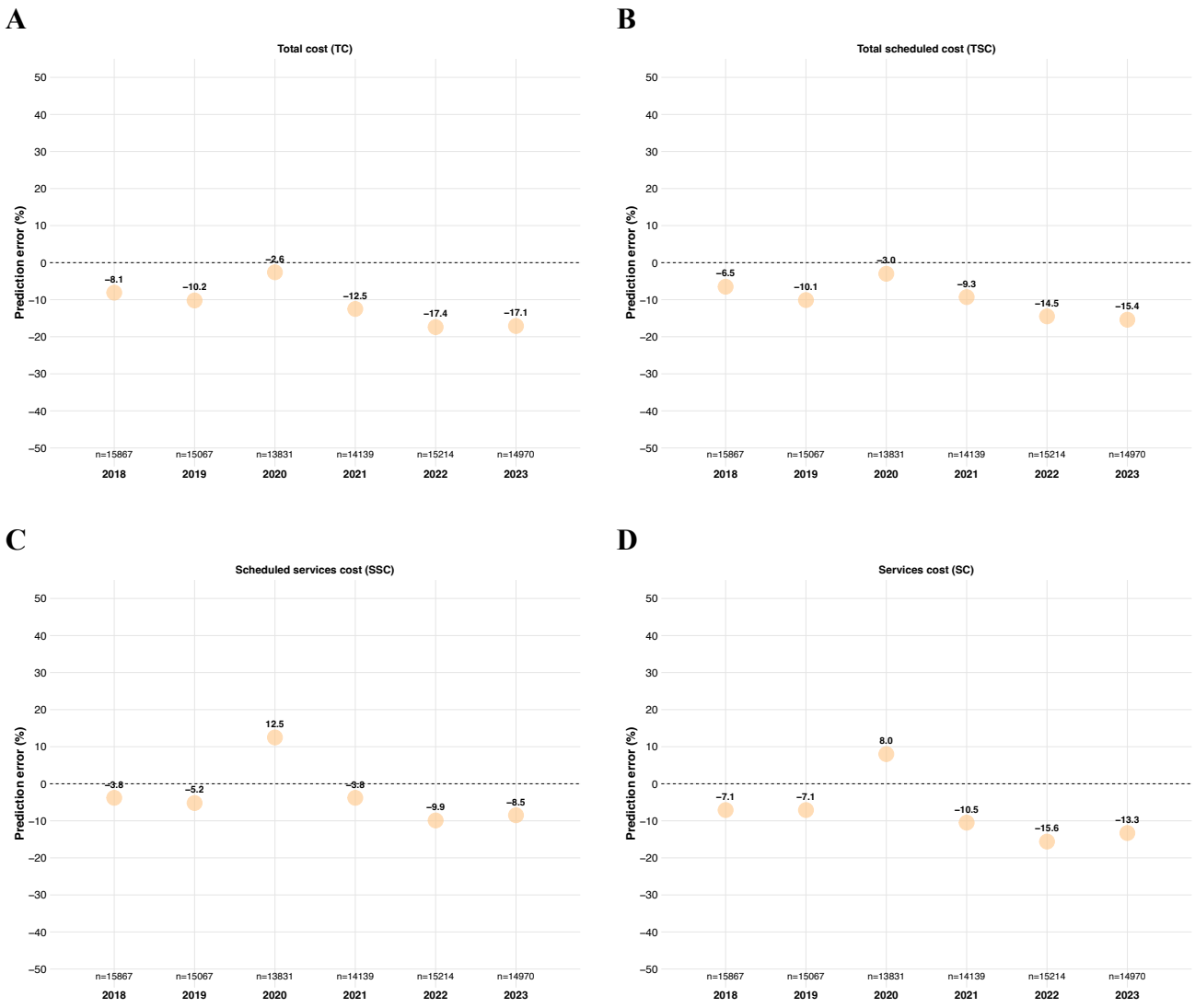


Figure S45. Prediction errors (PE) of TC, TSC, SSC, and SC, for the Heart failure "all" patients' group. Cost predictions derived from non-updated algorithms, including historical cost information in the predictors' set.

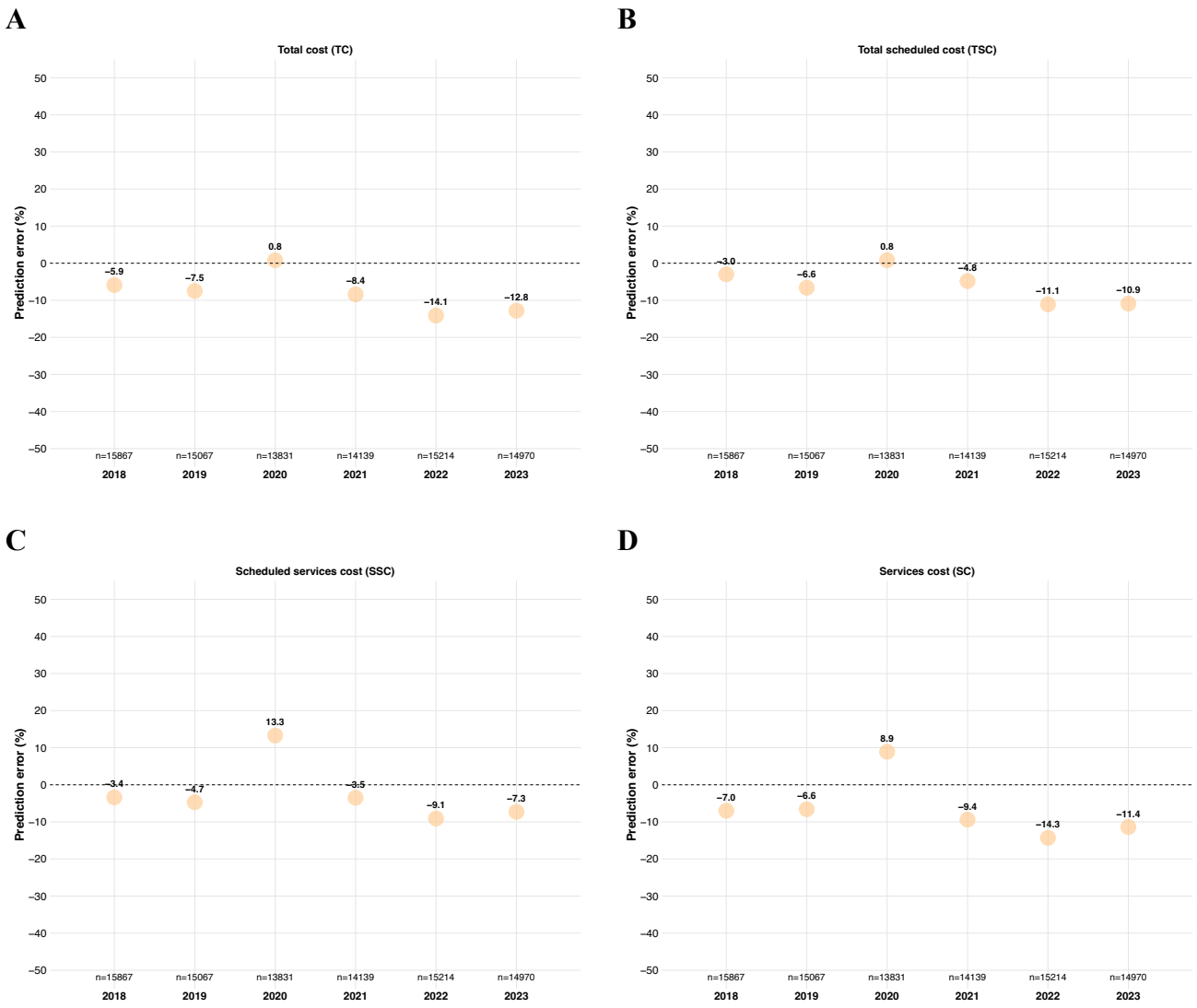
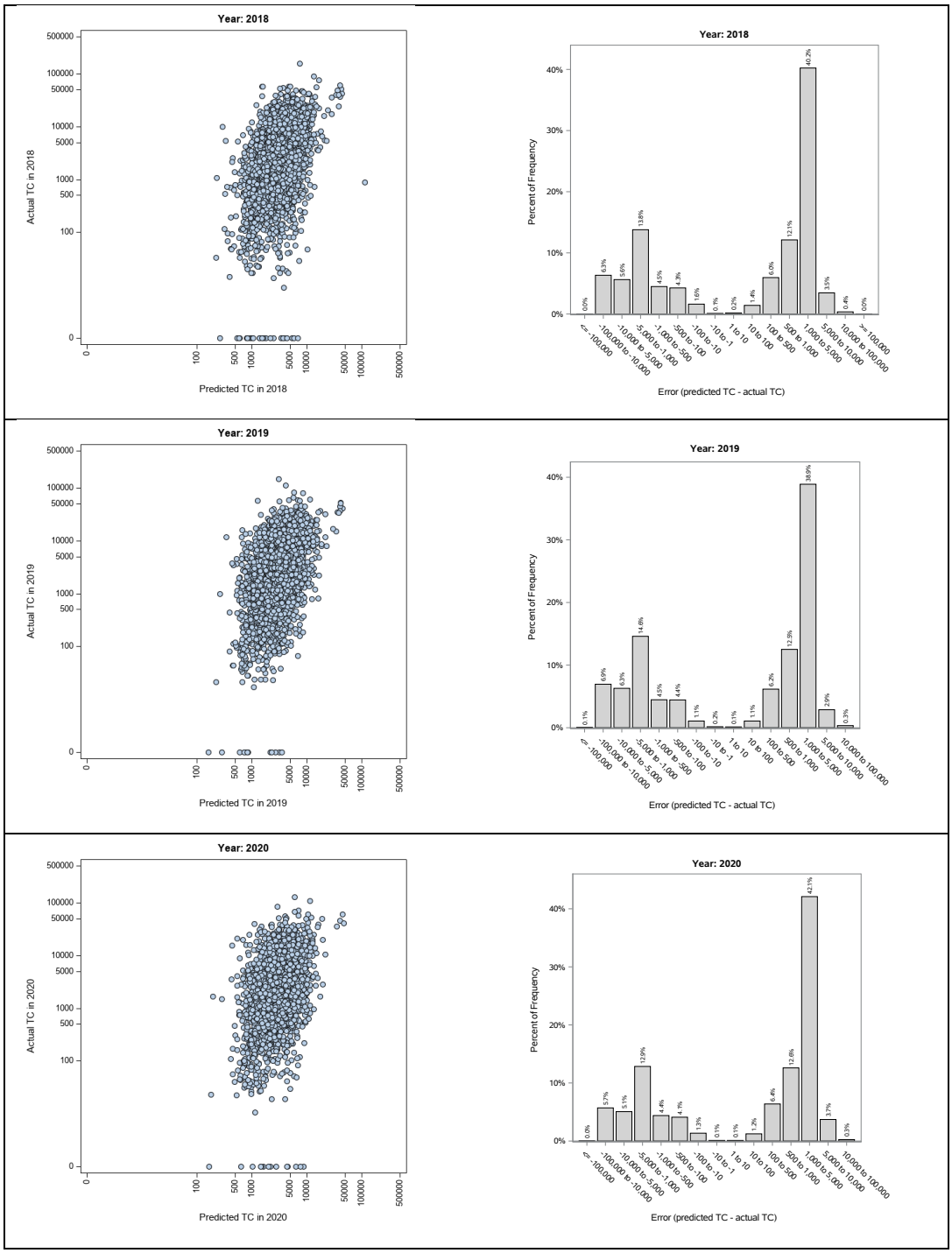


Figure S46. Scatter plots of actual TC vs. predicted TC (left-hand panels) and distributions of absolute errors (right-hand panels), in the Parkinson's disease/Parkinsonian syndromes "all" patients' group. Cost predictions derived from updated algorithms, excluding historical cost information from the predictors' set.



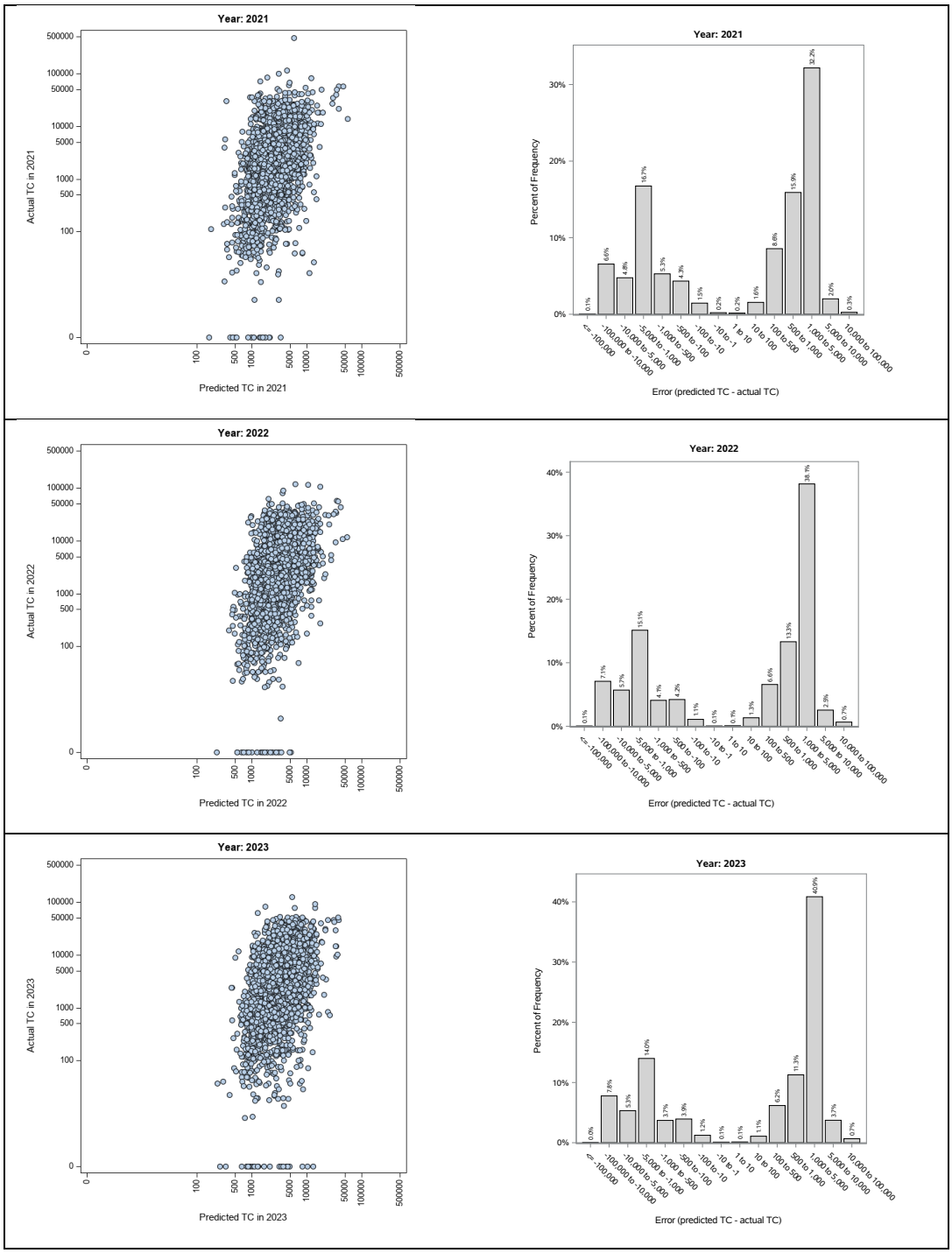

Year: 2021
Actual TC in 2021
Predicted TC in 2021
Year: 2021
Percent of Frequency
Error (predicted TC - actual TC)
Year: 2022
Actual TC in 2022
Predicted TC in 2022
Year: 2022
Percent of Frequency
Error (predicted TC - actual TC)
Year: 2023
Actual TC in 2023
Predicted TC in 2023
Year: 2023
Percent of Frequency
Error (predicted TC - actual TC)

Figure S47. Prediction errors (PE) of TC for the Parkinson's disease/Parkinsonian syndromes patients' groups. Cost predictions derived from updated algorithms, including historical cost information in the predictors' set.

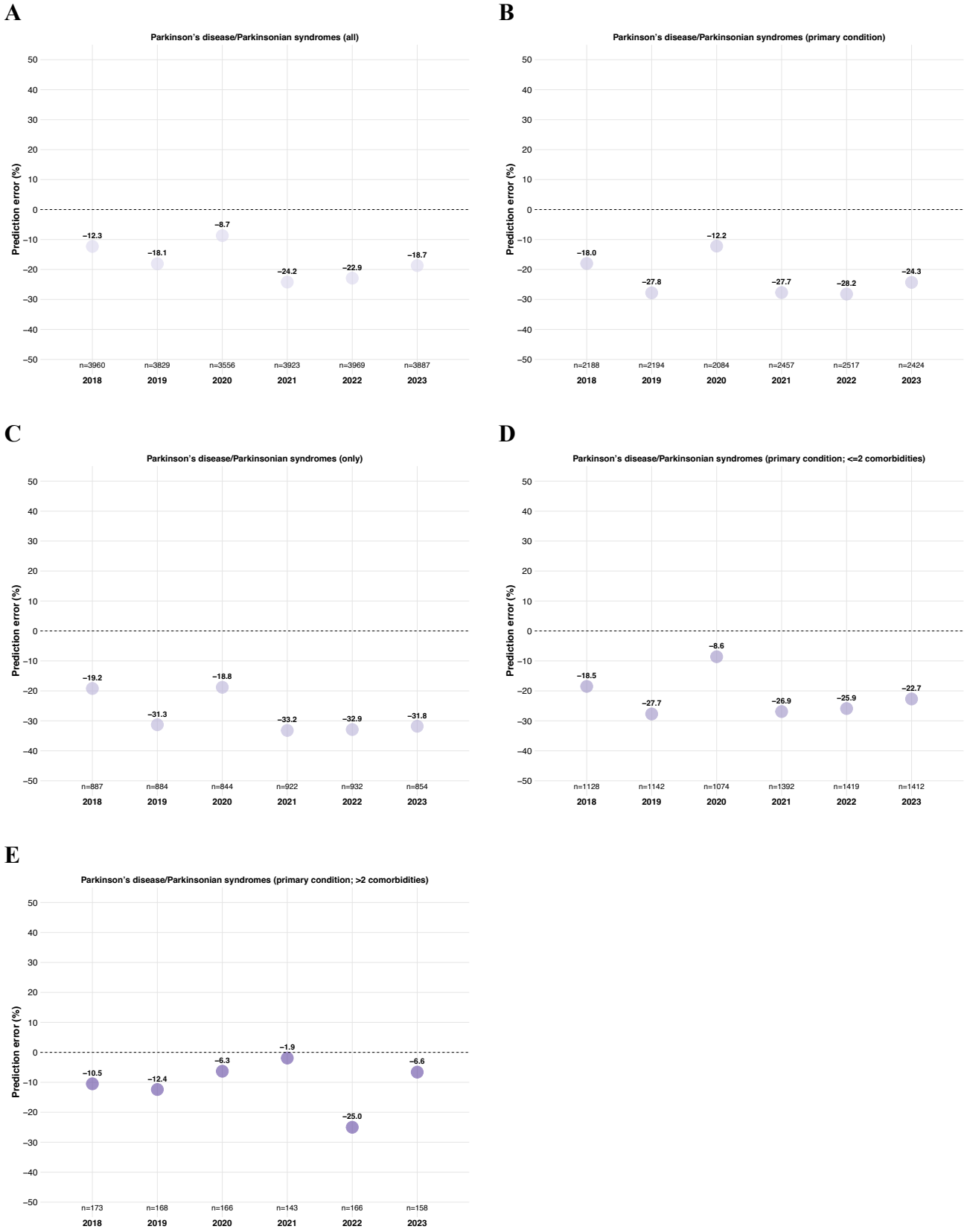


Figure S48. Annual predicted and actual mean TC for Parkinson's disease/Parkinsonian syndromes patients' groups. Cost predictions derived from updated algorithms, including historical cost information in the predictors' set.

Figure S49. Prediction errors (PE) of TSC for the Parkinson's disease/Parkinsonian syndromes patients' groups. Cost predictions derived from updated algorithms, including historical cost information in the predictors' set.

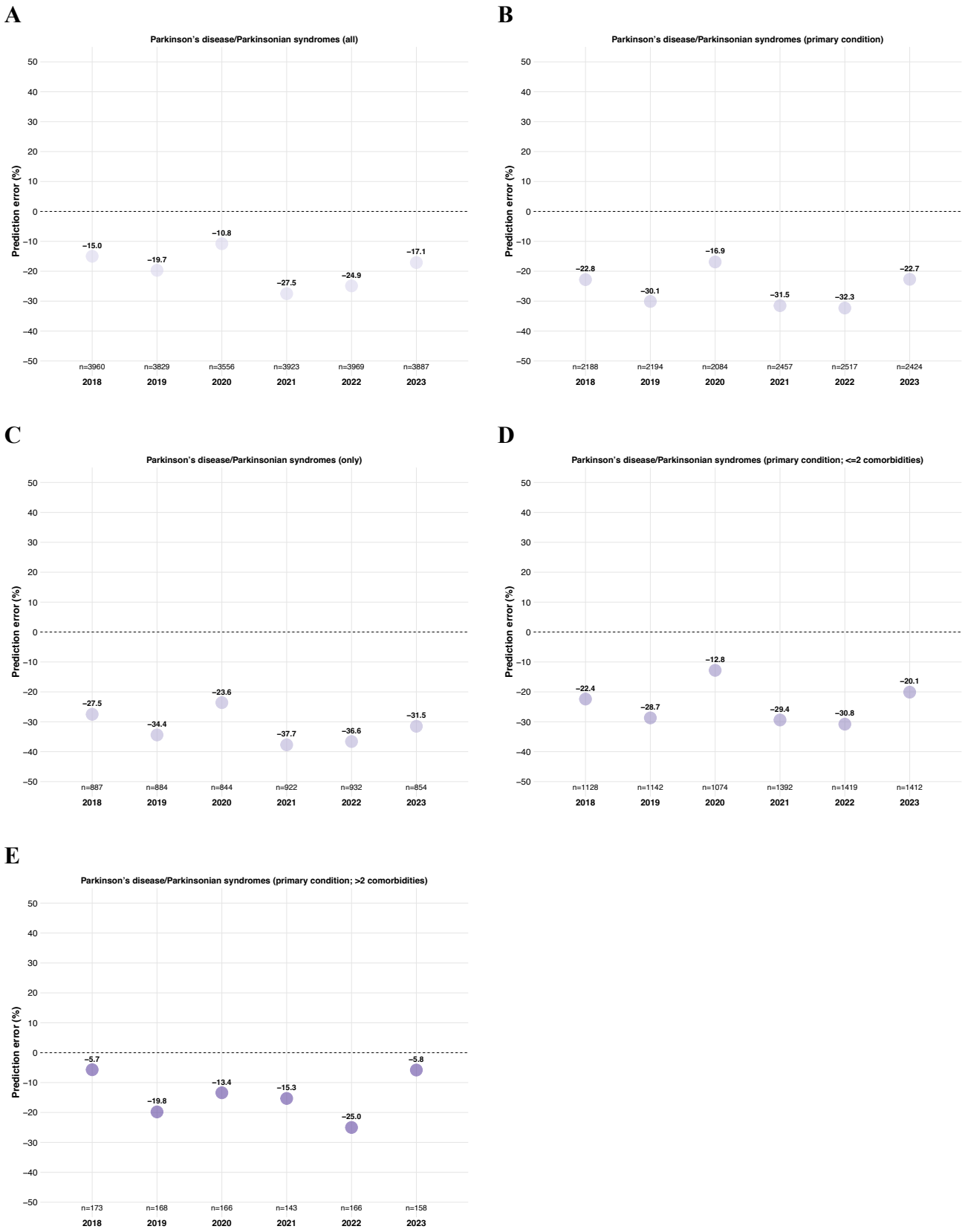


Figure S50. Annual predicted and actual mean TSC for Parkinson's disease/Parkinsonian syndromes patients' groups. Cost predictions derived from updated algorithms, including historical cost information in the predictors' set.

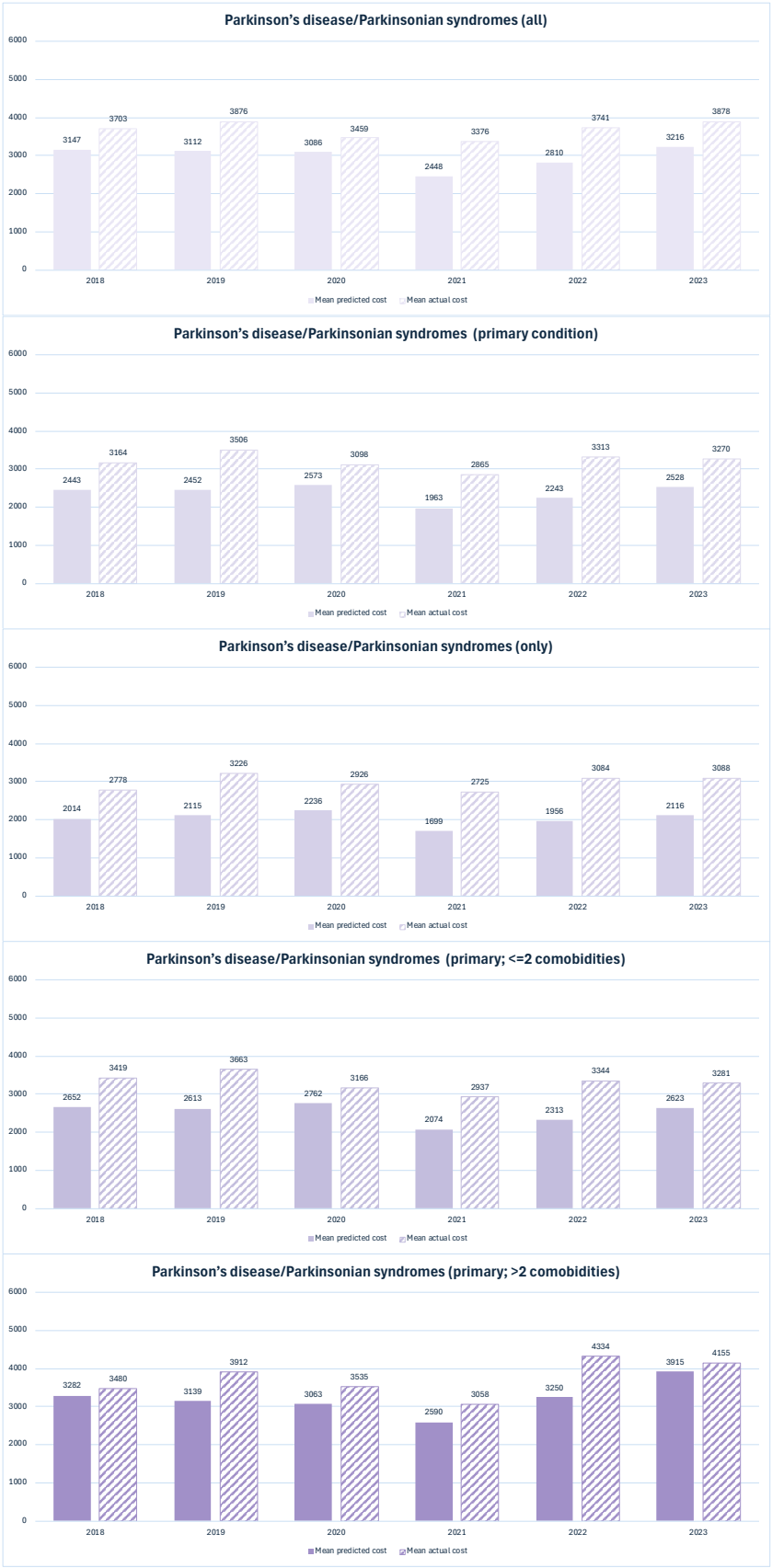


Figure S51. Prediction errors (PE) of SSC for the Parkinson's disease/Parkinsonian syndromes patients' groups. Cost predictions derived from updated algorithms, including historical cost information in the predictors' set.

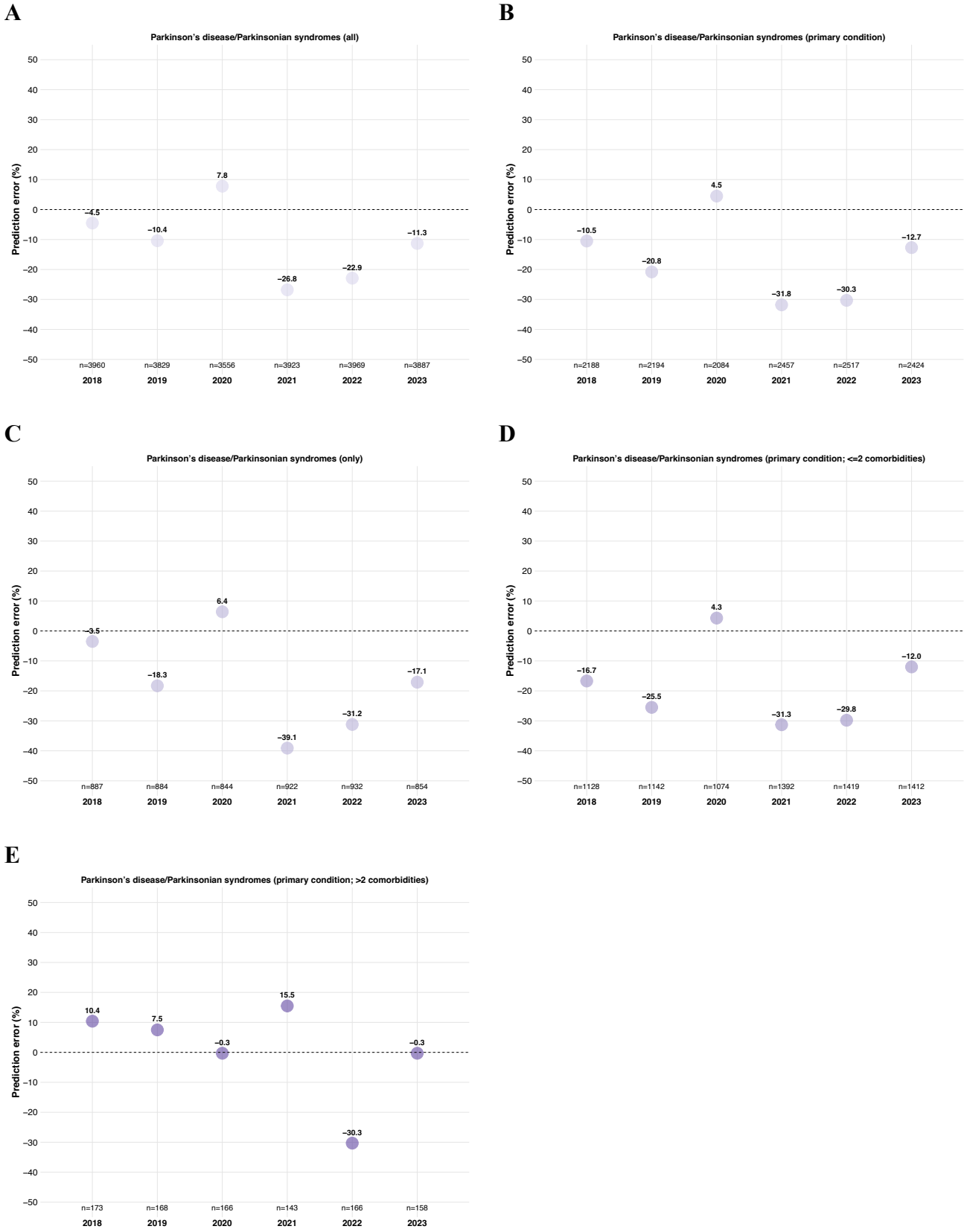


Figure S52. Annual predicted and actual mean SSC for Parkinson's disease/Parkinsonian syndromes patients' groups. Cost predictions derived from updated algorithms, including historical cost information in the predictors' set.

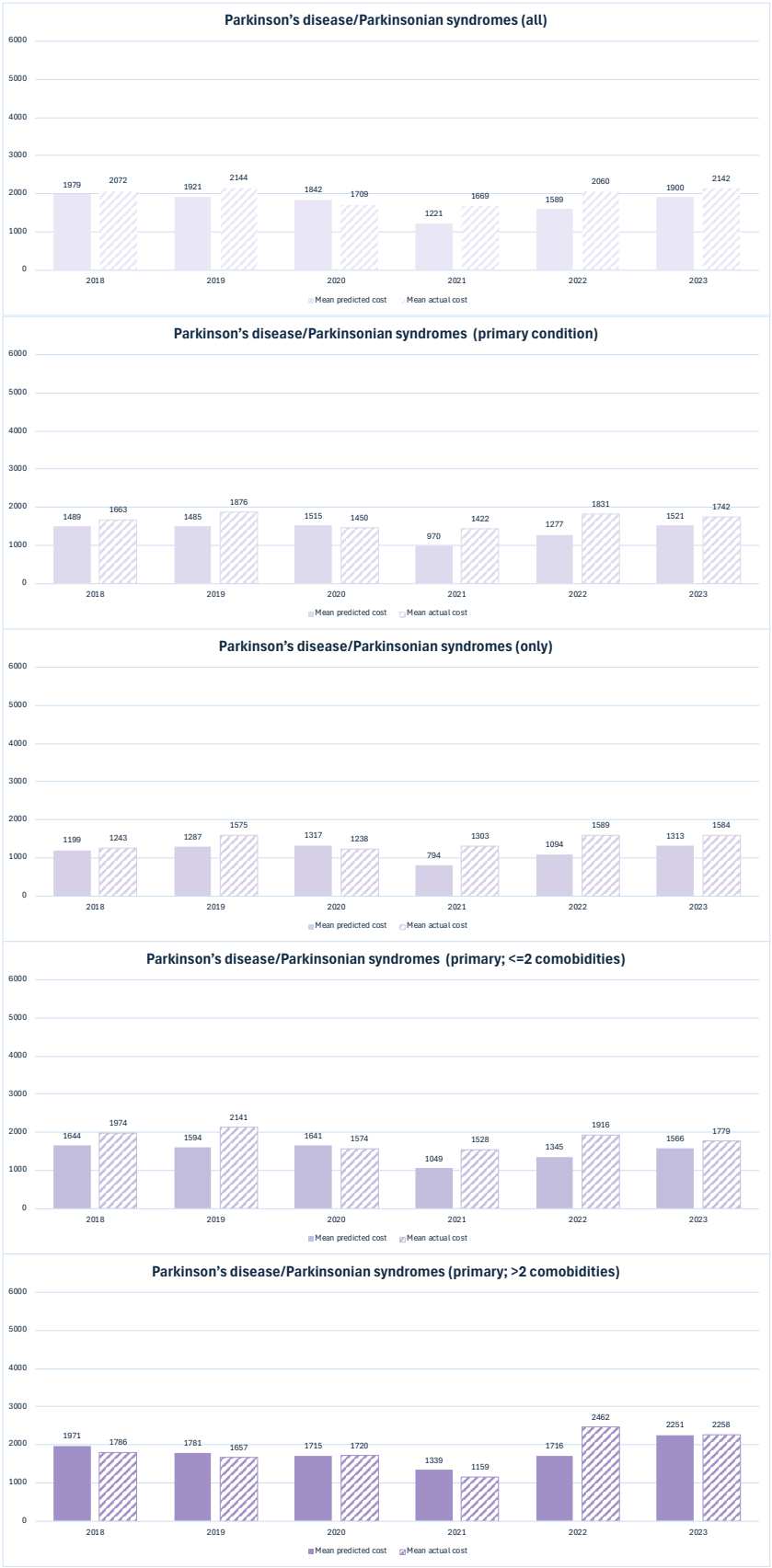


Figure S53. Prediction errors (PE) of SC for the Parkinson's disease/Parkinsonian syndromes patients' groups. Cost predictions derived from updated algorithms, including historical cost information in the predictors' set.

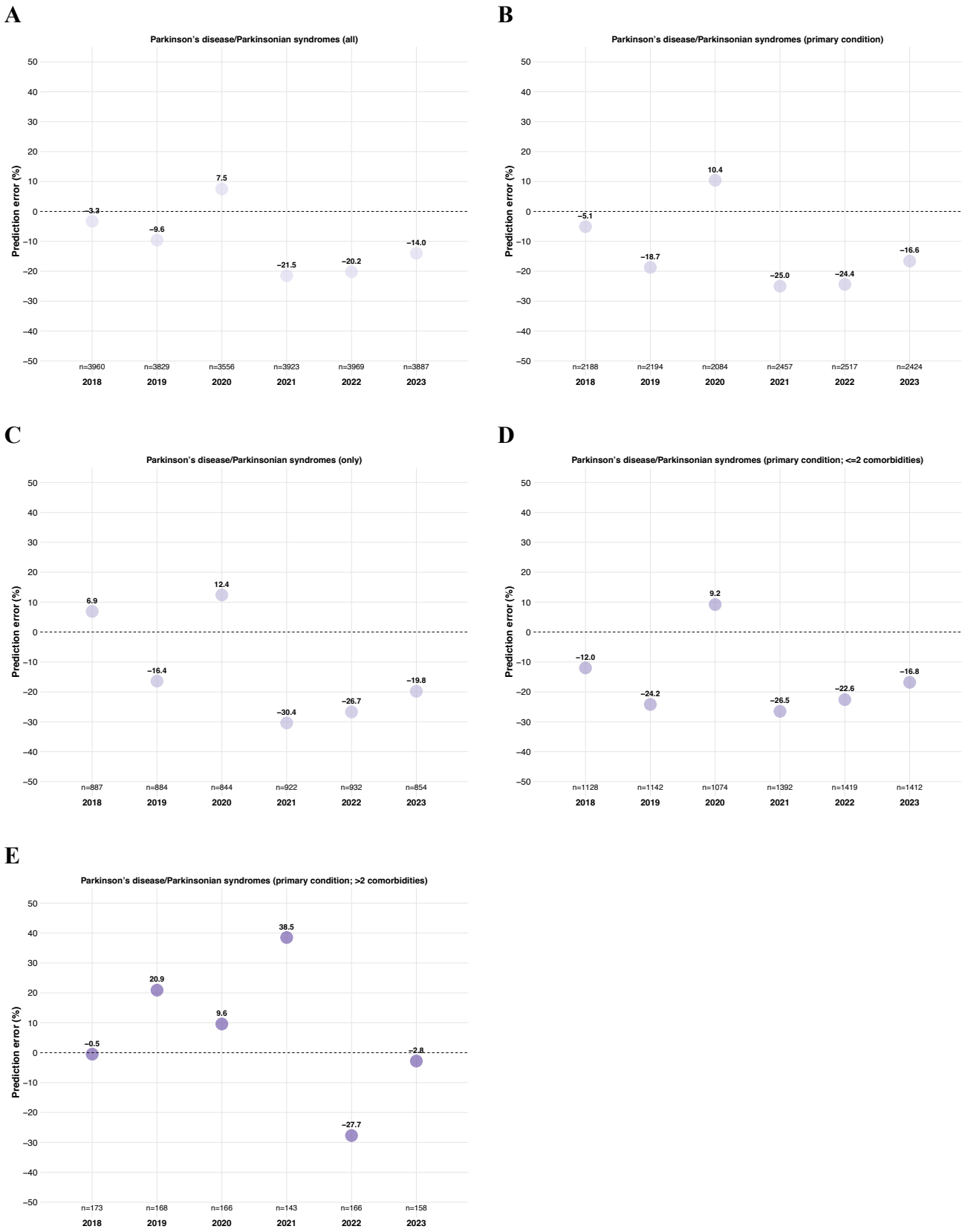


Figure S54. Annual predicted and actual mean SC for Parkinson's disease/Parkinsonian syndromes patients' groups. Cost predictions derived from updated algorithms, including historical cost information in the predictors' set.

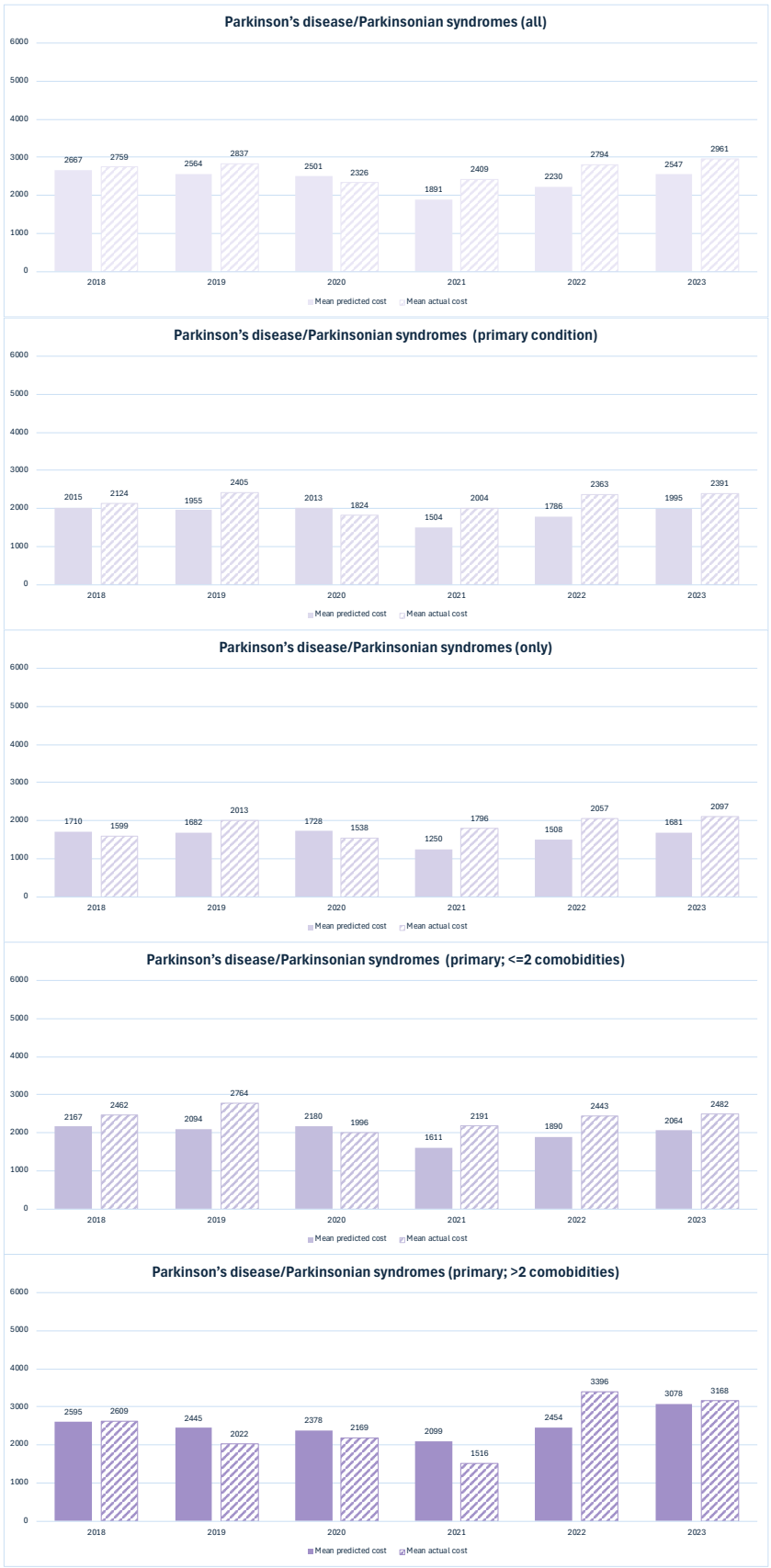


Figure S55. Prediction errors (PE) of TC, TSC, SSC, and SC, for the Parkinson's disease/Parkinsonian syndromes "all" patients' group. Cost predictions derived from non-updated algorithms, excluding historical cost information from the predictors' set.

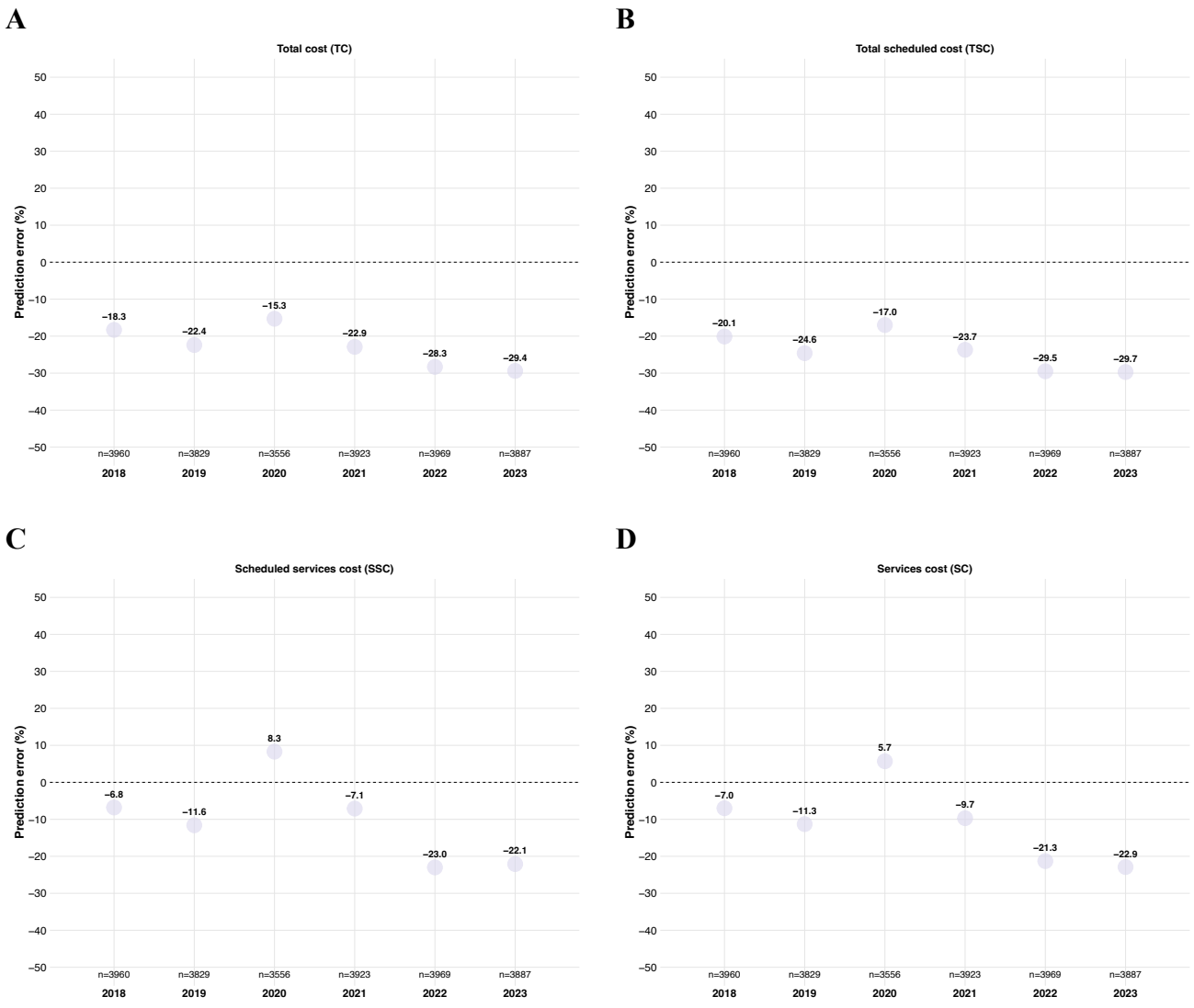


Figure S56. Prediction errors (PE) of TC, TSC, SSC, and SC, for the Parkinson's disease/Parkinsonian syndromes "all" patients' group. Cost predictions derived from non-updated algorithms, including historical cost information in the predictors' set.

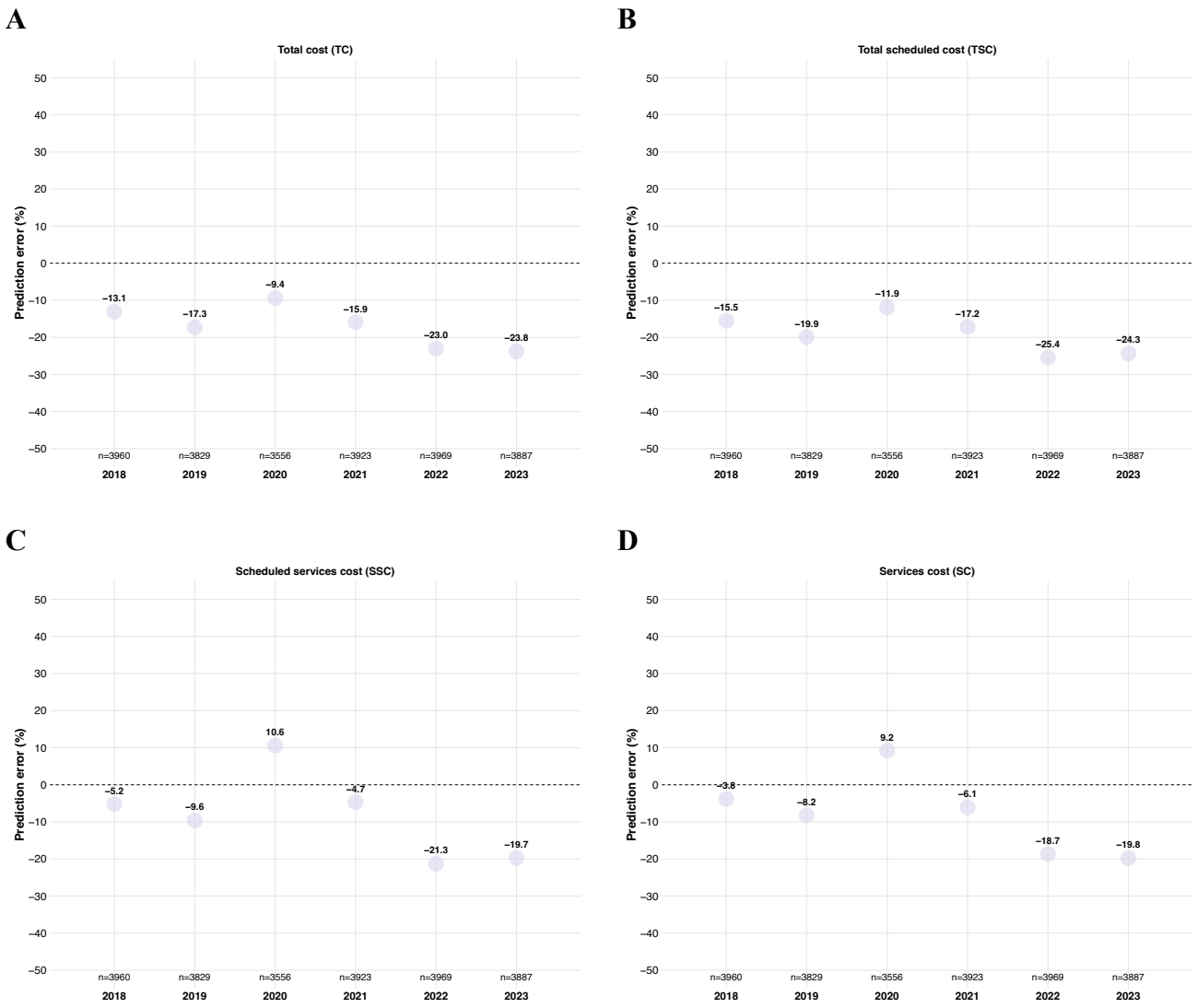
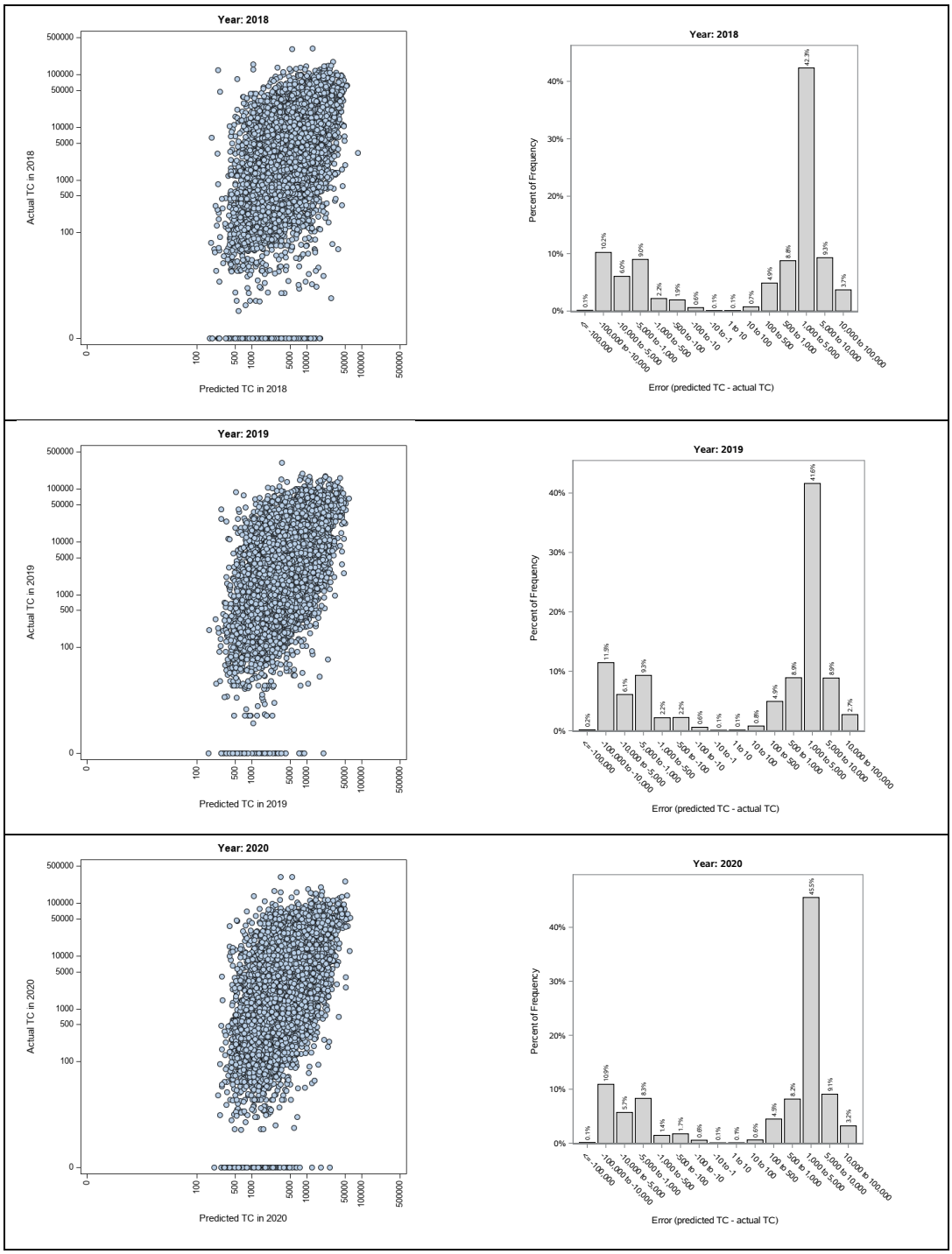


Figure S57. Scatter plots of actual TC vs. predicted TC (left-hand panels) and distributions of absolute errors (right-hand panels), in the Active neoplasia "all" patients' group. Cost predictions derived from updated algorithms, excluding historical cost information from the predictors' set.



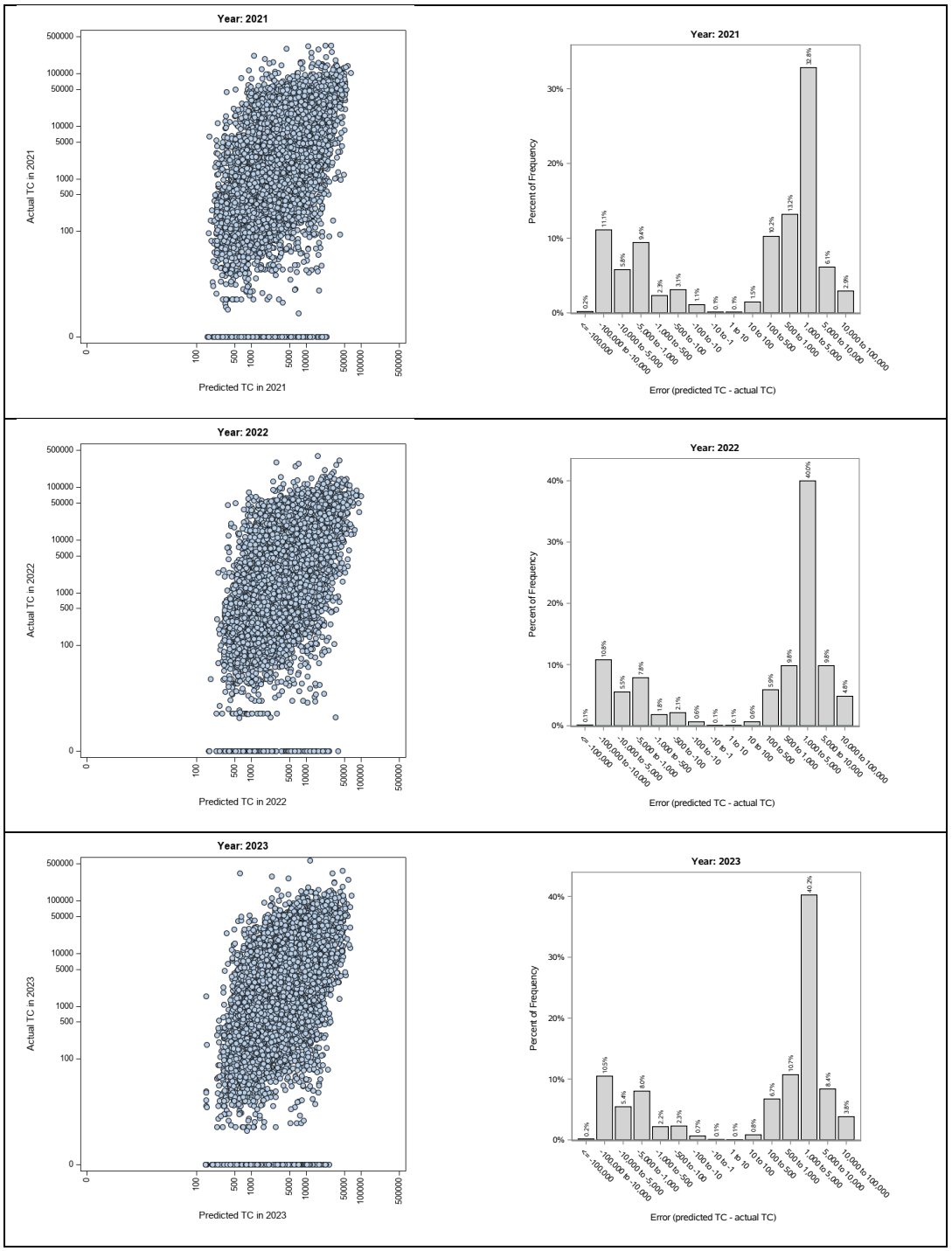

Year: 2021
Actual TC in 2021
Predicted TC in 2021
Year: 2021
Percent of Frequency
Error (predicted TC - actual TC)
Year: 2022
Actual TC in 2022
Predicted TC in 2022
Year: 2022
Percent of Frequency
Error (predicted TC - actual TC)
Year: 2023
Actual TC in 2023
Predicted TC in 2023
Year: 2023
Percent of Frequency
Error (predicted TC - actual TC)

Figure S58. Prediction errors (PE) of TC for the Active neoplasia patients' groups. Cost predictions derived from updated algorithms, including historical cost information in the predictors' set.

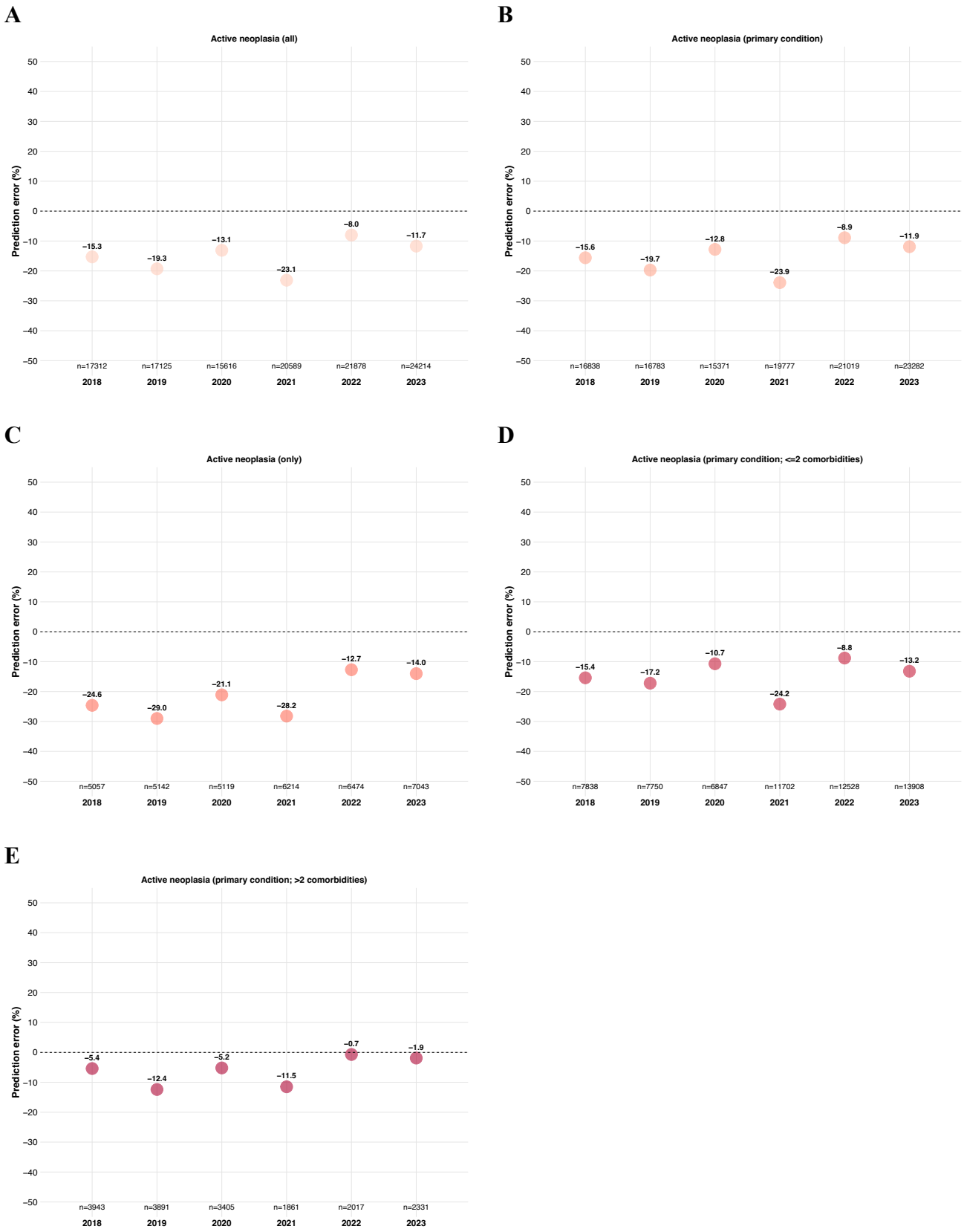


Figure S59. Annual predicted and actual mean TC for Active neoplasia patients' groups. Cost predictions derived from updated algorithms, including historical cost information in the predictors' set.

Figure S60. Prediction errors (PE) of TSC for the Active neoplasia patients' groups. Cost predictions derived from updated algorithms, including historical cost information in the predictors' set.

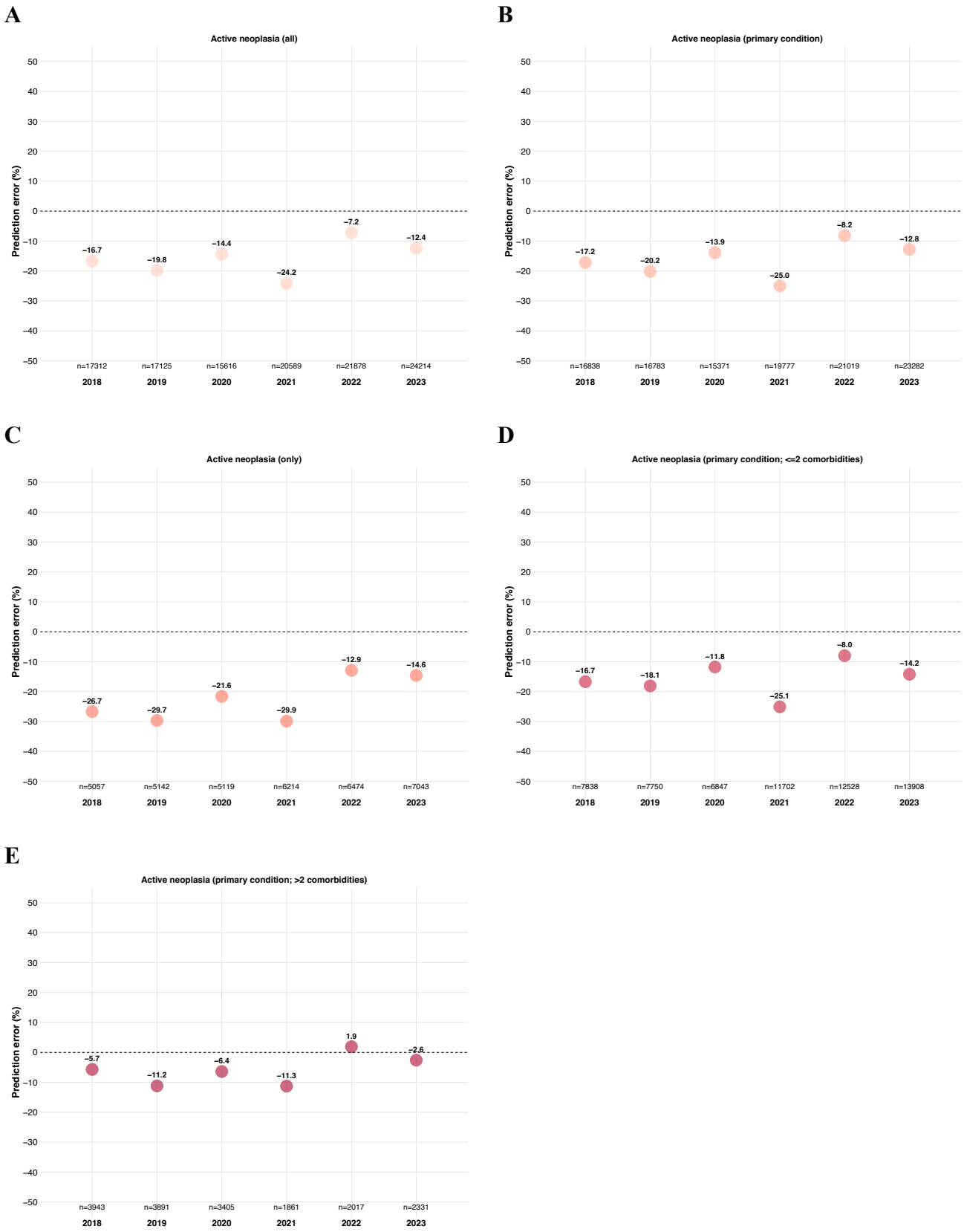


Figure S61. Annual predicted and actual mean TSC for Active neoplasia patients' groups. Cost predictions derived from updated algorithms, including historical cost information in the predictors' set.

Figure S62. Prediction errors (PE) of SSC for the Active neoplasia patients' groups. Cost predictions derived from updated algorithms, including historical cost information in the predictors' set.

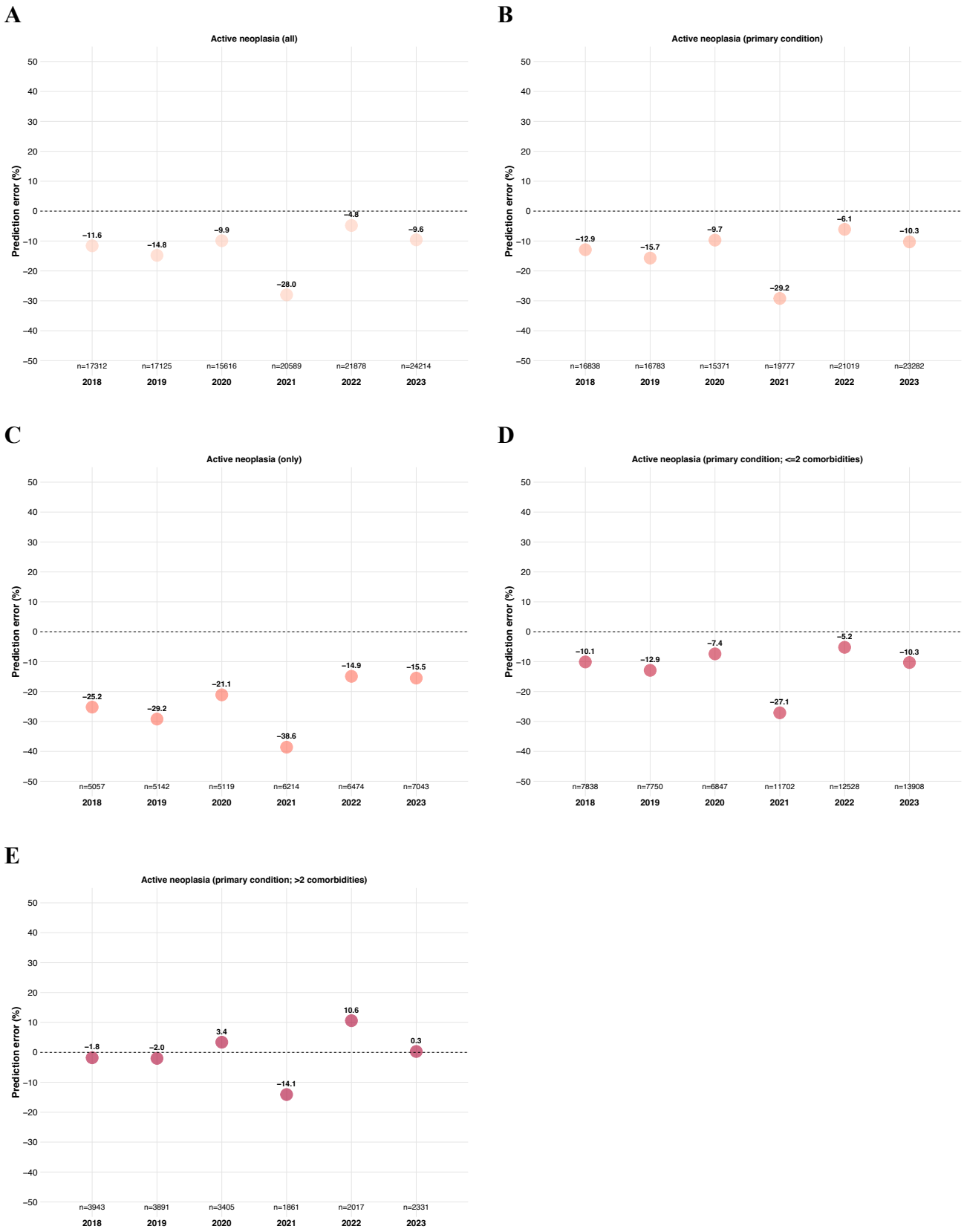


Figure S63. Annual predicted and actual mean SSC for Active neoplasia patients' groups. Cost predictions derived from updated algorithms, including historical cost information in the predictors' set.

Figure S64. Prediction errors (PE) of SC for the Active neoplasia patients' groups. Cost predictions derived from updated algorithms, including historical cost information in the predictors' set.

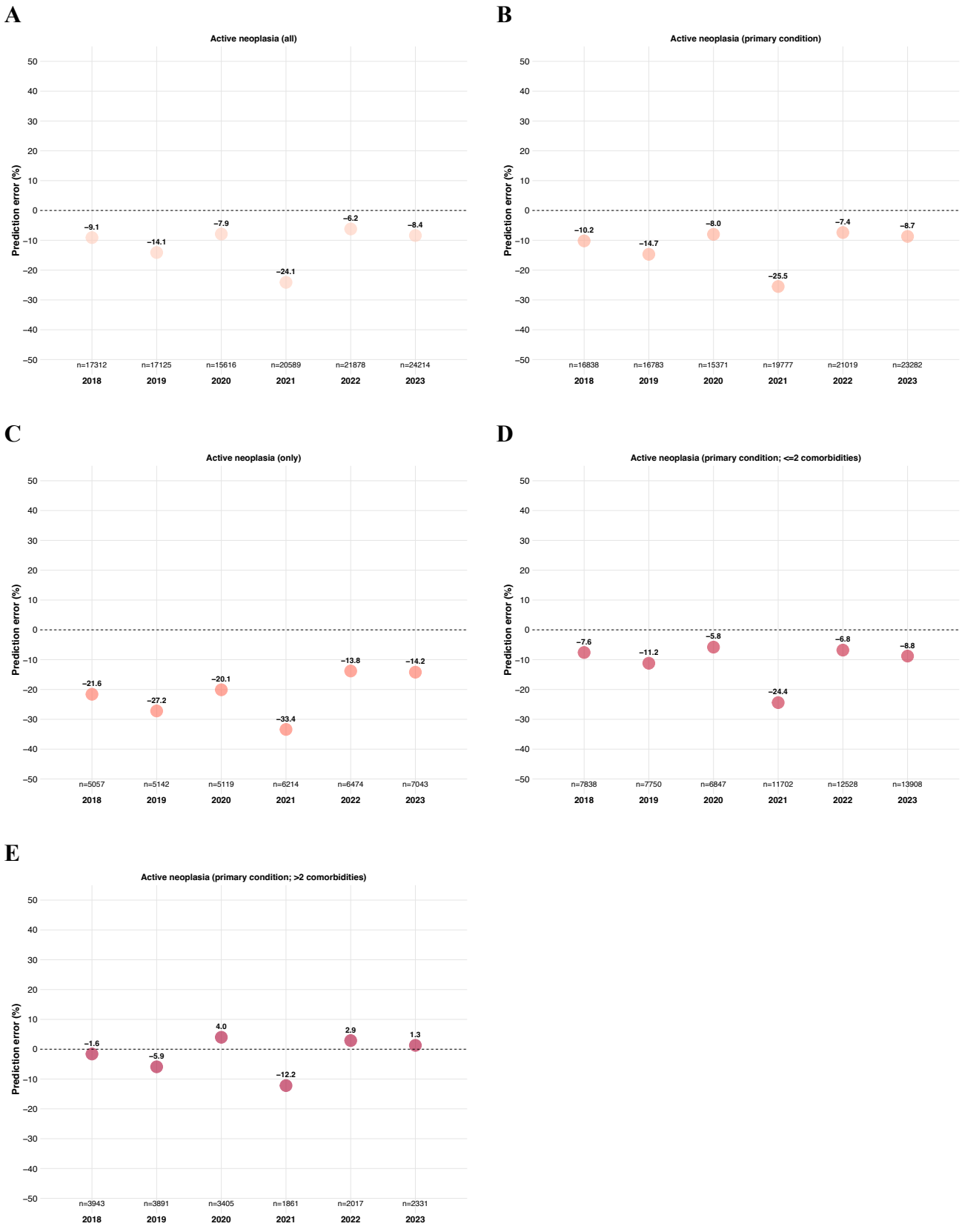


Figure S65. Annual predicted and actual mean SC for Active neoplasia patients' groups. Cost predictions derived from updated algorithms, including historical cost information in the predictors' set.

Figure S66. Prediction errors (PE) of TC, TSC, SSC, and SC, for the Active neoplasia "all" patients' group. Cost predictions derived from non-updated algorithms, excluding historical cost information from the predictors' set.

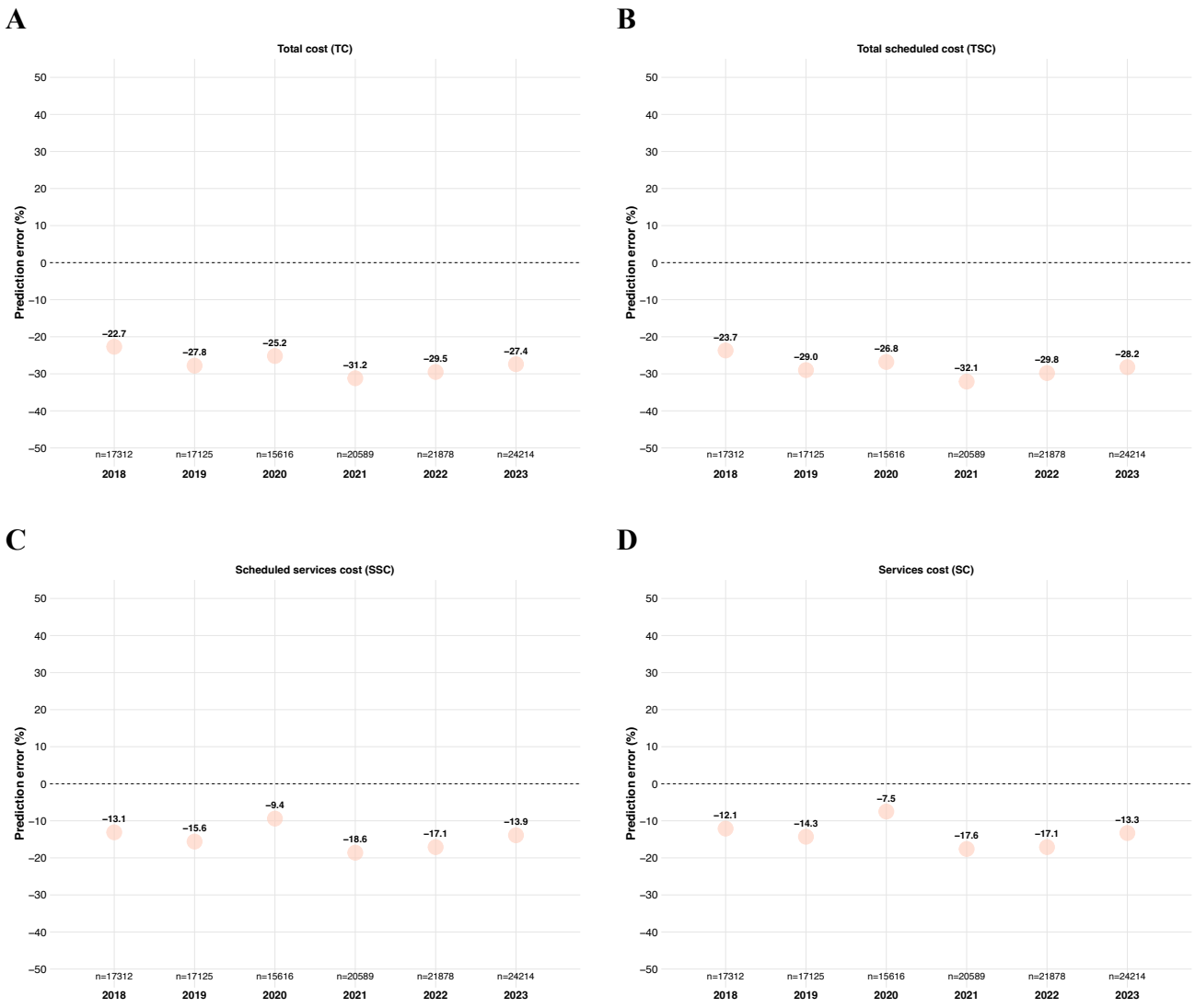
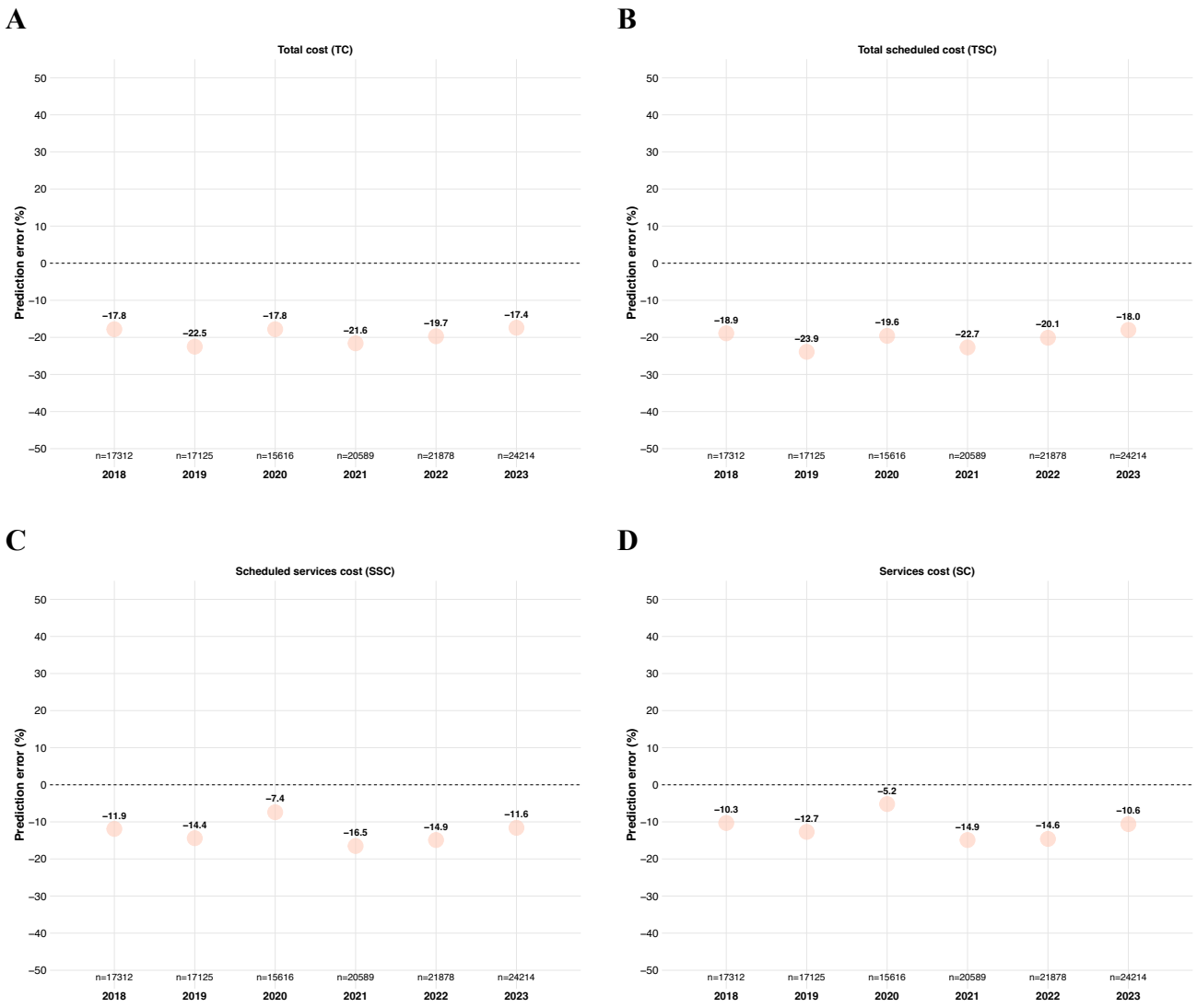


Figure S67. Prediction errors (PE) of TC, TSC, SSC, and SC, for the Active neoplasia "all" patients' group. Cost predictions derived from non-updated algorithms, including historical cost information in the predictors' set.



Acknowledgements

This project was carried out within a PhD scholarship co-funded by Next Generation EU Funds - PNRR and by Mediolanum Assicurazioni S.p.A. I am grateful to the Mediolanum team, especially to Dott. Massimo Grandis and Dott.ssa Federica Carlomagno, for their support and the trust they have placed in me throughout this work.

I would like to dedicate the final lines of this thesis to express my gratitude to all those who have been part of this journey. I feel fortunate to have the opportunity to work with brilliant colleagues. My sincere thanks go to Prof. Lorenzo Mantovani and Prof. Vincenzo Bagnardi for their continuous guidance and for involving me in such stimulating research projects. I am also deeply grateful to Dr. Sara Conti, whose constant support, precious feedback, and patience have been invaluable throughout these years.

La borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4 – Componente 1 – Riforma 4.1 Riforma dei Dottorati – Inv. 4.1
Borse PNRR patrimonio Culturale – CUP H41J22000220009