



Bridging the gap between oncology clinical trials and real-world data: evidence on replicability of efficacy results using German claims data

Journal of **Comparative Effectiveness Research**

Marco Ghiani^{*,1} , Ulf Maywald², Thomas Wilke¹  & Bart Heeg³

¹IPAM e.V., Alter Holzhafen 19, Wismar, 23966, Germany

²AOK PLUS, Dresden, 01067, Germany

³Cytel, Weena 316-318, Rotterdam, 3012 NJ, The Netherlands

*Author for correspondence: Tel.: +49 030 2759 5023; marco.ghiani@ipam-wismar.de

Aims: Using German claims, the authors replicated the CHAARTED trial in metastatic hormone-sensitive prostate cancer. **Methods:** The authors identified metastatic hormone-sensitive prostate cancer patients replicating the inclusion/exclusion criteria of CHAARTED. Patients treated with docetaxel in combination with androgen deprivation therapy (ADT) at first line (docetaxel group) were compared with patients treated with ADT monotherapy (ADT mono group). After propensity score matching, overall survival was compared between the matched cohorts. **Results:** The authors included 441 patients. After propensity score matching, two equally sized matched cohorts of 74 patients each were compared in terms of overall survival. The hazard ratio (HR) was 0.71 (95% CI: 0.42–1.19), comparable to the HR in CHAARTED (HR: 0.72; 95% CI: 0.59–0.89). **Conclusions:** Using early comparative evidence from real-world data for regulatory and health technology assessment decisions is useful.

First draft submitted: 22 September 2021; Accepted for publication: 18 February 2022; Published online: 22 March 2022

Keywords: oncology • prostate cancer • real-world data

While randomized controlled trials (RCTs) are considered the gold standard for estimating the efficacy and safety of a treatment, the use of real-world evidence (RWE) for both regulatory and health technology assessment (HTA) purposes is gaining momentum [1–3]. RWE is useful for purposes such as historical control arms, extrapolation of outcomes using RWE on standard-of-care treatments, and use of comparative effectiveness/safety data in regions with early market introduction for HTA purposes in late-adopting regions. However, the lack of randomization in the real world hinders trust in RWE, limiting its impact in regulatory and HTA decisions [3].

Techniques such as propensity score matching (PSM) aim to reproduce the effects of randomization in real-world data (RWD). However, such techniques rely on the assumption that all patient characteristics relevant for treatment assignment are observable or can be effectively proxied in the data [4]. This assumption is particularly problematic in oncology, where treatment decisions and outcomes largely depend on factors typically not observable in RWD, such as disease staging, performance status, results of mutational testing or even patient preference. Claims data, in particular, are less likely to contain clinical measures compared with electronic health records and cancer registry databases [1].

Emulating completed RCTs following a ‘target trial’ approach as outlined by Hernán and Robins [5] can help calibrate RWD against trials and shed some light on the opportunities and limitations of real-world studies [6]. With a focus on US data, prior research has reached mixed conclusions regarding the ability of observational studies to reproduce results from RCTs [7–11]. Using cancer registry data from the USA, Kumar *et al.* [9] systematically replicated 141 randomized oncology trials and found consistent hazard ratios (HRs) for overall survival (OS) in up to 70% of the analyses. However, their study included placebo-controlled trials and comparisons of surgical and medical treatments, which are difficult to reproduce without bias in RWD [1,12], and the questions of when and

why RWD fail to replicate RCT results remain. Moreover, while RWD pose constraints in matching trial eligibility criteria, it is unclear whether the resulting discrepancies between trial and real-world populations would result in differential efficacy estimates in the absence of major effect modification. The authors of this study aimed to fill this knowledge gap by examining the validity of using RWD in support of RCT data for efficacy analyses in oncology.

The authors sought to mimic a randomized trial in oncology using German claims data. The CHAARTED trial [13,14] was an RCT in metastatic hormone-sensitive prostate cancer (mHSPC) comparing docetaxel in combination with androgen deprivation therapy (ADT) versus ADT monotherapy. In the CHAARTED trial, treatment with docetaxel+ADT was associated with significantly longer median OS compared with ADT alone in the intention to treat (ITT) population ($p = 0.0018$). The authors aimed to replicate OS results using RWD.

Methods

Study design & data source

The CHAARTED trial ($n = 790$) compared docetaxel in combination with ADT ($n = 397$) and ADT monotherapy ($n = 393$) [13,14]. To replicate the trial, the authors of the present study used data from AOK PLUS, German statutory health insurance covering around 3.6 million patients in the regions of Saxony and Thuringia. The authors implemented a new-user cohort study [1] using data from 2012 to 2019. Patients were followed longitudinally, and data were collected on outpatient pharmacy utilization, age, gender, date of death, inpatient and outpatient diagnoses and procedures (including inpatient administration of selected drugs).

This work was a non-interventional, retrospective study based on anonymized data. Ethical approval and informed consent from patients were not required, in accordance with German laws and the policies of the institutions assessing patient-level data.

Study cohort

CHAARTED included patients with prostate cancer (PC) with an elevated prostate-specific antigen (PSA) level, radiologic evidence of metastatic disease and an Eastern Cooperative Oncology Group (ECOG) performance status score of 0, 1 or 2. No prior adjuvant ADT was allowed, unless it lasted 24 months or less and progression had occurred more than 12 months after completion of therapy, or there was no evidence of progression and treatment had commenced within 120 days before randomization [13,14]. The full list of criteria adopted in the trial is reported in [Supplementary Table 1](#). Patients were randomly assigned to ADT alone or to combination therapy with ADT plus docetaxel.

To emulate this design, the authors identified all patients with at least one inpatient or outpatient diagnosis of PC between 1/1/2014 and 31/12/2019 (inclusion period allowing for at least 2 years of baseline period). To identify patients with metastatic disease, the authors included only patients who had at least one inpatient or outpatient diagnosis of a metastasis within 365 days from a PC diagnosis in the inclusion period (see [Supplementary Table 2](#) for the full list of diagnostic codes). They then identified adult patients with at least one outpatient prescription or inpatient administration of docetaxel or ADT in the inclusion period and after first diagnosis of metastasis ([Supplementary Table 1](#)). Index date was set as the earliest date of docetaxel or ADT prescription/administration. Patients were excluded if they had an ADT prescription before 120 days prior to index. Prior adjuvant ADT was allowed if it started 24 months or less prior to index and ended at least 12 months prior to index. Similar to the trial protocol, patients were dropped if they received radiation therapy 30 days prior to index or underwent major surgery in the 28 days prior to index. In addition, patients who received prescriptions of abiraterone, enzalutamide, cabazitaxel, mitoxantrone or estramustine concomitantly to the index medication were dropped. To allow for complete medical history over the baseline period, the authors further dropped patients who were not continuously insured in the 2 years prior to index date ([Supplementary Table 1](#)).

For each of the inclusion and exclusion criteria in the trial, [Supplementary Table 1](#) reports the corresponding criterion in the real-world study or provides justification for the absence of a corresponding criterion.

Exposure groups

Within the real-world cohort of mHSPC patients, the authors identified patients who were treated concomitantly with docetaxel and ADT (docetaxel+ADT group) and patients who were treated with ADT monotherapy (ADT mono group). In accordance with the trial, where patients were randomized within 4 months from starting ADT, the authors included in the docetaxel+ADT group all patients who were prescribed docetaxel at index and ADT within 120 days, or ADT at index and docetaxel within 120 days.

To enhance the comparability of the two groups, the authors controlled for several baseline characteristics available in claims data proxying for prognostic factors that might have influenced the decision to treat with docetaxel (see [Supplementary Table 2](#) for the exact definition of each variable and the respective codes): age (based on birth year); number of inpatient and outpatient visits in the 2 years pre-index; number of hospitalization days during baseline; indicators for history of liver, biliary, renal and cardiac disease; an indicator for pleural effusion during baseline; and the Charlson comorbidity index.

Study end points

The primary end point in CHAARTED was OS in the ITT population. This analysis parallels the ITT analysis in the target trials by comparing groups according to whether they received a prescription of therapy at baseline or not, regardless of whether individuals continue on the same treatment afterward [5].

Follow-up

The authors followed patients from date of first ADT or docetaxel prescription until date of death. To avoid immortal time bias, patients in the docetaxel+ADT group were followed starting from the latest date between first docetaxel prescription and first ADT prescription.

Statistical analysis

Patients in the docetaxel+ADT and ADT mono groups were first compared in terms of baseline characteristics. The authors reported means for continuous variables and percentages for categorical variables. Differences in mean baseline characteristics were tested using a t-tests on the coefficient of a univariate regression of the variable on the group indicator. The authors then performed an unadjusted comparison of OS between the docetaxel+ADT and ADT mono groups using a univariate Cox regression model to estimate HRs and 95% CIs. They plotted Kaplan–Meier curves and tested for the proportional hazard assumption using Schoenfeld residuals.

To enhance the comparability of the two RWD cohorts in terms of OS, the authors subsequently implemented 1:1 nearest-neighbor PSM with maximum caliper equal to 0.2 of the standard deviation of the logit of the propensity score (PS) [15]. The PS was estimated from a logistic regression of an indicator of treatment assignment on all baseline characteristics. The predictive quality of the model was assessed using the c-statistic. In addition, the authors performed the following sensitivity analyses: they performed 1:2 nearest-neighbor PSM, they implemented inverse propensity score weighting (IPSW) using stabilized weights, they compared the unmatched cohorts using a multivariate Cox regression controlling for all baseline characteristics and they compared the unmatched RWD cohorts using a multivariate Cox regression controlling for the PS.

All analyses were performed using Stata (version 14.2) and MS-SQL (version 17.9.1).

Results

Of 46,621 patients with ≥ 1 inpatient or confirmed outpatient diagnosis of PC in the inclusion period, 8918 had at least one inpatient or outpatient diagnosis of a metastasis within 365 days from a PC diagnosis ([Figure 1](#)). Of these, 4408 had at least one prescription or inpatient administration of docetaxel or ADT agent after first diagnosis of a metastasis and within the inclusion period. After implementation of the inclusion and exclusion criteria based on the CHAARTED trial criteria, the authors included 441 patients (median follow-up: 22.3 months). 75 patients were included in the docetaxel+ADT group (median age: 67 years) and 366 patients in the ADT mono group (median age: 76 years).

Comparing the baseline characteristics of patients in the two groups ([Table 1](#)) showed that mean age was significantly lower in the docetaxel+ADT group (66.5 years) compared with the ADT mono group (74.4 years; $p < 0.001$). In addition, patients in the docetaxel+ADT group had on average a lower number of outpatient visits during baseline (6.8) compared with patients in the ADT mono group (7.4; $p = 0.04$). Patients in the two groups did not differ significantly in terms of other baseline characteristics.

When doing an unadjusted comparison of the two unmatched groups in terms of OS ([Table 2](#)), median survival was not reached in the docetaxel+ADT RWD group, and it was 35.2 months in the ADT mono RWD group, with an HR of 0.68 (95% CI: 0.45–1.03) ([Table 2](#) & [Figure 2](#)). In comparison ([Table 2](#)), the CHAARTED trial included younger patients (median age: 64 years), and at a median follow-up of 53.7 months, the median OS was 57.6 months for the docetaxel+ADT arm versus 47.2 months for the ADT mono arm (HR: 0.72; 95% CI: 0.59–0.89).

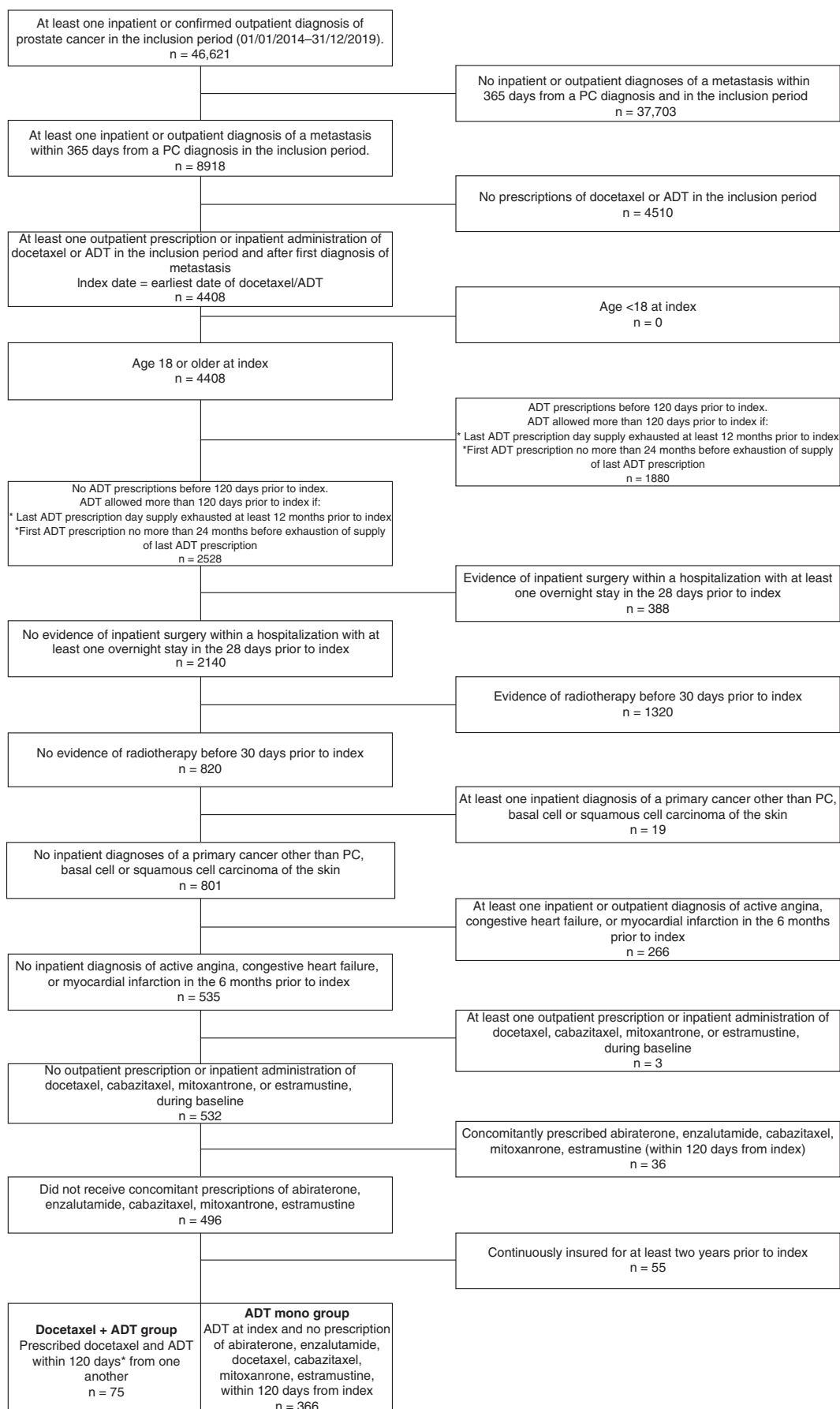


Figure 1. Patient attrition chart.
ADT: Androgen-deprivation therapy; PC: Prostate cancer.

Table 1. Comparison of baseline characteristics in matched and unmatched cohorts in real-world data.

Baseline characteristics	Mean		Bias		p-value
	Treated	Control	% bias	% reduction bias	
Age					
– Unmatched	66.507	74.448	-96.2		0.000
– Matched	66.703	66.149	6.7	93.0	0.635
CCI					
– Unmatched	9.56	10.055	-26.1		0.053
– Matched	9.5811	9.3649	11.4	56.3	0.464
Number inpatient visits					
– Unmatched	1.92	1.5902	21.7		0.077
– Matched	1.8784	1.7568	8	63.1	0.629
Number outpatient visits					
– Unmatched	6.8	7.3689	-26.3		0.038
– Matched	6.8243	6.1351	31.9	-21.2	0.100
Biliary disease					
– Unmatched	0.06667	0.06557	0.4		0.972
– Matched	0.06757	0.06757	0.0	100.0	1.000
Liver disease					
– Unmatched	0.05333	0.04372	4.5		0.716
– Matched	0.05405	0.04054	6.3	-40.5	0.701
Pleural effusion					
– Unmatched	0.01333	0.03005	-11.5		0.419
– Matched	0.01351	0.02703	-9.3	19.2	0.563
Renal disease					
– Unmatched	0.25333	0.28142	-6.3		0.419
– Matched	0.25676	0.21622	9.1	-44.3	0.563
Cardiac disease					
– Unmatched	0.08	0.10109	-7.3		0.576
– Matched	0.08108	0.02703	18.8	-156.3	0.148
Hospitalization days					
– Unmatched	14.547	11.74	13.5		0.211
– Matched	14.703	15.459	1.2	91.3	0.951
Boldface denotes statistical significance at the 95% CI. CCI: Charlson Comorbidity Index.					

Boldface denotes statistical significance at the 95% CI.

CCI: Charlson Comorbidity Index.

Table 2. Overall survival in the CHARTED trial and in German claims data, unmatched cohorts.

Group	CHAARTED Trial				German claims data (unmatched cohorts)			
	Median age	Median survival	Median follow-up	Hazard ratio (95% CI)	Median age	Median survival	Median follow-up	Hazard ratio (95% CI)
Androgen-deprivation therapy	63 years	47.2 months	53.7 months	0.72	76 years	35.2 months	22.3 months	0.68
Docetaxel	64 years	57.6 months		(0.59–0.89)	67 years	Not reached		(0.45–1.03)

Proportional hazard test (Schoenfeld residuals): $p = 0.551$.

After estimating the PS through a logistic regression (c -statistic = 0.8), patients in the two groups were matched using a 1:1 nearest-neighbor approach, resulting in two equally sized matched cohorts of 74 patients each. Patients in the matched cohorts did not differ in any of the baseline characteristics (Table 1), and the mean and median age in the docetaxel+ADT group (66.7 and 67 years, respectively) matched the mean and median age in the ADT mono group (66.1 and 65 years). In comparison, the median age of the CHARTED trial was 64 years (Table 3). In the matched RWD cohorts (Table 3), the median survival was not reached in the docetaxel+ADT RWD group, and it was 37.2 months in the ADT mono RWD group. The HR was 0.71 (95% CI: 0.42–1.19) (Figure 3 &

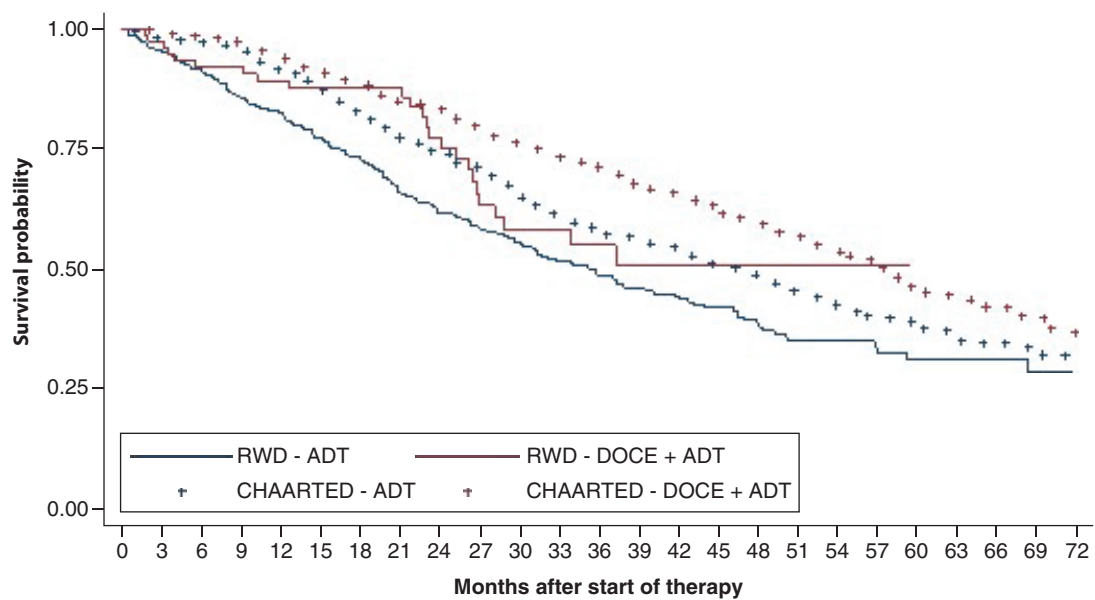


Figure 2. Overall survival in unmatched cohorts, Kaplan-Meier survival estimates.
RWD: HR (95% CI): 0.68 (0.45–1.03).
ADT: Androgen-deprivation therapy; DOCE: Docetaxel; HR: Hazard ratio; RWD: Real world data.
CHAARTED survival data [13] were obtained using Web Plot Digitizer (<https://automeris.io/WebPlotDigitizer/>).

Table 3. Overall survival in the CHAARTED trial and in German claims data, matched cohorts.

Group	CHAARTED Trial				German claims data (matched cohorts)			
	Median age	Median survival	Median follow-up	Hazard ratio (95% CI)	Median age	Median survival	Median follow-up	Hazard ratio (95% CI)
Androgen-deprivation therapy	63 years	47.2 months	53.7 months	0.72	65 years	37.2 months	22.4 months	0.71
Docetaxel	64 years	57.6 months		(0.59–0.89)	67 years	Not reached		(0.42–1.19)

Proportional hazard test (Schoenfeld residuals): p = 0.547.

Table 3). In comparison (Table 3), the median follow-up in the CHAARTED trial was 53.7 months, and the median OS was 57.6 months for the docetaxel+ADT arm versus 47.2 months for the ADT mono arm (HR: 0.72; 95% CI: 0.59–0.89).

Sensitivity analyses

First, the authors performed a 1:2 nearest-neighbor PSM with maximum caliper equal to 0.2 of the standard deviation of the logit of the PS [15]. The matched RWD cohorts included 74 patients in the docetaxel+ADT group and 104 in the ADT mono group, and they did not differ significantly in terms of baseline characteristics (Supplementary Table 3). In the matched cohorts (Supplementary Table 4), the median survival was not reached in the docetaxel+ADT group, and it was 37.9 months in the ADT mono group. The HR was 0.75 (95% CI: 0.44–1.25) (Supplementary Figure 1 & Supplementary Table 4).

Next, the authors implemented IPSW on the RWD using stabilized weights. One patient was dropped due to not being on the common support and the final RWD cohorts included 74 patients in the docetaxel+ADT group and 366 patients in the ADT mono group. Weighted cohorts did not differ significantly in any of the baseline characteristics, with the exception of age (Supplementary Table 5). To adjust for residual age differences after weighting, the authors included age as a covariate in the Cox regression. The median survival was 37.3 months in the docetaxel+ADT group and 37.0 months in the ADT mono group (Supplementary Table 6). The HR was 0.86 (95% CI: 0.52–1.43) (Supplementary Figure 2 & Supplementary Table 6).

Finally, the authors compared the unmatched cohorts using a multivariate Cox regression controlling for all baseline characteristics or, alternatively, the PS. The HR for death in the docetaxel+ADT group was 0.89 (95%

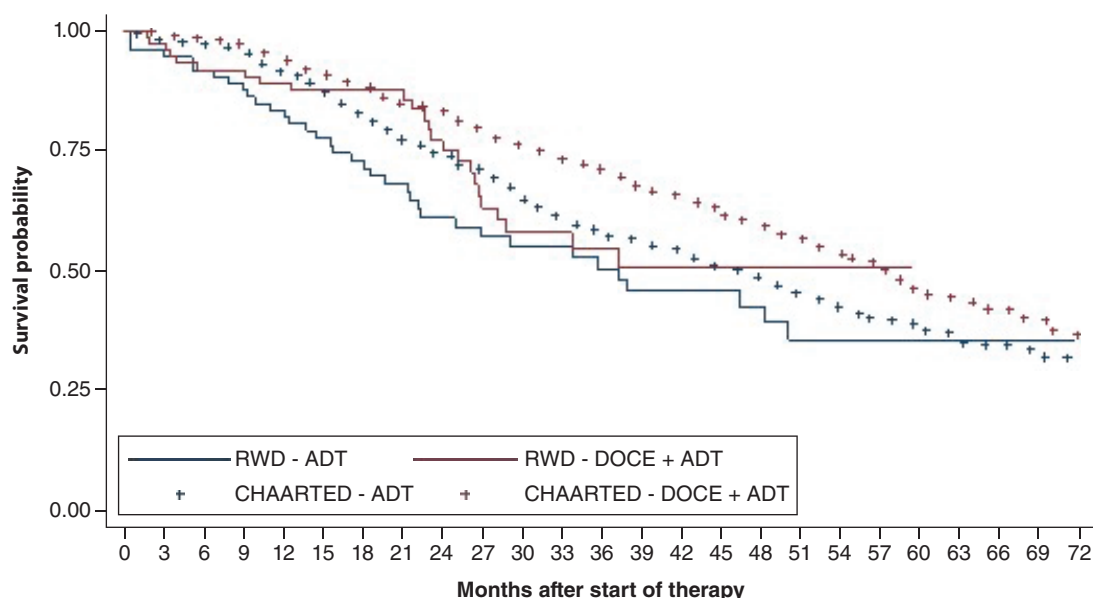


Figure 3. Overall survival in matched cohorts, Kaplan-Meier survival estimates.

HR (95% CI): 0.71 (0.42–1.19).

ADT: Androgen-deprivation therapy; DOCE: Docetaxel; HR: Hazard ratio; RWD: Real world data.

CHAARTED survival data [13] were obtained using Web Plot Digitizer (<https://automeris.io/WebPlotDigitizer/>).

CI: 0.56–1.40) when controlling for all baseline characteristics and 0.86 (95% CI: .55–1.34) when controlling for the PS.

Discussion

With the mandate of the 21st Century Cures Act, the US FDA established the RWE Program to foster the scientific evaluation of RWD in support of regulatory approvals and post-approval safety studies [16,17]. In Europe, the EMA formed the HMA/EMA Joint Big Data Task Force to establish a roadmap for the use of RWD in the regulatory assessment of drugs [18]. The framework for the FDA's RWE Program suggests that the replication of RCTs using RWD may provide insights into the opportunities and limitations of observational studies [17].

This work investigated the replicability of an oncology trial in mHSPC using German claims data. After implementation of inclusion and exclusion criteria closely aligned to the trial protocol, the real-world population included in the study was significantly older and had a shorter median OS compared with the trial population. However, the HR was concordant with the CHAARTED trial and, after adjusting for baseline characteristics with PSM, the point estimate from the RWD (0.71) was remarkably similar to the estimate in the trial (0.72). The results were robust to different types of adjustments (multivariate adjustment and IPSW) for baseline confounders, but PSM achieved the closest point estimate of the efficacy estimate in the trial. This is partially explained by the better balance in observable covariates achieved by PSM compared with IPSW and specifically the inability of weighting to balance cohorts in terms of age, with the weighted age in the treated group being significantly lower than the weighted age in the comparator group ($p = 0.011$). While in general PSM may result in greater unbalance compared with other methods [19,20], prior studies have shown that PSM can perform better than IPSW [21,22].

Prior studies focused primarily on electronic health records (EHRs) and registry data [9,23]. Compared with claims, EHRs and registries are more likely to provide clinical data and important confounders and thus represent a better source for comparative effectiveness research [1]. The current findings further contribute to the literature, providing an example on the replicability of RCTs in oncology using claims data, which are particularly disadvantageous for this purpose.

The existing literature reached mixed conclusions regarding the replicability of oncology trials using RWD, with some studies achieving the replicability of RCT results and other studies identifying significant discrepancies. The present results are in line with prior studies [7,24,25] that found largely similar results and sizes of treatment

effects between observational studies and RCTs and an association between the results of non-randomized versus randomized studies.

In contrast, a systematic MEDLINE search by Soni *et al.* [10] found that only 62% of the observational study HRs fell within the 95% CIs of the randomized trials. Kumar *et al.* [9] examined 141 RCTs covering eight tumor types. For each trial, they performed their own observational study using the National Cancer Database registry. Propensity-matched HRs for OS fell outside the trial CI 36% of the time. However, Kumar *et al.* included RCT for surgical procedures versus medical treatment, placebo control trials and phase II trials, which may be more difficult to replicate in the real world [1,12] and for which there might be good reasons for different effect estimates other than bias. Like Kumar *et al.*, this analysis found that confounding adjustment through PS can reduce bias and help replicate trial results.

The authors acknowledge that this analysis has some limitations. First, due to the stringency of the inclusion criteria adopted, the analytical sample in claims data was small, which reduced statistical power and led to an estimate non-significant at the 5% significance level. However, the proximity of the point estimate to the point estimate in the trial suggests that these results are concordant with the trial and are less precise merely due to sample size. Second, claims data do not provide important prognostic variables such as ECOG score, blood test results and PSA levels. Despite potential residual bias from unobservable confounders, after matching the authors were able to replicate RCT estimates of the treatment effect. These findings need further validation with additional data sources and larger sample sizes, as well as within different indications and treatments. If validated, these results suggest that after adjusting for observable confounders, claims data have the potential to emulate relative efficacy for OS from trial results, despite residual systematic differences between trial and real-world populations. In contrast, the lack of important prognostic factors impacts the estimate of absolute OS. Thus, these results confirm that the use of RWD to directly extrapolate absolute survival is not a proper endeavor. The use of claims for synthetic control arms in oncology is also discouraged based on these findings, while the use of RWD is encouraged to estimate relative efficacy in OS within Bayesian hierarchical network meta-analyses and to measure effectiveness in outcome-based reimbursement agreements. The results of this analysis indicate that using early comparative evidence from RWD sources for regulatory and HTA decision making it useful.

Summary points

- While real-world data (RWD) pose constraints in matching trial eligibility criteria, it is unclear whether the resulting discrepancies would result in differential efficacy estimates.
- This study aimed to examine the validity of using RWD in support of randomized controlled trial data for efficacy analyses in oncology. Specifically, this work investigated the replicability of an oncology trial in metastatic hormone-sensitive prostate cancer using German claims data.
- This study shows that after confounding adjustment, claims can emulate relative efficacy for overall survival (OS) from trial results, despite residual systematic differences between trial and real-world populations. In contrast, the lack of important prognostic factors impacts the estimate of absolute OS.
- While these results discourage the use of claims data for survival extrapolations and synthetic control arms in oncology, they encourage the use of RWD to estimate relative efficacy in OS. As such, using early comparative evidence from RWD sources for regulatory and health technology assessment decision making can be useful.

Supplementary data

To view the supplementary data that accompany this paper please visit the journal website at: www.futuremedicine.com/doi/suppl/10.2217/cer-2021-0224

Author contributions

Conception and design of the study: all; acquisition of the data: M Ghiani; analysis of the data: M Ghiani; interpretation of the data: all; drafting of the article: M Ghiani; review and finalization of the article: all. All authors have made substantial contributions to the study and writing of the article and have given their final approval of the version to be submitted.

Financial & competing interests disclosure

M Ghiani is a staff member of IPAM. U Maywald is an employee of AOK PLUS. T Wilke and B Heeg have received honoraria from several pharmaceutical/consultancy firms. No funding was received to conduct this research. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

References

Papers of special note have been highlighted as: ● of interest; ●● of considerable interest

1. Franklin JM, Schneeweiss S. When and how can real world data analyses substitute for randomized controlled trials? *Clin. Pharmacol. Ther.* 102(6), 924–933 (2017).
- **Addresses the issue of bridging the gap between real-world data and randomized controlled trials and provides a valuable overview to frame the efforts of the present manuscript.**
2. Oortwijn W, Sampietro-Colom L, Trowman R. How to deal with the inevitable: generating real-world data and using real-world evidence for HTA purposes – from theory to action. *Int. J. Technol. Assess. Health Care* 35(4), 346–350 (2019).
3. Sheffield KM, Dreyer NA, Murray JF, Faries DE, Klopchin MN. Replication of randomized clinical trial results using real-world data: paving the way for effectiveness decisions. *J. Comp. Eff. Res.* 9(15), 1043–1050 (2020).
4. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1), 41–55 (1983).
5. Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. *Am. J. Epidemiol.* 183(8), 758–764 (2016).
- **The theoretical framework for trial emulation.**
6. Schneeweiss S. Real-world evidence of treatment effects: the useful and the misleading. *Clin. Pharmacol. Ther.* 106(1), 43–44 (2019).
7. Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. *N. Engl. J. Med.* 342(25), 1878–1886 (2000).
8. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N. Engl. J. Med.* 342(25), 1887–1892 (2000).
9. Kumar A, Guss ZD, Courtney PT *et al.* Evaluation of the use of cancer registry data for comparative effectiveness research. *JAMA Netw. Open* 3(7), e2011985 (2020).
- **A prior attempt to investigate the replicability of oncology randomized controlled trial results using real-world data.**
10. Soni PD, Hartman HE, Dess RT *et al.* Comparison of population-based observational studies with randomized trials in oncology. *J. Clin. Oncol.* 37(14), 1209–1216 (2019).
- **Evidence on the replicability of oncology randomized controlled trial results using real-world data.**
11. Banerjee R, Prasad V. Are observational, real-world studies suitable to make cancer treatment recommendations? *JAMA Netw. Open* 3(7), e2012119 (2020).
12. Spoendlin J, Desai RJ, Franklin JM, Glynn RJ, Payne E, Schneeweiss S. Using healthcare databases to replicate trial findings for supplemental indications: adalimumab in patients with ulcerative colitis. *Clin. Pharmacol. Ther.* 108(4), 874–884 (2020).
13. Sweeney CJ, Chen Y-H, Carducci M *et al.* Chemohormonal therapy in metastatic hormone-sensitive prostate cancer. *N. Engl. J. Med.* 373(8), 737–746 (2015).
14. Kyriakopoulos CE, Chen YH, Carducci MA *et al.* Chemohormonal therapy in metastatic hormone-sensitive prostate cancer: long-term survival analysis of the randomized phase III E3805 CHARTED trial. *J. Clin. Oncol.* 36(11), 1080–1087 (2018).
15. Austin PC. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharm. Stat.* 10(2), 150–161 (2011).
16. Bonamici S. H.R.34–21st Century Cures Act. 114th Congress (2015–2016) [Internet] (2016).
17. US Food and Drug Administration. Framework for FDA's Real-World Evidence Program (2018). www.fda.gov/media/120060/download
18. Heads of Medicines Agencies. HMA/EMA Joint Big Data Steering Group (2020). www.hma.eu/509.html
19. King G, Nielsen R. Why propensity scores should not be used for matching. *Polit. Anal.* 27(4), 435–454 (2019).
20. Allan V, Ramagopalan SV, Mardekian J *et al.* Propensity score matching and inverse probability of treatment weighting to address confounding by indication in comparative effectiveness research of oral anticoagulants. *J. Comp. Eff. Res.* 9(9), 603–614 (2020).
21. Elze MC, Gregson J, Baber U *et al.* Comparison of propensity score methods and covariate adjustment: evaluation in 4 cardiovascular studies. *J. Am. Coll. Cardiol.* 69(3), 345–357 (2017).
22. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat. Med.* 28(25), 3083–3107 (2009).
23. Soni PD, Hartman HE, Dess RT *et al.* Comparison of population-based observational studies with randomized trials in oncology. *J. Clin. Oncol.* 37(14), 1209–1216 (2019).
24. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N. Engl. J. Med.* 342(25), 1887–1892 (2000).
25. Ioannidis JPA, Contopoulos-Ioannidis DG, Haidich AB *et al.* Comparison of evidence of treatment effects in randomized and nonrandomized studies. *J. Am. Med. Assoc.* 286(7), 821–830 (2001).