Received: 2 November 2021 | Revised: 2 June 2022 | Accepted: 9 June 2022

DOI: 10.1002/hec.4562

Health Economics WILEY

SUPPLEMENT PAPER

How real can we get in generating real world evidence? Exploring the opportunities of routinely collected administrative data for evaluation of medical devices

Benedetta Pongiglione | **Aleksandra Torbica**

Centre for Research on Health and Social Care Management, Bocconi University, Milano, Lombardia, Italy

Correspondence
Benedetta Pongiglione, Centre for Research on Health and Social Care Management, Bocconi University, Milano, Lombardia, Italy.
Email: benedetta.pongiglione@unibocconi.it

Funding information
European Union's Horizon 2020 research and innovation program, Grant/Award Number: 779306; Regione Lombardia
Open Access Funding provided by Universita Bocconi within the CRUI-CARE Agreement.

Abstract

Real-world data are considered a potentially valuable source of evidence for assessing medical technologies in clinical practice, but their widespread use is hampered by numerous challenges. Using the case of coronary stents in Italy, we investigate the potential of administrative databases for estimating costs and health outcomes associated with the use of medical devices in real world conditions. An administrative dataset was created ad hoc by merging hospital records from patients admitted between 2013 and 2019 for stent implantations with ambulatory records, pharmaceutical use data and vital statistics. Health outcomes were multifold: all-cause and cardiac mortality and myocardial infarction, within 30 days, 1, 2, 5 years. Costs were estimated from the National Health System perspective. We used multivariable Cox models and propensity score (PS) methods (PS matching; stratification on PS; inverse probability of treatment weighting using PS; PS adjustment). 257,907 coronary stents were implanted in 113,912 patients. For all health outcomes and follow-up times, and across all methods, patients receiving drug-eluting stents (DES) presented lower risk. For all-cause mortality, the DES patient advantage over bare-metal stent (BMS) patients declined over time but remained significant even at 5 years. For myocardial infarction, results remained quite stable. The DES group presented lower cumulative total costs (ranging from 3264 to 2363 Euros less depending on methods). Our results confirm the consolidated evidence of the benefits of DES compared to BMS. The consistency of results across methods suggests internal validity of the study, while highlighting strengths and limitations of each depending on research context. Administrative data yield great potential to perform comparative effectiveness and cost-effectiveness analysis of medical devices provided certain conditions are met.

KEYWORDS
administrative data, medical device, propensity score, real-world data, real-world evidence

1 | INTRODUCTION

In the past decade we have witnessed increasing calls for greater use of real-world data (RWD) and real-world evidence (RWE) by decision makers, both for regulatory reasons (i.e., enabling faster provision of innovative technologies to patients) and to support health technology assessment (HTA) and decisions regarding market access and reimbursement. The reasons for this increased attention are multiple. First, – especially in the case of highly innovative products -clinical trial-generated evidence may not be available at all, leaving decision makers with the option of delaying the decision or making the assessment on available evidence. Second, evidence stemming from RWD is expected to address some of the limitations associated with the use of trial-generated efficacy data. It has been argued that, due to the complexity of treatment regimens, demographic and clinical heterogeneity of patients receiving treatments, and the long time frame of many treatments, trial-generated data present limitations to predict the relative effectiveness and economic impact of health technologies in routine clinical practice (Makady et al., 2017).

Contemporarily with this greater opening by HTA agencies to RWD, the quantity and availability of RWD has grown exponentially, thus providing more opportunities to researchers to generate RWE. Although interest is observed across all types of health technologies, the greatest opportunity for wider use of RWD may be in the medical device field. This is due to intrinsic characteristics of medical devices that make their impact particularly sensitive to the real-life conditions in which they are used. The issues include iterative and rapid changes in device design, the need to account for the role of operator expertize in clinical outcomes (i.e., the effectiveness of a device is likely to differ largely from its efficacy due to use-specific effects, like the presence of a learning curve), and challenges in the realization of an experimental study which does not grant the elimination of biases from selection or awareness (lack of blinding). In addition, diffusion of medical devices in clinical practice, in the vast majority of cases, is not bound to experimental studies, which compels decision-makers to decide upon other types of evidence (Tarricone, Torbica, & Drummond, 2017).

While the growing availability of RWD could be seen as a prime opportunity to respond to the changing needs of decision makers, the enormous variety of data sources could lead to confusion. RWD can include an infinite variety of sources from medical health records, registries, biobanks, administrative data, health surveys, observational studies, health insurance data to data generated from mobile applications, etc. Thanks to growing attention on the matter recently, there has been some effort to reach consensus on the definition of RWD that has brought some clarity (Makady, de Boer, Hillege, Klungel, & Goettsch, 2017).

With specific focus on medical devices, a plethora of potential sources of data stems from routine clinical practice. A recent study systematically mapped RWD sources in 15 European countries for three different medical devices. The number of sources varied substantially across countries for each case study, and there was enormous heterogeneity in the quality and quantity of data collected. Depending on the characteristics of the devices and the stage of the product lifecycle, certain types of RWD sources were more commonly available: registries are the most important source for mature technologies, but they present a series of challenges for generating RWE (Pongiglione et al., 2021).

Digital innovation has permitted the generation, collection and storage of a large volume of health-related data that can be employed to track patient health and monitor health service delivery and technology use during all stages of the life cycle. In recent years, electronic administrative databases have become increasingly popular as a source of RWD for several reasons. First, the virtually unselected nature of the target population allows for results that can be generalized to real-world routine clinical practice. Second, the availability of large temporal series makes it possible to study the long-term outcomes of chronic treatments. Third, data are routinely collected so that prompt availability of data reduces the time and costs of research (Corrao & Mancia, 2015).

While all these features hold the promise of better matching policymakers' needs, use of administrative data for evaluation of medical devices presents numerous challenges that need to be adequately addressed (Tarricone et al., 2017). In recent years, methodological research has made extraordinary progress with the development of methods, algorithms, and designs that overcome, or at least rationally take into account, the pitfalls of observational research, including those based on administrative data sets (Berger et al., 2017; Gansen, 2018; Mendlowitz et al., 2020; Wang et al., 2017).

The general aim of this study is to assess the feasibility and applicability of using administrative data to generate useful evidence for informing HTA of medical devices. As mentioned above, administrative data routinely collected by healthcare systems provide unique datasets that represent several advantages in addition to being readily accessible by decision makers. However, the extent to which this evidence and the methods used can support decision making on medical devices remains unclear.

We address this gap by conducting an empirical analysis of a large administrative database from the Lombardy region, the largest and most affluent region in Italy, that systematically collects data on demographic, clinical and economic information of

patients and healthcare structures in the region, allowing for an analysis of the entire history of diagnostic and therapeutic pathways of each individual. Our empirical analysis is aimed at providing a proof-of-concept case study on the opportunities and challenges of using administrative data for cost and outcome evaluation of medical devices. Specifically, we use the case study of coronary stents to compare the effectiveness on health outcomes of two types of medical devices and assess relative costs; this translates into answering the following research question, "Is beginning treatment with drug-eluting stents more effective than beginning treatment with bare-metal stents in reducing mortality and myocardial infarction observed up to 5 years?" "Are there differences in cumulative medical costs incurred up to 5 years by the two groups?"

2 | DATA AND METHODS

2.1 | Lombardy administrative databases

The Lombardy Region collects and analyses healthcare data to evaluate and compare the appropriateness, effectiveness and efficiency of healthcare provided over population groups or structures. The regional databases contain sensitive data according to the definition of Legislative Decree 196/2003, and we could access them through a formal agreement under the premise of a European Horizon 2020 project COMED (https://cordis.europa.eu/project/id/779306).

More specifically, we had access to the following administrative databases: hospital discharge records ('Scheda di dimissione ospedaliera' (SDO)) which includes a special appendix for endoprostheses implanted during the admission, called "SDO4"; ambulatory records; pharmaceutical use data and vital statistics. Discharge records include information on diagnoses, procedures, basic demographics, length of stay. Diagnoses and procedures are classified according to the International Classification of Diseases-ninth revision- Clinical Modification (ICD-9-CM) codes. Reimbursement to healthcare providers is based in the Diagnosis Related Groups (DRG) ICD-9-CM version for hospitalizations, and on a specific list of services and drugs for outpatient care (Mazzali et al., 2016).

Each database includes a unique, anonymized ID for each patient, which allows for follow up over time as well as linking the records from different datasets to combine patient demographic and clinical characteristics (from SDO) with information on the device implanted (from SDO4), medications prescribed and used, outpatient visits performed, and vital statistics. Data from SDO, ambulatory records and pharmaceutical use data have been available since 2000, while SDO4 since 2010. As described in the next section, we define as baseline the first stent implantation registered between 2013 and 2019; hence, we are able to control for previous hospitalizations as well as implantations.

2.2 | The case study: Coronary stents

The choice of using coronary stents as a case study, comparing bare-metal stents (BMS) and drug-eluting stents (DES), was based on the characteristics and clinical history of those medical devices. The scientific evidence and expert opinion on these technologies have changed substantially over time. DES entered the market in the early 2000s and was initially considered superior to BMS, resulting in quick adoption in the United States (US) and in most European countries (Epstein et al., 2012; Packer et al., 2006). In Italy, DES procedures comprised about 38% of all percutaneous transluminal coronary angioplasty (PTCA) procedures between 2003 and 2007, rising from 30% in 2004 to more than 50% in 2006 (Cappellaro et al., 2011). However, in the initial phase of DES use, an increase in the incidence of reported sub-acute stent thrombosis and higher long-term mortality rates were recorded, and the preference toward the use of DES declined after 2005 (Chitkara & Pujara, 2010). With the entrance into the market of second generation DES, with considerable improvements over the previous version, their use increased again (Epstein et al., 2012), and contemporary DES were generally preferred to BMS due to a lower risk of restenosis (Bønaa et al., 2016). In 2012, the European Society of Cardiology (ESC) issued a Class IIa, Level A recommendation stating that for the treatment of patients with a STEMI diagnosis, DES should generally be preferred over BMS unless there are specific contraindications (Steg et al., 2012). The more recent ESC guidelines, published in 2014 (Kolh et al., 2014) and 2018 (Neumann et al., 2019), recommended DES use even further, making it a Class Ia recommendation.

From the cost perspective, the purchasing costs for DES are generally higher compared to BMS, but their long-term outcomes make them more cost-effective than BMS (Cheng et al., 2019).

2.3 | Study design

When designing non-randomized studies for comparative effectiveness, a useful structural approach is the so-called 'target trial' approach proposed by Hernan and Robins (Hernán et al., 2016). It suggests designing any observational data analysis as if one would design a randomized clinical trial and outlines a framework for comparative effectiveness research using RWD that revolves around the explicit description and emulation of the target trial. Although the purpose of this study it is not to emulate a specific existing RCT, it is useful to structure the research following the main steps recommended by Hernan. We developed a study protocol accordingly, available as supplementary material, and have organized the description of methods according to the elements of the framework.

2.3.1 | Population study and eligibility criteria

To identify hospitalizations for percutaneous coronary interventions involving coronary stent implantation, we extracted data from the dataset "Endoprosthesis admission" (SDO4) when the device implanted was either a drug eluting coronary stent or a non-drug eluting coronary stent, for the years between 2013 and 2019. If a SDO4 record was not available, but hospital admissions reported a stent insertion procedure, patients were also eligible.

Our sample consisted of patients admitted to hospital between 2013 and 2019 for a stent implantation (i.e., we considered as baseline the first admission since 2013, where it was the first hospital admission ever for a patient or the first during the observation time). Patients receiving both types of stents simultaneously were not included.

2.3.2 | Treatment strategy and assignment

Two treatment strategies were compared: DES implantation and BMS implantation. The intention was to compare a treatment strategy that begins with DES compared to one that begins with BMS, irrespective of what subsequent treatment was received (e.g., including patients who received BMS and subsequently received DES).

The treatments in the observational study are identified from the SDO4 data and defined using codes specific to this dataset. Specifically, the codes 02B and 02H were used for DES (in two cases, the device is the same, the way the service is delivered is different, i.e., outsourced vs. not outsourced) and the codes 02A and 02G for BMS.[1]

When data from SDO4 were not available, but the corresponding hospital admission record was available, we identified treatments based on ICD-9-CM codes 36.06 (Insertion of non-drug-eluting coronary artery stent(s)) for BMS and 36.07 (Insertion of drug-eluting coronary artery stent(s)) for DES.

Subjects were assigned to a treatment arm at the time of their first hospital admission for a stent implantation between 2013 and 2019. There was no Grace period -time during which treatment initiation can happen-in the treatment strategy.

2.3.3 | Follow-up and clinical outcomes

The study baseline is at the time of eligibility, which is when individuals are admitted for stent implantation. The study's follow-up period starts at baseline and ends at the end of the 5-year follow-up period unless right censoring occurs. The three endpoints of interest for our study were time-to-event in nature and consisted of all-cause and cardiac mortality and the occurrence of myocardial infarction (MI) at different time points: 30 days, 1 year, 2 and 5 years after discharge for the baseline procedure. Data on deaths were retrieved from the regional dataset on vital statistics. Cardiac mortality was defined based on ICD10 classification selecting all deaths whose cause was under chapter *I* "cardiac rheumatoid mortality". MI was identified through hospital admissions with ICD-9-CM diagnosis codes 410. x.

2.4 | Economic outcomes

Costs were estimated from the National Health System perspective and are reported in Euros. Resource utilization included *baseline costs*, consisting of the cost of the first hospital admission and cost of implanted device(s); and *follow-up costs* up to

5 years including acute heart disease-related health care costs for inpatient admissions and general and heart-related health care costs for outpatient visits, costs of coronary stents implanted subsequently, and general and heart disease-related medications since baseline admission.

We identified heart disease-related hospital admissions based on the literature and discussions with a clinician expert in cardiology, and selected the following diagnosis codes: 410.x (acute myocardial infarction), 412 (old myocardial infarction), 429.7x (certain sequelae of MI not elsewhere classified); and procedure codes: 36.1x (Bypass anastomosis for heart revascularization), 00.6x (procedures on blood vessels), 36.0x (Removal of coronary artery obstruction and insertion of stent(s)), 39.50 (Angioplasty or atherectomy of other non-coronary vessel(s)).

For outpatient visits, the ICD-9-CM codes are not a compulsory field in the discharge form, hence we selected events based on the compulsory field "specialty code"; in particular: cardiology, vascular surgery for angiology, vascular surgery for angiology in outpatient surgery, and check-up visit.

The use of medications was measured using Anatomical Therapeutic Chemical (ATC) codes. The selection was based not only on the compounds specifically prescribed for coronary diseases, but also those used for the control of the traditional risk factors and the most frequent comorbidities. For medical therapy after PTCA with stent implantation, we selected aspirin (B01AC06), clopidogrel (B01AC04), Ticagrelor (B01AC24), Prasugrel (B01AC22), combinations (B01AC30), proton-pump inhibitor (A02BC*), acetylsalicylic acid, combinations with proton pump inhibitors (B01AC56), ticlopidine (B01AC05) (Brilakis, Patel, & Banerjee, 2013); we also considered the following extra medications: oral hypoglycemic drugs and/or insulin (A10), antihypertensive drugs (C02, C03, C08), cardiac drugs (C01), statins (C10AA, C10BA02), beta-blocking agents (C07), angiotensin-converting enzyme (ACE)-inhibitors/angiotensin II receptor antagonists (AAIIs) (C09) (Brilakis et al., 2013; Degli Esposti, Perrone, Veronesi, Buda, & Rossini, 2018).

2.5 | Baseline characteristics

Demographic characteristics of patients were available from SDO and included sex and age in years. Clinical variables were selected based on the literature (Bønaa et al., 2016; Piccolo et al., 2019) and availability from administrative records. It is indeed recommended to select potentially prognostically important covariates based on subject-matter expertize and existing literature, rather than on formal statistical hypothesis testing in the study sample (Austin, 2014).

The selected covariates comprised main diagnoses at the time of indicator admission, in particular angina (identified through ICD-9-CM 413), MI (ICD-9-CM 410. x), no ST-segment elevation (ICD-9-CM 410.7); ST-segment elevation (ICD-9-CM 410.* excluded 410.7*); and co-morbidities distinguishing diabetes mellitus (ICD-9-CM 250.x), hypertension (ICD-9-CM 401–405), hypercholesterolemia (ICD-9-CM 272.0) and hyperlipidemia (ICD-9-CM 272.2, 272); number of stents implanted per patient (retrieved from SDO4). In addition, we estimated the Charlson Comorbidity Index (CCI) (Charlson, Pompei, Ales, & MacKenzie, 1987; Gasparini, 2018). Then, we considered an indicator of previous hospital admission within 5 years before the baseline admission, as a proxy for general health before baseline; specific medical history within the previous five years included diabetes mellitus measured from hospital admissions (ICD-9-CM 250. x) and drug prescriptions (ATC A10x); hypertension (ICD-9-CM 401–405 and/or drug prescription for ATC C02) and hypercholesterolemia (ICD-9-CM 272.0 and/or drug prescription for ATC C10); previous myocardial infarction; previous stroke (ICD-9-CM 434.9), previous coronary-artery bypass grafting (ICD-9-CM 36.1).

2.6 | Statistical analyses

Continuous variables were presented as the mean ± standard deviation and categorical variables as numbers and percentages. Clinical and demographic characteristics between the study groups (DES and BMS) were compared with the Chi-squared test for categorical variables and *t*-test for continuous variables.

2.6.1 | Naïve analyses

First, we present univariate Kaplan and Meier (KM) survival curves at 30-day, 1-year and 5 years for all-cause mortality and MI by stent type for the selected sample. Symmetrically, we estimated unadjusted hazard ratios (HRs) with 95% confidence intervals (CIs) from Cox proportional hazard regression model for population differences.

These results are prone to bias, but are helpful to compare an expectedly wrong analysis to the improved adjusted analyses, illustrated below. We expect that the adjusted analyses will give a different result to the naïve analyses, and, specifically, HRs closer to 1, which suggests they are to some extent controlling for prognostic differences in patients who received the different treatments.

2.6.2 | Multivariable-adjusted models

We then used Cox proportional hazards regression models to provide adjusted hazard ratios (HRs) with 95% confidence intervals (CIs) of the indicator variable denoting stent treatment status for all-cause mortality and MI, adjusting for a set of baseline covariates described in Section 2.6. For cardiac deaths, as a sensitivity analysis, we also used a Fine and Gray's sub-distribution hazard model considering non-cardiac deaths as a competing risk (Fine & Gray, 1999).

After exponentiation, the regression coefficients derived from an adjusted Cox proportional hazards regression model denotes the conditional HR: the relative change in the hazard of the occurrence of a time-to-event outcome because of treatment. A conditional treatment effect is the average effect, at the individual level, of changing a subject's treatment status from untreated to treated.

Differently from real-world studies, randomized trials by using a Cox proportional hazards model allow for estimating marginal treatment effects, that is, the difference in outcomes between two populations that are identical in all respects, except that in one population everyone is treated, while in the second population everyone is untreated (Austin, 2014, p. 1244).

Statistical methods, such as propensity score methods, can be applied in real-world settings to estimate marginal effects, whereas conventional regression adjustment estimates conditional effects (Austin, 2014). However, when propensity score methods, such as propensity score matching, are used with prognostic covariates re-introduced in the final Cox model, the effect estimated is conditional.

2.6.3 | Propensity score methods

The propensity score is the probability of receiving the treatment, given a set of observed covariates. Propensity scores are used to reduce confounding and match groups based on these covariates. Following the approach of Austin (2014) in the use of propensity score methods with time-to-event outcomes in real-world settings, we selected four different propensity score methods for reducing the effects of confounding when estimating the effects of treatment on outcomes: propensity score matching, inverse probability of treatment weighting (IPTW) using the propensity score, propensity score subclassification and covariate adjustment using the propensity score.

2.6.4 | Propensity score matching

Propensity score was first estimated fitting a logistic regression for the probability of receiving a DES versus BMS, conditional on the demographic and clinical covariates. We then used nearest neighbor with replacement, which is particularly helpful in settings where there are few controls compared to the treated individuals (Stuart, 2010)- with a caliper matching of 0.2 σ_p (standard deviation of the estimated propensity scores of the sample), already applied to the case of coronary stents (Austin, 2011; Iqbal et al., 2016) and 1:3 ratio. Practically, given the peculiarity of treatment in the period considered in this study, such that cases largely and progressively outnumber controls, we derived propensity score using a logistic regression which modeled the probability of receiving BMS versus DES. Nearest neighbor matching usually estimates the average treatment effect on treated (ATT), as it matches control individuals to the treated group and discards controls who are not selected as matches (Stuart, 2010). In this case, however, because a caliper was used, the estimand corresponds to neither the ATT nor the average treatment effect (ATE), but rather to an average treatment effect in the remaining matched sample (ATM) (Greifer, 2022). The success of matching was examined using weighted standardized differences in the distribution of the potential confounders between the two stent groups, and a difference of 10% between the 2 groups was considered acceptable (See Table A1 of the Supplementary Material for balance diagnostics and Figure A2 for propensity score distribution after matching).

Finally, univariate Cox proportional hazards regression models were run to compare the risk of outcomes with DES versus BMS use in the propensity-matched cohorts (marginal effect), and then adjusting for all the covariates used in the multivariable Cox PH models, estimating conditional effects.

2.6.5 | Propensity score subclassification

Subclassification forms groups of individuals who are similar as defined by the decile of the propensity score distribution (Stuart, 2010). Propensity score subclassification provides ATT. Austin (2013) examined the performance of several propensity score subclassification-based estimators of the marginal effects, all relying on pooling stratum-specific effects, and found high bias highlighting that the method does not estimate marginal effects, but conditional measures of effect. To overcome this bias, we used the stratum weights to estimate marginal effects, as suggested by Greifer (2022). Then we estimated conditional effects introducing into the model all the covariates included in the multivariable Cox model described above.

2.6.6 | Inverse probability of treatment weighting using the propensity score

Inverse probability of treatment weighting uses weights based on the propensity score to create a synthetic sample in which the distribution of measured baseline covariates is independent of treatment assignment (Austin, 2014). In a point treatment situation (time-invariant covariates), we can adjust for a set of confounders C when estimating the effect of discrete exposure A by weighting observations i by the inverse probability weights. We used stabilized weights (Hernán, Miguel Ángel, Brumback, & Robins, 2000; van der Wal, Willem M & Geskus, 2011), that in a point treatment situation are expressed as

$$w_i = \frac{P(A_i = a_i)}{P(A_i = a_i | C_i = c_i)}$$

The denominator contains the probability of the observed exposure level given the observed values of covariates C, and the numerator contains the probability of the observed exposure level, which is just the observed frequency.

These weights were assigned to each patient, with IPTW = $1/w_i$ for treated patients (DES); and IPTW = $1/(1-w_i)$ for control patients (BMS). They allow for estimating the ATE. Moreover, we trimmed weights that exceed the 99.5 percentile to that value (Austin & Stuart, 2015).

In order to also estimate ATT, we assigned weight 1 to treated and w_i equal to $\frac{\hat{e}_i}{1-\hat{e}_i}$ to untreated $\left(w_i = A_i + (1 - A_i)\frac{\hat{e}_i}{1-\hat{e}_i}\right)$.

Figure A3 of the appendix illustrates ATE stabilized weights and ATT weight distribution plots.

Finally, Cox proportional hazards regression with robust standard errors (Austin, 2016) were used to determine the effect of DES use on selected outcomes incorporating all covariates included in the multivariable models and weighting for IPTW in the final model.

2.6.7 | Covariate adjustment using the propensity score

Finally, we used covariate adjustment using the propensity score. Here we estimated the propensity score for receiving DES versus BMS using logistic regression and then run a Cox model to estimate the treatment effect on time-to-event outcomes controlling for the propensity score. Austin (2013) showed via a Montecarlo simulation that the estimation of marginal hazard ratios with adjustment using PS is substantially biased, and when estimating conditional hazard ratios, the magnitude of bias seems to depend on settings (Austin 2013).

2.6.8 | Kaplan-Meier Sample Average

To estimate and compare the cumulative medical costs between DES and BMS patients, we considered the matched sample, obtained from 1:3 PSM and conducted Kaplan-Meier Sample Average (KMSA) estimates (Chen, Rolfes, & Zhao, 2015; Cheng et al., 2019; Lin et al., 1997). This technique accounts for censoring due to loss at follow-up and/or occurrence of the event (all-cause mortality) and partitions the follow-up period (5 years) for each patient into time-intervals, that we set at 1-year expressed in days. Medical costs are then calculated as the sum of the Kaplan-Meier estimates of the probability of survival to the beginning of each interval, multiplied by the corresponding estimates of costs, during the interval, being conditional on survival to the beginning of the interval. We produced four types of estimates, two "naïve" and two "weighted estimators". The first two are named "Available Sample estimator" (AS) and "Complete Case estimator" (CC) and are a weighted average; the

first considers all individuals ($\widehat{\mu_{AS}} = \frac{\sum_1^n M_i}{n}$ where M_i are total costs and n total individuals), the second estimator accounts for events ($\widehat{\mu_{CC}} = \frac{\sum_1^n \Delta_i M_i}{\sum_1^n \Delta_i}$ where Δ is the event indicator). The other estimators account for censoring as just described. The Bang and Tsiatis estimator (also known as Weighted Complete Case estimator) is $\widehat{\mu_{BT}} = \frac{1}{n}\sum_1^n \frac{\Delta_i M_i}{\hat{K}(T_i)}$ where $\hat{K}(T_i)$ is the Kaplan-Meier estimator of the probability of censoring at time Ti. The Zhao and Tian's estimator (also known as Weighted Available Sample estimator) is $\widehat{\mu_{ZT}} = \frac{1}{n}\sum_1^n \left[\Delta_i \frac{M_i}{\hat{K}(T_i)} + (1-\Delta_i)\frac{M_i - \overline{M(C_i)}}{\hat{K}(T_i)}\right]$ where $M(Ci)$ is the average of cost until time Ci among individuals with event time later than Ci. This is explained in the manual of the *R* package we used for the estimation (see below).

We produced these calculations for the total costs registered since baseline admission and then separately for each cost registered at baseline and at follow-up.

All analyses were run using *R* version 3.6.0. In particular, the package *survival* was used for Cox regression model (Therneau & Lumley, 2015) and *cmprsk* for competing risk (Gray & Gray, 2020), *ipw* for inverse probability of treatment weighting (van der Wal et al., 2011), *MatchIt* for propensity score matching (Ho, 2027) and *ccostr* for Kaplan-Meier sample average (Børty, Brøndum, & Bøgsted, 2019).

3 | RESULTS

3.1 | Sample

Out of 157,385 patients selected from SDO data, 113,912 had information on coronary stent implantation from SDO4. Excluding cases where both types of devices were contemporarily implanted ($N = 1821$) and including cases where the endoprosthesis record was not available from SDO4, but classification was retrievable from hospital records (DES corresponding to ICD-9-CM 36.07; BMS to 36.06) ($N = 2426$), we ended up with a sample equal to 114,517 (see figure A1 in the Supplementary Material), of which 114,502 had complete records on selected confounders.

Table 1 reports patient characteristics by type of stent. Compared to BMS patients, patients receiving DES on average were younger, and a higher percentage was male; they more frequently had diabetes; hypertension, hypercholesterolemia and hyperlipidemia. Their cardiac diagnosis was more frequently angina and MI with NSTEMI and less frequently MI with STEMI. DES patients (that were younger as observed) tended to have a better health history, with fewer previous admissions, lower CCI, lower prevalence of previous conditions including MI and hypertension, but higher prevalence of previous diabetes and hypercholesterolemia. The number of implanted stents at admission baseline was higher among DES patients (1.9 cardiac stents implanted on average vs. 1.5 in BMS patients), and their length of stay at first admission was shorter (6.8 vs. 9.3 days for BMS patients).

3.2 | Stents

Between 2013 and 2019, the endoprosthesis data (SDO4) included 257,907 coronary stents implanted on 113,912 unique patients. The distribution of the two types, drug-eluting and not drug-eluting, over time is illustrated in Figure 1, while Figure 2 reports the average purchase cost of device for the two types of stents over the 7 years. As expected, the use of DES was already predominant at the beginning of the study period, and in the following years it further increased (Figure 1), with BMS use declining from 20% in 2013 to less than 1% at the end of 2019. Conversely, over the study period the average device cost declined by about 50% for DES (837 Euro in 2013, 404 Euro in 2019) and increased by 87% for BMS (from 382 Euro in 2013 to 716 in 2019).

3.3 | Outcome analysis

Figure 3 reports the unadjusted Kaplan-Meier survival curves at 30-day, 1-year and 5 years all-cause mortality and MI by stent type for the selected sample (DES = 106,86, BMD = 8416). The curves do not account for potential differences in the populations receiving the two types of device, and therefore results are interpretable along with the descriptive statistics of health outcomes in Table 1: re-admissions for MI are much more frequent than mortality in the short term (30 days), and after two

TABLE 1 Baseline demographic and procedural characteristics for the total study population

	DES	BMS	*p*-value
N	106,086 (92.6)	8416 (7.4)	
Age (s.d.)	67.5 (11.4)	73.0 (12.0)	<0.0001
Male sex	80,851 (76.2)	5823 (69.2)	<0.0001
Comorbidities			
Diabetes mellitus	25,572 (24.1)	1697 (20.2)	<0.0001
Hypertension	26,117 (24.6)	1947 (23.1)	<0.0001
Hypercholesterolemia	93,792 (88.4)	6811 (80.9)	<0.0001
Hyperlipidemia	93,600 (88.2)	6805 (80.9)	<0.0001
Cardiac diagnosis			
Angina	17,559 (16.6)	712 (8.5)	<0.0001
Myocardial infarction No ST-segment elevation	23,012 (21.7)	1691 (20.1)	0.0004
Myocardial infarction ST-segment elevation	31,850 (30.0)	3951 (46.9)	<0.0001
Charlson comorbidity index (s.d.)	0.91 (1.01)	1.23 (1.17)	<0.0001
No. of stents implanted per patient (s.d.)	1.88 (1.22)	1.48 (0.91)	<0.0001
Previous conditions			
Previous hospital admission (yes)	45,172 (42.6)	4374 (52.0)	<0.0001
Myocardial infarction[a]	3428 (3.2)	412 (4.9)	<0.0001
Previous stroke[a]	126 (0.11)	15 (0.18)	0.2076
Coronary-artery bypass grafting[a]	1451 (1.4)	116 (1.4)	0.9362
Diabetes mellitus	23,113 (21.8)	1615 (19.2)	<0.0001
Hypertension	14,621 (13.8)	1596 (19.0)	<0.0001
Hypercholesterolemia	46,709 (44.0)	3558) (42.3)	<0.0001
Length of stay at first admission in days (s.d.)	6.84 (6.64)	9.29 (8.58)	<0.0001
Outcomes			
Mortality 30 days	1374 (1.3)	549 (7.1)	<0.0001
Mortality 1 year	3229 (3.0)	1142 (13.6)	<0.0001
Mortality 2 years	4503 (4.2)	1461 (17.4)	<0.0001
Mortality 5 years	6523 (6.1)	2169 (25.8)	<0.0001
Cardiovascular mortality 30 days	1254 (1.2)	547 (6.5)	<0.0001
Cardiovascular mortality 1 year	2302 (2.2)	852 (10.1)	<0.0001
Cardiovascular mortality 2 years	2878 (2.7)	985 (11.7)	<0.0001
Cardiovascular mortality 5 years	3676 (3.5)	1280 (15.2)	<0.0001
Myocardial infarction in 30 days	9041 (8.5)	1221 (14.5)	<0.0001
Myocardial infarction in 1 year	11,084 (10.4)	1567 (18.6)	<0.0001
Myocardial infarction in 2 years	12,148 (11.5)	1666 (19.8)	<0.0001
Myocardial infarction in 5 years	13,526 (12.8)	1841 (21.9)	<0.0001

Note: Values are numbers (percentage) of the total sample ($N = 114,571$) unless stated otherwise.

Abbreviations: BMS, bare-metal stents; DES, Drug-eluting stent; s.d., standard deviation.

[a]Percentage calculated among those having previous admissions.

and five years of follow-up; especially for BMS, mortality becomes a more frequent event. In terms of differences between the survival curves of the two types of stents, they are more marked for mortality and are all significantly different based on log-rank tests.

Table 2 synthesizes the results obtained for each health outcome using different methodologies. Different methods produced different estimands and treatment effects, as described in Section 2.7.3 and reported in the table. Hence not all results are directly comparable. Here we proceed discussing the univariable model first and then the other models, discussing results comparable to one another.

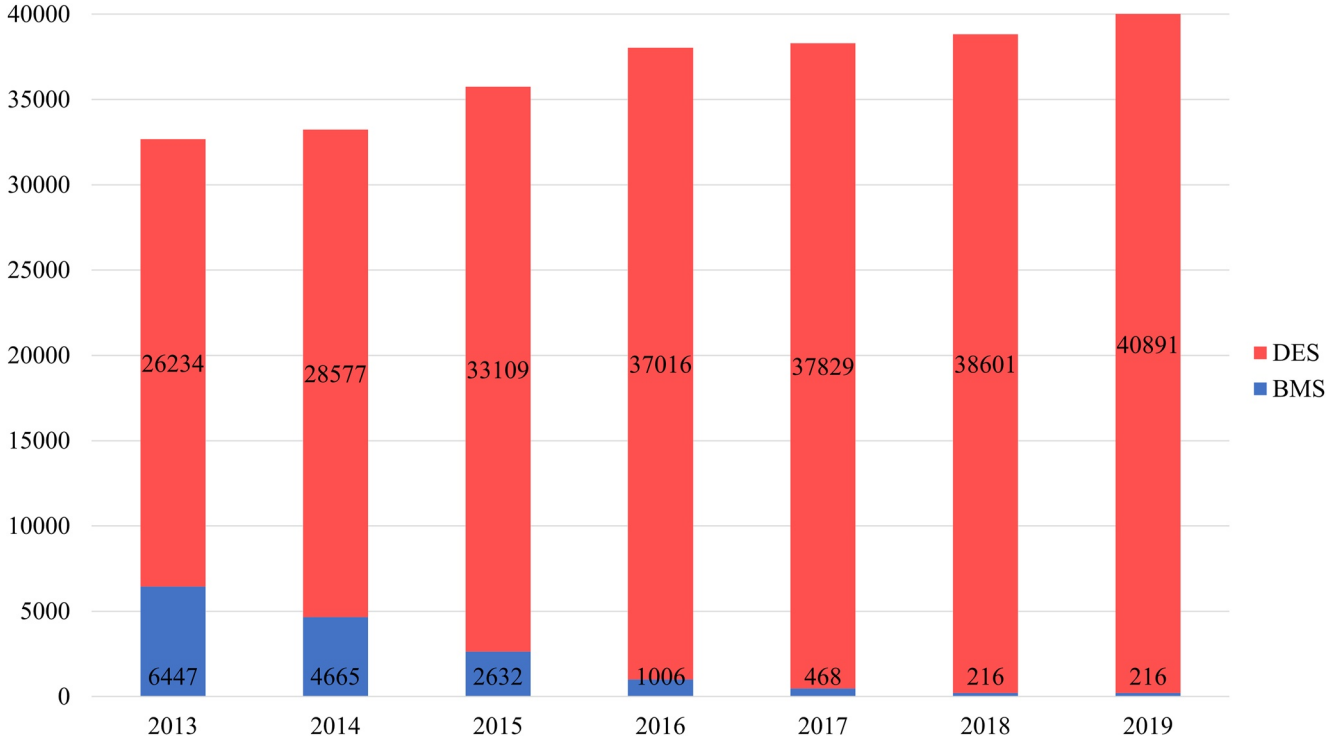

FIGURE 1 Percutaneous coronary stents with drug-eluting and bare metal stents between 2013 and 2019 ($N = 257{,}907$). The number of implanted stents is reported within or above bars [Colour figure can be viewed at wileyonlinelibrary.com]

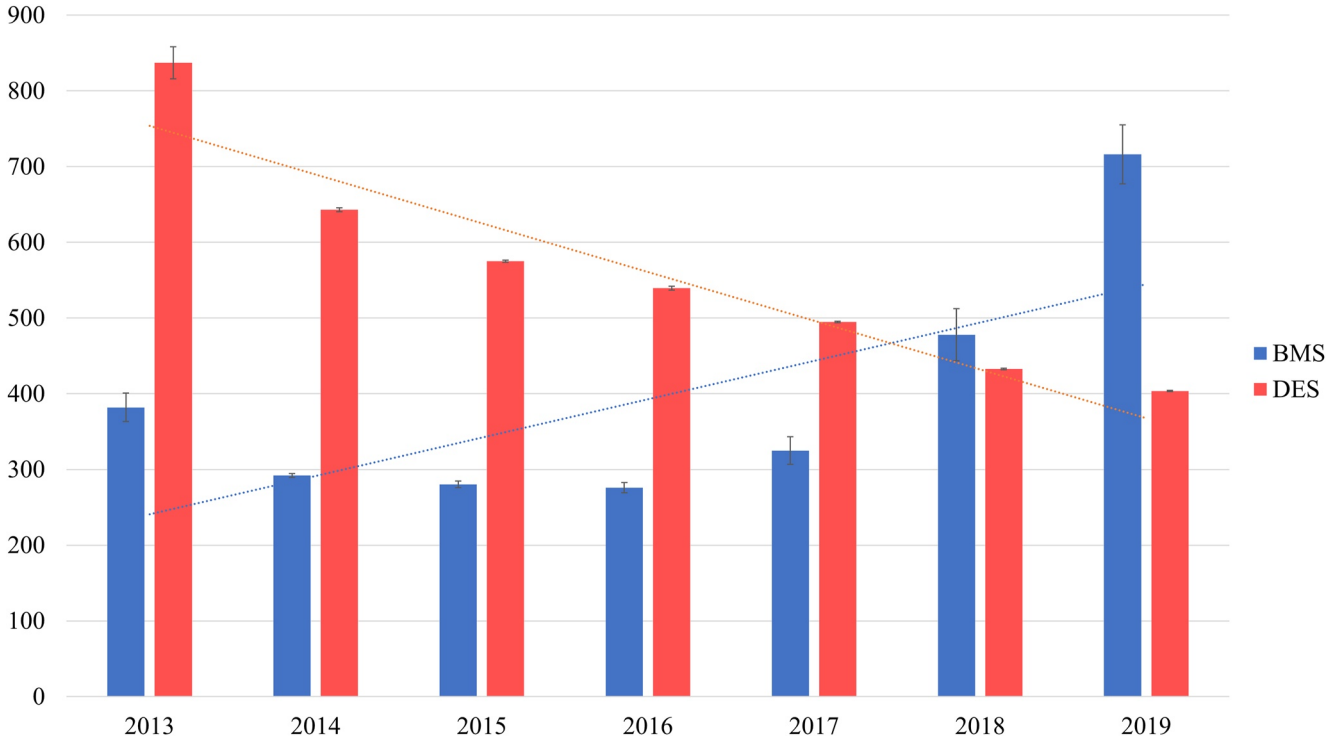

FIGURE 2 Average purchase cost of device (95% CI) by stent type and year [Colour figure can be viewed at wileyonlinelibrary.com]

The univariable Cox model presented much lower hazard ratios for all health outcomes. This suggests that DES and BMS patients differed in several clinical and demographic characteristics related to the considered outcomes. Once we accounted for main confounders, the HRs from the multivariable Cox model were much higher compared to univariable model, but still all

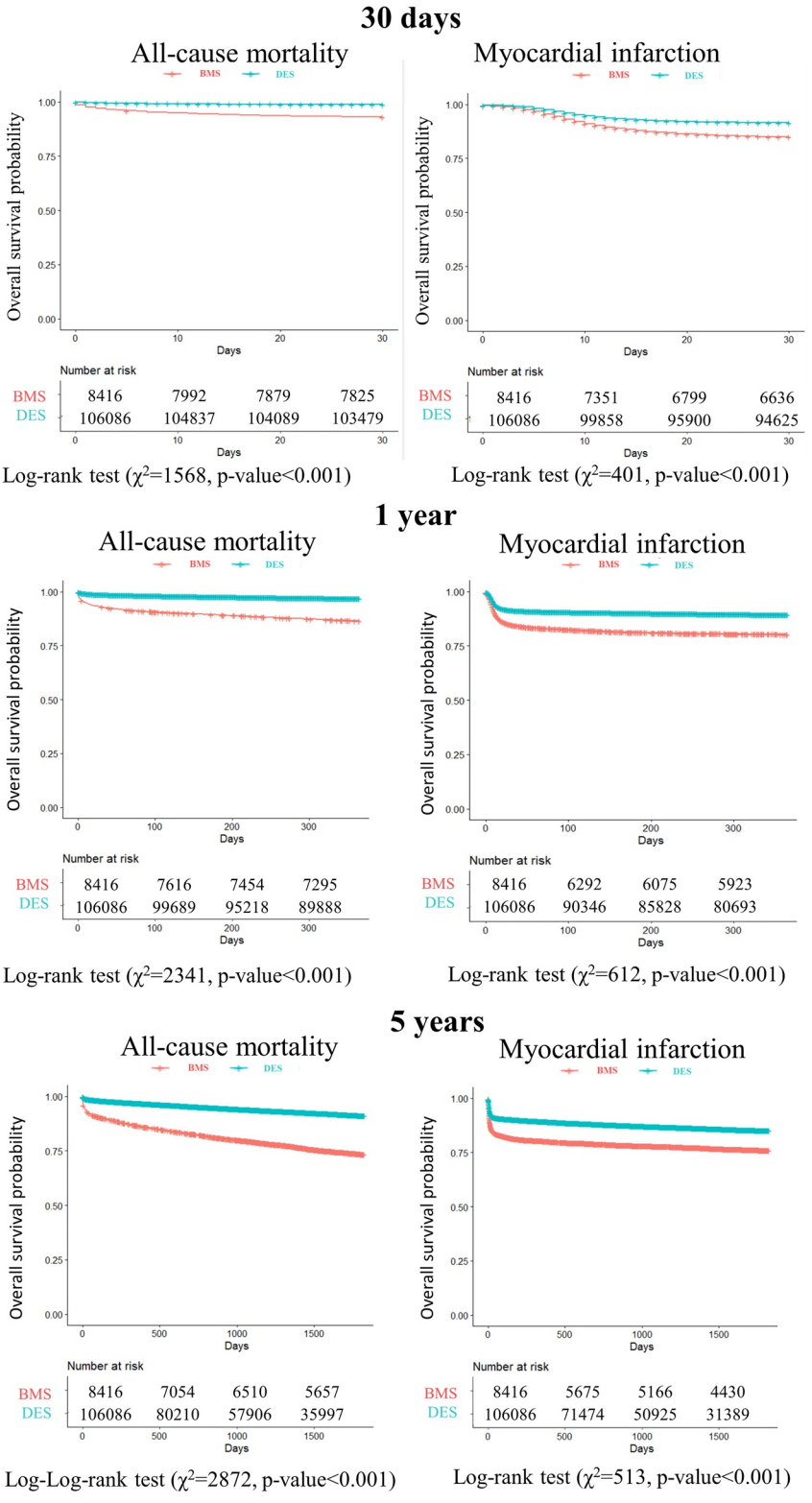

FIGURE 3 Univariate Kaplan-Meier survival curves of 30-day, 1-year and 5-year clinical outcomes after baseline percutaneous coronary intervention by stent type for the following endpoints: all-cause death (a), myocardial infarction (b). BMS = bare metal stent; DES = drug-eluting stent [Colour figure can be viewed at wileyonlinelibrary.com]

significantly lower than 1, indicating a largely reduced mortality risk for DES patients compared to BMS patients, as well as lower risks for myocardial infarction and cardiac mortality, for all follow up times.

Estimates of the marginal effects (i.e., outcome difference in populations identical in all respects but treatment), produced by IPTW, PSM and PS subclassification, ranged considerably, for all outcomes and follow up time points, depending on methods. Overall, PSM produced the marginal HRs closest to 1, suggesting that the matched sample, that included a much smaller group of treated, may be composed of more similar patients overall. The effect estimated here is an average treatment effect of the matched sample, which is neither an ATE nor an ATT, and despite being smaller compared to other results, still suggests

TABLE 2 Adjusted outcomes after percutaneous coronary intervention

	Hazard ratio (95% CI) 30 days	Hazard ratio (95% CI) 1 year	Hazard ratio (95% CI) 2 years	Hazard ratio (95% CI) 5 years
All-cause mortality				
Univariable cox model	0.18 (0.16–0.20)	0.22 (0.21–0.24)	0.25 (0.23–0.26)	0.29 (0.27–0.30)
Multivariable cox model	0.48 (0.43–0.54)	0.52 (0.48–0.57)	0.58 (0.54–0.62)	0.62 (0.59–0.66)
IPTW ATE marginal	0.40 (0.36–0.45)	0.40 (0.37–0.45)	0.44 (0.41–0.47)	0.50 (0.47–0.53)
IPTW ATE conditional	0.41 (0.36–0.46)	0.38 (0.35–0.41)	0.42 (0.39–0.45)	0.48 (0.45–0.51)
IPTW ATT marginal	0.38 (0.36–0.41)	0.37 (0.36–0.39)	0.41 (0.40–0.43)	0.47 (0.46–0.49)
IPTW ATT conditional	0.40 (0.32–0.50)	0.36 (0.26–0.48)	0.40 (0.31–0.50)	0.45 (0.37–0.55)
Adjustment using PS	0.45 (0.40–0.51)	0.51 (0.47–0.55)	0.55 (0.51–0.59)	0.59 (0.56–0.63)
PSM nearest 1:3 caliper 0.2 marginal[a]	0.53 (0.47–0.59)	0.56 (0.52–0.61)	0.60 (0.56–0.65)	0.64 (0.61–0.68)
PSM nearest 1:3 caliper 0.2 conditional[a]	0.55 (0.49–0.62)	0.58 (0.53–0.62)	0.62 (0.58–0.66)	0.65 (0.61–0.69)
PS Subclassification marginal	0.48 (0.42–0.55)	0.54 (0.49–0.59)	0.58 (0.53–0.62)	0.63 (0.59–0.67)
PS Subclassification conditional	0.55 (0.47–0.63)	0.58 (0.53–0.64)	0.62 (0.57–0.67)	0.66 (0.62–0.70)
Myocardial infarction				
Univariable cox model	0.55 (0.51–0.58)	0.52 (0.49–0.54)	0.54 (0.51–0.57)	0.57 (0.55–0.60)
Multivariable cox model	0.79 (0.74–0.85)	0.73 (0.69–0.78)	0.75 (0.71–0.79)	0.78 (0.74–0.82)
IPTW ATE marginal	0.69 (0.65–0.73)	0.63 (0.60–0.66)	0.64 (0.61–0.67)	0.67 (0.63–0.70)
IPTW ATE conditional	0.62 (0.59–0.66)	0.57 (0.54–0.60)	0.59 (0.56–0.62)	0.62 (0.59–0.65)
IPTW ATT marginal	0.67 (0.66–0.69)	0.61 (0.60–0.63)	0.62 (0.61–0.64)	0.65 (0.64–0.67)
IPTW ATT conditional	0.61 (0.50–0.75)	0.56 (0.47–0.66)	0.57 (0.48–0.68)	0.60 (0.52–0.71)
Adjustment using PS	0.80 (0.74–0.85)	0.73 (0.69–0.78)	0.75 (0.71–0.79)	0.77 (0.73–0.81)
PSM nearest 1:3 caliper 0.2 marginal[a]	0.83 (0.78–0.89)	0.77 (0.73–0.82)	0.79 (0.74–0.84)	0.81 (0.76–0.86)
PSM nearest 1:3 caliper 0.2 conditional[a]	0.83 (0.77–0.89)	0.77 (0.72–0.82)	0.78 (0.74–0.83)	0.80 (0.76–0.85)
PS Subclassification marginal	0.81 (0.76–0.87)	0.76 (0.71–0.81)	0.77 (0.73–0.82)	0.79 (0.74–0.83)
PS Subclassification conditional	0.82 (0.76–0.88)	0.76 (0.71–0.81)	0.77 (0.73–0.82)	0.79 (0.74–0.84)
Cardiac mortality				
Univariable cox model	0.18 (0.16–0.20)	0.21 (0.19–0.23)	0.23 (0.22–0.25)	0.26 (0.24–0.27)
Multivariable cox model	0.49 (0.43–0.55)	0.53 (0.49–0.59)	0.59 (0.54–0.64)	0.62 (0.58–0.67)
IPTW ATE marginal	0.41 (0.37–0.47)	0.48 (0.44–0.54)	0.52 (0.48–0.58)	0.57 (0.52–0.62)
IPTW ATE conditional	0.42 (0.34–0.52)	0.47 (0.40–0.55)	0.51 (0.43–0.59)	0.55 (0.48–0.62)
IPTW ATT marginal	0.40 (0.37–0.42)	0.47 (0.44–0.49)	0.51 (0.48–0.53)	0.55 (0.53–0.57)
IPTW ATT conditional	0.42 (0.33–0.52)	0.46 (0.38–0.55)	0.49 (0.42–0.58)	0.53 (0.46–0.61)
Adjustment using PS	0.46 (0.41–0.52)	0.51 (0.46–0.56)	0.55 (0.51–0.60)	0.58 (0.54–0.62)
PSM nearest 1:3 caliper 0.2 marginal[a]	0.49 (0.44–0.56)	0.54 (0.49–0.59)	0.59 (0.54–0.64)	0.63 (0.58–0.68)
PSM nearest 1:3 caliper 0.2 conditional[a]	0.52 (0.46–0.59)	0.54 (0.49–0.59)	0.58 (0.54–0.64)	0.62 (0.58–0.67)
PS Subclassification marginal	0.48 (0.42–0.55)	0.54 (0.48–0.60)	0.58 (0.53–0.64)	0.62 (0.57–0.67)
PS Subclassification conditional	0.55 (0.47–0.64)	0.59 (0.52–0.66)	0.63 (0.57–0.70)	0.65 (0.60–0.71)

Note: Hazard ratios (95% confidence intervals) for drug-eluting versus bare metal stents.

[a]Sample size = 24,899.

a remarkably lower mortality HR for DES patients as well as for myocardial infarction, although less markedly lower. On the other hand, the IPTW produced the lowest HRs (closest to 0) for both ATE and ATT marginal estimands. PS subclassification produced ATT marginal effects, that, compared to ATT marginal effects produced by IPTW, are much closer to 1, especially for MI (up to 14% points higher 0.67 (0.66–0.69) versus 0.81 (0.76–0.87)).

Estimates of conditional effects (i.e., average effect at the individual level of changing a subject's treatment status) were estimated by adjustment for PS and by PSM and PS subclassification adding the prognostic control variables included in the multivariable Cox model. Here too, results produced by PSM and PS subclassification were closer in magnitude compared to IPTW. For example, the HR for all-cause mortality at 30 days ranged from 0.40 (IPTW ATT), to 0.45 (PS adjustment), to 0.55 from PSM and PS subclassification.

Looking at follow-up time, for all-cause mortality, the DES patient advantage over BMS patients declined over time across all effects and estimands but remained significant even 5 years after baseline implantation. For the MI endpoint, the lower hazard of DES patients remained quite stable in magnitude, with a lower HR observed in the medium term (1 and 2 years follow-up). For cardiac mortality, the Fine and Gray's hazard model was also run for the univariable and multivariable Cox models, accounting for competing risks; results are presented in Table A2 of the Supplementary Material and are consistent with those shown in Table 2.

3.4 | Cumulative medical costs

We applied the KMSA on the 1:3 matched sample in a 5-year time frame, results are shown in Table 3. The DES group was found to have lower cumulative total costs, according to all methods (as low as 3264 Euros according to AS, or 2362 based on the CC method). Looking at single components of costs, the main driver of the lower costs of DES were less expensive hospitalizations, both at baseline -where the average length of stay was 9.3 days for BMS (8.6 sd) and 7.4 days for DES (7.1 sd)- and at follow-up. The average cost of the devices implanted at baseline and also those implanted at follow-up was higher for DES patients. Ambulatory costs at follow-up were very similar for the two groups, while the cost of medication use after the baseline visits related to percutaneous interventions was higher for patients receiving DES, as well as the cost of general medications. Another study (Cheng et al., 2019) found higher cumulative outpatient, inpatient, and cardiovascular diseases related outpatient costs for DES patients compared to BMS patients, and a lower cumulative inpatient cost related to cardiovascular disease.

4 | DISCUSSION

This paper aims to provide a proof-of-concept case study on the opportunities and challenges of using administrative data to inform the evaluation of medical devices. To do so, we exploited a privileged permission to access multiple administrative databases at the patient level of a major region in Italy, Lombardy, linked through an anonymized identification number. We used the case of coronary stent implantation to assess whether administrative data are potentially suitable for investigating costs and health outcomes of medical technologies and whether existing methodologies for observational data can be applied using such data sources. Our primary objective indeed was to assess the feasibility of using this data for informing comparative effectiveness studies and, more generally, health technology assessments, shedding the light on the challenges (as well as advantages) these sources of data may bring. To our knowledge, this is the first attempt of this kind in the literature.

Therefore, before we discuss our results, we must put front and center the challenges we incur and their implications. The nature of data we use in our analysis (large administrative database) doesn't allow us to be certain about the causality and perfect comparability between the two groups. This is intrinsic limitation to all sources of real-world data, and even more evident in routinely collected datasets for other purposes. Some of the challenges are clearly linked to and magnified by the technology chosen for example, the disproportionally higher number of patients receiving the treatment over time (significantly reduced in BMS group) makes the overlap in characteristics between DES and BMS patients particularly problematic in the analysis. This and other issues, however, are common, at different extents, across different case studies and observational studies (presence of unmeasured confounders, selection bias). It could be then questioned whether it is worth at all undertaking this type of analysis, and we strongly believe it is, for different reasons. We believe it was worthwhile comparing the BMS and DES using the administrative dataset in order to study the advantages and limitations of administrative data for evaluation purposes in the absence/scarcity of recent trials given that, although minimal, a proportion of patients still receive BMS. Whether we can call causal or not the association between the type of implanted stent and clinical endpoints, it is important to know whether a group of patients experience disadvantages to inform policy makers and clinicians so to put in place adequate interventions to reduce the gap (e.g., more frequent follow up visits). The other reason why we believe this research has important implications despite its limitations is that it provides insights on how to maximize the use of administrative databases, that are existing and available sources of information whose analysis implies no additional costs, and how to interpret obtained results, investigating appropriate methodologies suitable to face challenges that prevent establishing a causality link.

TABLE 3 Cumulative medical costs (EURO) estimated by Kaplan–Meier sample average method for the 1:3 matched sample

		Total	Baseline visit			Costs 5-year follow up				
Methods	**N**	**Total costs**	**Hospital 1st admission**	**Length of stay**	**Cost of device 1st implanted**	**Ambulatory costs**	**Re-admission**	**Stents subsequently implated**	**Medicines related to stent**	**Medicine general**
AS	BMS ($N = 8408$)	**25,436.1**	9276.28	9.29 (8.58)	624.45	532.93	12,435.02	183.76	533.41	1850.25
CC		**25,870.95**	9429.69		643.53	542.09	12,610.43	194.25	547.46	1903.5
BT		**26,006.14**	9256.31		632.63	568.04	12,766.94	202.50	576.84	2002.86
ZT		**26,103.26**	9316.70		639.37	568.04	12,766.94	205.47	576.84	2002.86
AS	DES ($N = 16,549$)	**22,171.97**	7906.48	7.53 (7.06)	1187.623	525.58	9566.61	219.74	709.15	2056.78
CC		**23,508.26**	8094.68		1253.275	566.20	10,349.24	241.84	735.48	2267.54
BT		**23,368.54**	7925.73		1255.582	585.71	10,241.3	249.66	761.16	2349.39
ZT		**23,543.78**	8027.65		1267.449	585.71	10,241.3	250.48	761.16	2349.39
AS	$Costs_{DES}$-$costs_{BMS}$	**−3264.13**	−1369.80		563.173	−7.35	−2868.41	35.98	175.74	206.53
CC		**−2362.69**	−1335.01		609.745	24.11	−2261.19	47.59	188.02	364.04
BT		**−2637.6**	−1330.58		622.952	17.67	−2525.64	47.16	184.32	346.53
ZT		**−2559.48**	−1289.05		628.079	17.67	−2525.64	45.01	184.32	346.53

Note: The bold indicates total costs include all costs considered in the study, corresponding to the sum of columns titled "Hospital 1st admission" + "Cost of 1st device implanted" + costs 5-year follow up (ambulatory costs + re-admission + stents subsequently implanted + medicine related stents + medicine general).

Abbreviations: AS, naive "Available Sample" estimator; BT, "Bang and Tsiatis's" estimator; CC, "Complete Case" estimator; ZT, "Zhao and Tian's" estimator.

The different methodologies applied in this study confirmed the advantage of DES use compared to BMS use, widely identified in observational studies and in RCTs for selected health outcomes, although some differences in magnitude of the results emerged. Comparisons with other studies are challenging, due to different eligibility criteria, methods applied, data sources and covariates used, but some correspondence between existing and our findings emerged. For example, the study by Iqbal et al (2016) that analyzed patients with acute STEMI who received either DES or BMS during 2008–2014 found substantially similar results (e.g., based on multivariable Cox model, the HR for DES vs. BMS in all-cause mortality was 0.41 at 30 days, 0.48 at 1 year and 0.53 at 2 years; based on IPTW ATE HRs for mortality at 30 days was 0.41, 0.48 at 1 year and 0.53 at 2 years). Ours and their study used different observational data sources and available control variables were different and so the capability of comparing the two groups; for example, they controlled for specific details on the procedures, that are often available in cardiac registries but not in administrative records. Some evidence based on previous years (1998–2007), and produced using administrative records, was also similar (Yeh, Chandra, McCulloch, & Go, 2011), with DES patients presenting significantly lower rates of death at 30 days (odds ratio (OR) based on PS adjustment 0.49, based on PSM 0.55) and 1 year (OR 0.58 and 0.70, respectively), as well as lower rate of MI at 1 year (OR 0.72 and 0.94 not significant).

A previous meta-analysis which considered studies published or presented up to February 2008 found that in observational studies DES were associated with significant reductions in mortality and MI (median follow-up 2.5 years); interestingly, compared to more recent evidence, while direction of findings is the same, the magnitude of the advantage was smaller (HR 0.78 and 0.87, respectively). Indeed, the comparison of existing evidence, even among studies which use similar sources of data, is challenging, because the timing (years) of implantation is particularly important for this medical technology. As shown in Figure 1, the percentage of BMS use has dramatically declined in past years, hence the use of BMS in the present may depend on different circumstances compared to a decade earlier. A recent study based on the National Inpatient Sample database (United States) explored which patients receive BMS in the era of second generation DES (2010–2014) and found that BMS was used in certain high-risk procedures and in patients with multiple comorbidity (Doshi et al., 2018), while this was not necessarily the case in earlier years. As a matter of fact, the ESC guidelines for the treatment of patients with a STEMI diagnosis since 2012 indicates that DES should generally be preferred over BMS unless there are specific contraindications. This is coherent with the higher DES advantage found in more recent studies and ours compared to previous findings. At the same time, it warns that the two treatment groups present different characteristics that may not be captured with the information available from administrative records and, as a consequence, it is particularly challenging to identify the treatment effect. Moreover, in this study we compare patients starting treatment with DES versus BMS and do not consider changes in treatment. In a context where guidelines recommend DES over BMS and DES use has become increasingly more common overtime, we may expect change in treatment due to clinical guidelines, rather than the effect of treatments (i.e., not because BMS was ineffective), and this may alter the representation of the treatments. Future research may be directed to explore precisely this aspect and assess changes in treatment strategy, applying methods such as inverse-probability-of-censoring weighting (IPCW) or the g-methods (Robins & Hernán, 2009) to deal with time-dependent confounding, to disentangle the effect of clinical guidelines and the effect of treatments.

When we compare RWE from our study, as well as from others, with RCT evidence, results are in line only for a subsample of endpoints. For all-cause and cardiac mortality, results differ, with no significant differences observed in RCTs between DES or BMS use, and reduced risk associated with DES use found in real-world observational studies (Kirtane et al., 2009; Piccolo et al., 2019; Suh et al., 2011). An attempt to explain the inconsistency of results was undertaken in the meta-analysis of Kirtane and colleagues (2009). They juxtaposed the generalizability and larger sample size which leads to more power for observational studies with their vulnerability to confounding that can be mitigated by multivariable adjustment and/or propensity matching, but not entirely ruled out. Indeed, as said above, this precisely applies to our setting, which, compared to other RWE such as that using cardiac registries, suffers particularly from vulnerability to omitted variables. Not all characteristics considered at baseline in RCT studies were available, and, for example, we had no information on smoking status, body mass index, lesion characteristics and procedural details. Moreover, the available data on medical history and comorbidities were not directly comparable with those collected in RCTs or even in registries for various reasons; at time of admission up to five secondary diagnoses are registered, and for patients affected by multiple conditions there may be conditions that are not documented (e.g., hypertension or diabetes), though this was at least partly attenuated using data from pharmaceutical use. The same applies to medical history, which we retrieved from previous admissions; but if a patient suffered from a disease and was not admitted or prescribed with drugs, we could not record his/her condition. This limited our ability to match/control and fully address the sample selection bias, as it is common for observational data sources. Considering that in recent years, BMS were used mostly in certain high-risk procedures and in patients with multiple comorbidity, this becomes a particularly crucial aspect. We echo the hypothesis proposed by Kirtane and colleagues (2009) and remark that trials and real-world populations are not comparable. Indeed, in real practice BMS have been implanted despite a wealth of data demonstrating the superiority of DES. A study in 2013 investigated why BMS

were still used despite that the superiority of DES was widely accepted; the authors prospectively collected data from 31 centers in Europe and Asia and identified the main reason for implantation of BMS rather than DES in 744 consecutive percutaneous coronary interventions performed in 2012 (Morice et al., 2013). Reasons for using BMS, in order of importance, were large vessel diameter, ST-segment elevation myocardial infarction, reimbursement or regulatory reasons, advanced age, concomitant oral anticoagulant treatment, increased bleeding risk, cancer, or anemia, planned noncardiac surgery within the next year, and anticipated poor dual antiplatelet therapy compliance. As evidence of that, the descriptive statistics shown in Table 1 indicates that BMS patients are older and a much larger proportion of them suffered from ST-segment elevation myocardial infarction compared to DES. As said above, not all factors identified by Morice et al. could be accounted for. Moreover, the fact that DES patients are more likely than BMS to be multimorbid (Table 1) may suggest that, as clinicians saw the benefit of DES over BMS, they may have applied the technology to a wider (and higher risk) population who are more likely to accrue a mortality benefit; this may explain the mortality benefit of DES patients in RWE, an effect not necessarily seen in RCTs.

Taking a step back, the debate on the usefulness and correctness of replicating RCT results in real-world settings is lively. A recent commentary (Thompson, 2020) refers to the phenomenon known as 'efficacy–effectiveness gap' existing between RCTs and RWE, to suggest that a discrepancy between the two may and should be expected. The value of RWE should be recognized distinctly from the purposes and research questions of RCTs; indeed, RWE can both complement RCTs and be used to make causal inferences (Olson et al., 2021). On the other hand, there can still be some value in the use of RWD to complement RCTs, because RWE can also be used to understand why results differ when there is disagreement, and efforts to replicate RCT results are to better understand the sources of variation that influence the ability to close the efficacy–effectiveness gap (Crown & Bierer, 2021). Our work did not aim to replicate results from existing RCT, but rather to test the usability of large administrative databases for assessing costs and health outcomes, which is a fundamental step to advance the ongoing debate on comparability of RWE and RCT. We assessed whether different methodologies developed for observational studies are applicable to our setting and how results differed, thereby allowing to test their internal validity and identification of best practices -one of the goals listed by Thompson (2020).

In discussing applicability of methods and their internal validity, two considerations emerged. The first is specific to our case study, the second is general. The comparison of costs and health outcomes between DES and BMS use in the past decade is a challenging one, in that we compare two technologies with mature recommendations for use, where DES have been established as the preferred technology for several years, while the use of BMS persists but in a very small proportion and particular cases; hence, the treated group (DES patients) is much larger than control group. This has methodological implications for the creation of a matching sample, for example, and has limited the matching methods applicable (e.g., pair matching requires that the sample of untreated subjects be larger than the sample of treated subjects (Austin, 2014)). Moreover, the exceptional use of BMS compared to the recommended use of DES makes the availability of clinical and procedural information particularly important to compare the two groups. This pushed us to maximize the available information on patients, but at the same time highlighted a limitation of the data source we used.

The general consideration pertains to the importance of establishing a priori the effect of primary interest to answer a study research question. When dealing with non-collapsible outcome measures, marginal versus conditional treatment effects differ, and this must be considered; for example, if the attempt is to replicate RCT results, marginal effects should be estimated. Then, which estimand, between ATE or ATT, especially when outcome is time-to-event in nature, is more relevant may depend on the research question; for instance, when interested in the effect of a treatment accessible to any patient, ATE may be preferable, while to assess the effect of a treatment/program for which only a subpopulation is eligible, ATT may be more interesting. Indeed, different propensity score methods may be more suited for estimating certain types of estimand, and effects are not directly comparable. For example, conditional effects estimated through PSM analysis may be conditional on the PS and the covariates included in the outcomes model, whereas the IPTW is conditional on the covariates in the weighted population. These conditional effects are not the same. Balancing out pros and cons of each method should also be done depending on the research setting and measure of effect of interest. All propensity score methods are implemented beginning with the estimation of the propensity score. Weighting may be more sensitive to misspecification of the propensity score (Austin, 2014). In a setting like ours, where the treatment and control groups are likely to differ on comorbidity and procedural risks that are not always measurable, the correct specification of propensity score may be particularly challenging for IPTW. The HRs estimated with both IPTW ATE and ATT were lower compared to HRs obtained from the other methods, and this may depend on that. The 0.5–99.5 quintile range of stabilized ATE weights went from 0.13 to 2.65, but the full range went up to 122. As additional analysis, we trimmed weights (results not included into the paper, but available upon request), and findings were sensible to such an approach (e.g., HR for all-cause mortality at 30 days based on ATE stabilized weight was 0.40 vs. HR = 0.43 with trimmed weights). On the other hand, using PSM with caliper and a 1:3 ratio, likely allowed us to estimate more homogeneous groups, but dramatically decreased the treated size and prevented estimating either ATE or ATT, as explained above. More generally,

when incomplete matching takes place, and our case study is an extreme example of that, if the unmatched treated subjects are very different from the matched treated, ATT estimates are likely to be biased (Austin, 2014). Our case study permitted identification and remark on the challenges of each method, and the RWD source we used made possible to perform the analysis, but also brought in some additional limitations.

Cost analysis represent one of the strengths of our study, in that we had access to high quality data on different categories of costs on a very large sample of patients, which are not commonly available, especially in Italy. A further unique feature was access to the "SDO4", the ad hoc dataset for implanted devices, which includes rich information on the device, including its cost. More generally, the characteristics of the data represent an asset in that they allow for long-term follow-up to track patients over time, and, as is typical of administrative databases, sample size is large. However, some limitations stem from the data that represent challenges in their use for comparative evaluation, as described above, that equally affect estimations of effects and comparisons of costs between the two groups.

To conclude, we believe that our work represents an important step that has demonstrated that the Lombardy administrative databases -and hence all comparable data sources-shows great potential to evaluate the real-world effectiveness and costs of medical technologies. However, there is still a lot that can and should be done in order to make these sources of data valuable for evaluation purposes. To maximize the use of such existing data sources for HTA purposes and beyond, it is crucial to understand their content in terms of endpoints, costs and possible controls. Having easily access to the metadata on the variables collected in the administrative databases or to similar document would be very useful to advance the use of these sources of data more rapidly with no additional costs. A reasonable investment could be made to include a minimum set of variables considered important to account for confounders in these databases and would represent a remarkable extension of the use of existing sources. In this study, we have shown that by accounting for confounding, results change significantly, and once this was done different methods produced consistent results. Obviously, we cannot rule out the risk of omitted confounders from observational data, but we believe that studies like this one will bring us a step closer to the use of RWE in HTA. This trend is very important bot for the scientific community as well as decision makers. Faced with challenges, researchers will continue to strive for methodological improvements to produce reliable evidence base. Policy makers, on the other hand, should recognize the potential of routinely collected data and adopt measures to overcome the limitations in access and the information collected. The final aim being that of generating value for the healthcare systems, by increasing returns on investments made to set up and run data collections at different levels.

ACKNOWLEDGMENT

We wish to thank Dr. Aldo Pietro Maggioni for clinical advice on the choice of codes for hospital admissions, outpatient appointments and medications as well as selection of clinical variables at baseline and representing patient's clinical history. We are grateful to Regione Lombardia for granting access to administrative databases of the Directorate General for Welfare, and to Aria S. p.a. for facilitating access to data, in particular to Antonio Barone, Giuseppe Preziosi, Alfredo Bevilacqua, Rosella dell'Orto and Giovanna Rigotti. We are very much thankful to the reviewers that helped us tremendously to improve the manuscript from its original version. This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 779306 for the project "Pushing the boundaries of cost and outcome analysis of medical technologies—COMED".

Open Access Funding provided by Universita Bocconi within the CRUI-CARE Agreement.

[Correction added on 06 October 2022, after first online publication: CRUI-CARE funding statement has been added.]

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data used in this study were made available by the Directorate General for Welfare of Lombardy Region, under a special license agreement with Bocconi University.

ORCID

Benedetta Pongiglione https://orcid.org/0000-0001-8539-1554

ENDNOTE

[1] 02B = Drug Eluting Coronary Stent (not supplied in service); 02H = Drug Eluting Coronary Stent (supplied in service with in-use haemodynamics); 02A = Coronary Stent Without Drug Delivery (not supplied in service); 02G = Non-drug-eluting Coronary Stent (supplied in service with in-use haemodynamics).

REFERENCES

Austin, P. C. (2011). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics*, *10*(2), 150–161.

Austin, P. C. (2013). The performance of different propensity score methods for estimating marginal hazard ratios. *Statistics in medicine*, *32*(16), 2837–2849.

Austin, P. C. (2014). The use of propensity score methods with survival or time-to-event outcomes: Reporting measures of effect similar to those used in randomized experiments. *Statistics in Medicine*, *33*(7), 1242–1258.

Austin, P. C. (2016). Variance estimation when using inverse probability of treatment weighting (IPTW) with survival analysis. *Statistics in Medicine*, *35*(30), 5642–5655.

Austin, P. C., & Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine*, *34*(28), 3661–3679.

Berger, M. L., Sox, H., Willke, R. J., Brixner, D. L., Eichler, H., Goettsch, W., Madigan, D., Makady, A., Schneeweiss, S., Wang, S. V., Watkins, J., Mullins, C. D., & Tarricone, R. (2017). Good practices for real-world data studies of treatment and/or comparative effectiveness: Recommendations from the joint ISPOR-ISPE special task force on real-world evidence in health care decision making. *Value in Health*, *20*(8), 1003–1008.

Bønaa, K. H., Mannsverk, J., Wiseth, R., Aaberge, L., Myreng, Y., Nygård, O., Nilsen, D. W., Klow, N. E., Uchto, M., Bendz, B., Stavnes, S., Bjornerheim, R., Larsen, A. I., Slette, M., Steigen, T., Jakobsen, O. J., Bleie, O., Fossum, E., & Trovik, T. (2016). Drug-eluting or bare-metal stents for coronary artery disease. *New England Journal of Medicine*, *375*(13), 1242–1252.

Børty, L., Brøndum, R., & Bøgsted, M. (2019). Ccostr: An R package for estimating mean costs with censored data. *Journal of Open Source Software*, *4*(41), 1593. https://doi.org/10.21105/joss.01593

Brilakis, E. S., Patel, V. G., & Banerjee, S. (2013). Medical management after coronary stent implantation: A review. *JAMA*, *310*(2), 189–198.

Cappellaro, G., Ghislandi, S., & Anessi-Pessina, E. (2011). Diffusion of medical technology: The role of financing. *Health Policy*, *100*(1), 51–59.

Charlson, M. E., Pompei, P., Ales, K. L., & MacKenzie, C. R. (1987). A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. *Journal of Chronic Diseases*, *40*(5), 373–383.

Chen, S., Rolfes, J., & Zhao, H. (2015). Estimation of mean health care costs and incremental cost-effectiveness ratios with possibly censored data. *STATA Journal*, *15*(3), 698–711.

Cheng, H., Chiou, L., Chen, T., Sung, S., Chen, C., & Lang, H. (2019). Real-world cost-effectiveness of drug-eluting stents vs. bare-metal stents for coronary heart disease—A five-year follow-up study. *Health Policy*, *123*(2), 229–234.

Chitkara, K., & Pujara, K. (2010). Drug-eluting stents in acute coronary syndrome: Is there a risk of stent thrombosis with second-generation stents? *The European Journal of Cardiovascular Medicine*, *1*(2), 20. https://doi.org/10.5083/ejcm.20424884.10

Corrao, G., & Mancia, G. (2015). Generating evidence from computerized healthcare utilization databases. *Hypertension*, *65*(3), 490–498.

Crown, W. H., & Bierer, B. E. (2021). Real-world evidence: Understanding sources of variability through empirical analysis. *Value in Health*, *24*(1), 116–117.

Degli Esposti, L., Perrone, V., Veronesi, C., Buda, S., & Rossini, R. (2018). All-cause mortality, cardiovascular events, and health care costs after 12 months of dual platelet aggregation inhibition after acute myocardial infarction in real-world patients: Findings from the platelet-aggregation inhibition: Persistence with treatment and cardiovascular events in real world (PIPER) study. *Vascular Health and Risk Management*, *14*, 383–392. https://doi.org/10.2147/vhrm.s162004

Doshi, R., Shah, J., Jauhar, V., Decter, D., Jauhar, R., & Meraj, P. (2018). Comparison of drug eluting stents (DESs) and bare metal stents (BMSs) with STEMI: Who received BMS in the era of 2nd generation DES? *Heart & Lung*, *47*(3), 231–236.

Epstein, A. J., Ketcham, J. D., Rathore, S. S., & Groeneveld, P. W. (2012). Variations in the use of an innovative technology by payer: The case of drug-eluting stents. *Medical Care*, *50*(1), 1. https://doi.org/10.1097/mlr.0b013e31822d5de9

Fine, J. P., & Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, *94*(446), 496–509.

Gansen, F. M. (2018). Health economic evaluations based on routine data in Germany: A systematic review. *BMC Health Services Research*, *18*(1), 1–11.

Gasparini, A. (2018). Comorbidity: An R package for computing comorbidity scores. *Journal of Open Source Software*, *3*(23), 648. https://doi.org/10.21105/joss.00648

Gray, B., & Gray, M. B. (2020). Package 'cmprsk'.

Greifer, N. (2022). Estimating effects after matching. Retrieved from https://cran.r-project.org/web/packages/MatchIt/vignettes/estimating-effects.html#after-pair-matching-with-replacement

Hernán, M. Á, Brumback, B., & Robins, J. M. (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*, (5), 561–570. https://doi.org/10.1097/00001648-200009000-00012

Hernán, M. A., & Robins, J. M. (2016). Using big data to emulate a target trial when a randomized trial is not available. *American Journal of Epidemiology*, *183*(8), 758–764.

Ho, D. E., Imai, K., King, G., Stuart, E. A. (2027). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*. URL Http://Gking.Harvard.Edu/Matchit, *42*(8). https://doi.org/10.18637/jss.v042.i08

Iqbal, M. B., Nadra, I. J., Ding, L., Fung, A., Aymong, E., Chan, A. W., Hodge, S., Robinson, S. D., Siega, A. D., & British Columbia Cardiac Registry Investigators. (2016). Long-term outcomes following drug-eluting stents versus bare metal stents for primary percutaneous coronary intervention: A real-world analysis of 11, 181 patients from the British columbia cardiac registry. *Catheterization and Cardiovascular Interventions*, *88*(1), 24–35.

Kirtane, A. J., Gupta, A., Iyengar, S., Moses, J. W., Leon, M. B., Applegate, R., Brodie, B., Hannan, E., Harjai, K., Park, S. J., Perry, R., Racz, M., Saia, F., Tu, J. V., Waksman, R., Lansky, A. J., Mehran, R., Stone, G. W., & Jensen, L. O. (2009). Safety and efficacy of drug-eluting and bare metal stents: Comprehensive meta-analysis of randomized trials and observational studies. *Circulation*, *119*(25), 3198–3206.

Kolh, P., Windecker, S., Alfonso, F., Collet, J., Cremer, J., Falk, V., ., & Hamm, C. (2014). 2014 ESC/EACTS guidelines on myocardial revascularization: The task force on myocardial revascularization of the European society of cardiology (ESC) and the European association for cardio-thoracic surgery (EACTS) developed with the special contribution of the European association of percutaneous cardiovascular interventions (EAPCI). *European Journal of Cardio-Thoracic Surgery*, *46*(4), 517–592.

Lin, D. Y., Feuer, E. J., Etzioni, R., & Wax, Y. (1997). Estimating medical costs from incomplete follow-up data. *Biometrics*, (2), 419–434. https://doi.org/10.2307/2533947

Makady, A., deBoer, A., Hillege, H., Klungel, O., & Goettsch, W. (2017). What is real-world data? A review of definitions based on literature and stakeholder interviews. *Value in Health*, *20*(7), 858–865.

Mazzali, C., Paganoni, A. M., Ieva, F., Masella, C., Maistrello, M., Agostoni, O., Scalvini, S., & Frigerio, M. (2016). Methodological issues on the use of administrative data in healthcare research: The case of heart failure hospitalizations in Lombardy region, 2000 to 2012. *BMC Health Services Research*, *16*(1), 1–10.

Mendlowitz, A., Croxford, R., MacLagan, L., Ritcey, G., & Isaranuwatchai, W. (2020). Usage of primary and administrative data to measure the economic impact of quality improvement projects. *BMJ Open Quality*, *9*(2), e000712. https://doi.org/10.1136/bmjoq-2019-000712

Morice, M., Urban, P., Greene, S., Schuler, G., & Chevalier, B. (2013). Why are we still using coronary bare-metal stents? *Journal of the American College of Cardiology*, *61*(10), 1122–1123.

Neumann, F., Sousa-Uva, M., Ahlsson, A., Alfonso, F., Banning, A. P., Benedetto, U., ., & Head, S. J. (2019). 2018 ESC/EACTS guidelines on myocardial revascularization. *European Heart Journal*, *40*(2), 87–165.

Olson, M., Kahler, K. H., & Rudolph, A. E. (2021). To replicate or not to replicate? Insights and interpretations from a randomized trial duplication initiative. *Journal of Comparative Effectiveness Research*, *10*(13), 953–956. https://doi.org/10.2217/cer-2021-0137

Packer, C., Simpson, S., & Stevens, A. (2006). International diffusion of new health technologies: A ten-country analysis of six health technologies. *International Journal of Technology Assessment in Health Care*, *22*(4), 419–428.

Piccolo, R., Bonaa, K. H., Efthimiou, O., Varenne, O., Baldo, A., Urban, P., Kaiser, C., Remkes, W., Raber, L., de Belder, A., van 't Hof, A. W. J., Stankovic, G., Lemos, P. A., Wilsgaard, T., Reifart, J., Rodriguez, A. E., Ribeiro, E. E., Serruys, P. W. J. C., Abizaid, A., & de Belder, A. (2019). Drug-eluting or bare-metal stents for percutaneous coronary intervention: A systematic review and individual patient data meta-analysis of randomised clinical trials. *The Lancet*, *393*(10190), 2503–2510.

Pongiglione, B., Torbica, A., Blommestein, H., deGroot, S., Ciani, O., Walker, S., Dams, F., Blankart, R., Mollenkamp, M., Kovacs, S., Tarricone, R., Drummond, M., & Kovács, S. (2021). Do existing real-world data sources generate suitable evidence for the HTA of medical devices in Europe? Mapping and critical appraisal. *International Journal of Technology Assessment in Health Care*, *37*(1). https://doi.org/10.1017/s0266462321000301

Robins, J., & Hernán, M. (2009). Estimation of the causal effects of time-varying exposures. In G. Fitzmaurice, M. Davidian, G. Verbeke, & G. Molenberghs (Eds.), *Advances in longitudinal data analysis* (pp. 553–599). Chapman & Hall.

Steg, P. G., James, S. K., Atar, D., Badano, L. P., Lundqvist, C. B., Borger, M. A., & Ducrocq, G. (2012). ESC guidelines for the management of acute myocardial infarction in patients presenting with ST-segment elevation: The task force on the management of ST-segment elevation acute myocardial infarction of the European society of cardiology (ESC). *European Heart Journal*, *33*(20), 2569–2619.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, *25*(1), 1. https://doi.org/10.1214/09-sts313

Suh, H. S., Song, H. J., Choi, J. E., Jang, E. J., Son, H. J., Lee, S. M., Kim, J. S., & Choi, D. (2011). Drug-eluting stents versus bare-metal stents in acute myocardial infarction: A systematic review and meta-analysis. *International Journal of Technology Assessment in Health Care*, *27*(1), 11–22.

Tarricone, R., Torbica, A., & Drummond, M. (2017). Challenges in the assessment of medical devices: The MedtecHTA project. *Health Economics*, *26*, 5–12. https://doi.org/10.1002/hec.3469

Therneau, T. M., & Lumley, T. (2015). Package survival. *R Top Doc*, *128*(10), 28–33.

Thompson, D. (2020). Replication of randomized, controlled trials using real-world data: What could go wrong?, *Value in health*, *24*(1), 112–115.

Van Der, W., Willem, M., & Geskus, R. B. (2011). Ipw: An R package for inverse probability weighting. *Journal of Statistical Software*, *43*(13), 1–23.

Wang, S. V., Schneeweiss, S., Berger, M. L., Brown, J., deVries, F., Douglas, I., Gagne, J. J., Gini, R., Klungel, O., Nguyen, M. D., Rassen, J. A., Smeeth, L., Sturkenboom, M., & Mullins, C. D. (2017). Reporting to improve reproducibility and facilitate validity assessment for healthcare database studies V1. 0. *Value in Health*, *20*(8), 1009–1022.

Yeh, R. W., Chandra, M., McCulloch, C. E., & Go, A. S. (2011). Accounting for the mortality benefit of drug-eluting stents in percutaneous coronary intervention: A comparison of methods in a retrospective cohort study. *BMC Medicine*, *9*(1), 1–9.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Pongiglione, B., & Torbica, A. (2022). How real can we get in generating real world evidence? Exploring the opportunities of routinely collected administrative data for evaluation of medical devices. *Health Economics*, *31*(S1), 25–43. https://doi.org/10.1002/hec.4562