Reweighting Randomized Controlled Trial Evidence to Better Reflect Real Life – A Case Study of the Innovative Medicines Initiative

Michael Happich[1,*], Alan Brnabic[2], Douglas Faries[3], Keith Abrams[4], Katherine B. Winfree[3], Allicia Girvan[3], Pall Jonsson[5], Joseph Johnston[3], Mark Belger[1] and IMI GetReal Work Package 1

Evidence from randomized controlled trials available for timely health technology assessments of new pharmacological treatments and regulatory decision making may not be generalizable to local patient populations, often resulting in decisions being made under uncertainty. In recent years, several reweighting approaches have been explored to address this important question of generalizability to a target population. We present a case study of the Innovative Medicines Initiative to illustrate the inverse propensity score reweighting methodology, which may allow us to estimate the expected treatment benefit if a clinical trial had been run in a broader real-world target population. We learned that identifying treatment effect modifiers, understanding and managing differences between patient characteristic data sets, and balancing the closeness of trial and target patient populations with effective sample size are key to successfully using this methodology and potentially mitigating some of this uncertainty around local decision making.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?

☑ The limited external validity of randomized controlled trial (RCT) evidence poses a challenge for healthcare decision makers.

WHAT QUESTION DID THIS STUDY ADDRESS?

☑ Could an exploratory reweighting approach to generalizing RCT data to local real-world patient populations allow us to estimate the expected treatment benefit had the clinical trial been run in a broader real-world target population?

WHAT DOES THIS STUDY ADD TO OUR KNOWLEDGE?

☑ Identifying important treatment effect modifiers, understanding and managing differences between definitions of these patient characteristics in available data sets, and balancing the closeness of RCTs and target patient populations with the associated impact on the effective sample size available for analysis are key to successfully using this inverse propensity score reweighting methodology.

HOW MIGHT THIS CHANGE CLINICAL PHARMACOLOGY OR TRANSLATIONAL SCIENCE?

☑ Health technology assessment and regulatory decision making at the time of market authorization is ultimately executed under conditions of uncertainty. In certain settings, this reweighting approach could help to mitigate this uncertainty with respect to estimating the benefit of new interventions in real-world clinical practice.

The Innovative Medicines Initiative (IMI) is Europe's largest public–private partnership and aims to improve the drug development process by supporting more efficient discovery and development of better and safer medicines for patients. The IMI supports a number of collaborative research projects, among them GetReal, which aims to show how new methods for real-world evidence (RWE) collection and synthesis could be incorporated earlier into pharmaceutical research and development and healthcare decision making processes.[1]

It is widely accepted that large randomized controlled trials (RCTs) and meta-analyses of RCTs top the hierarchy of evidence for the efficacy of new pharmacological treatments. However, RCTs may have limited "external validity" (i.e., results may not be generalizable to the full range of patients likely to be treated with the new drug in real-world clinical practice).[2] An RCT is designed to evaluate the efficacy of a new treatment in a well-defined and controlled setting, where restrictive patient inclusion/exclusion criteria isolate the population in whom the benefit can most clearly be attributed to the treatment. This approach minimizes variability, narrows the confidence interval (CI), and maximizes "internal validity" (i.e., the confidence we can place in the cause and effect relationship) due to selection bias. Conversely, although observational studies are generally less susceptible to selection bias, RWE is nonrandomized, so no causal inference can typically be made

[1]Lilly Research Centre, Eli Lilly and Company, Surrey, UK; [2]Eli Lilly and Company, Sydney, New South Wales, Australia; [3]Lilly Corporate Center, Eli Lilly and Company, Indianapolis, Indiana, USA; [4]Department of Health Sciences, University of Leicester, Leicester, UK; [5]National Institute for Health and Care Excellence (NICE), Manchester, UK. *Correspondence: Michael Happich (happich_michael@lilly.com)

Received November 5, 2019; accepted March 31, 2020. doi:10.1002/cpt.1854

based on RWE alone. Therefore, in order for healthcare decision makers to make early postmarketing authorization access and reimbursement decisions at the national and/or local level, evidence from both RCTs and real-world observational studies is needed to address concerns regarding external validity.[3] However, RWE is rarely, if ever, available for new treatments at the time of regulatory approval/marketing authorization.

As a result, healthcare decision makers often question whether available RCT data reflect the patient population in their locality. In fact, many national health technology assessment (HTA) body guidance documents highlight the importance of and/or the need for the generalizability of data to local real-world patient populations.[4–8] Several approaches have been explored in the literature to address this question of generalizability to a target population, including comparing characteristics of RCT patients with those of real-world patients likely to be considered for treatment[2]; RCT subpopulation analyses relating to the region or country of interest; identification of treatment effect modifiers through subgroup analyses[9,10]; the confidence-profile method[11]; decision modeling[12]; graphical techniques to display the effect of population differences[13]; use of principle stratification[14]; or pragmatic trials.[15,16] However, potential limitations, including insufficient power, use of only a univariate approach, and/or operational challenges during implementation, lessen their applicability. Therefore, social science and public health research has more recently explored reweighting approaches to calibrate RCTs to be more reflective of target populations.[17–21]

In the context of regulatory decision making and HTA, we consider reweighting as a novel exploratory approach to generalizing RCT data to local real-world patient populations. The inverse propensity score (IPS) methodology described in this case study is used to reweight available RCT data based on existing observational baseline patient characteristic data from a target population. This allowed us to estimate the expected treatment benefit had the clinical trial been run in a broader real-world target population. It should be noted that this IPS methodology is well-established for addressing issues of confounding and has been published elsewhere.[22] Using the same concept in this setting, RCT outcomes of patients who are more representative of clinical practice receive higher weights, whereas RCT outcomes of patients who are less representative are discounted. Therefore, a weighted estimate of expected treatment benefit more reflective of the makeup of the real-world target population is generated. The advantage of this approach is that it only requires observational baseline patient characteristic data, which are more likely to be available at the time of regulatory approval/marketing authorization, not observational outcome data. It potentially allows us to answer the question "Would the trial results likely have been different if the patients enrolled in the RCT were more like 'local' patients?" This would allow healthcare decision makers to make early postmarketing authorization treatment access and reimbursement decisions based on what the results of an RCT might look like in the real-world population in their local area without having to wait for RWE or a trial in a local population.

Our objective is to illustrate this exploratory reweighting approach by applying this method to the JMDB trial, a pivotal phase III study that compared cisplatin plus pemetrexed with cisplatin plus gemcitabine as a first-line treatment for patients with advanced stage non-small cell lung cancer (NSCLC),[23] against a targeted real-world population derived using baseline characteristics from the FRAME observational study.[24,25] This illustrative example is not meant to imply any clinical meaning.

MATERIALS AND METHODS

Trial population

The pivotal JMDB trial was a noninferiority, phase III, randomized study that enrolled 1,725 chemotherapy-naïve patients aged ≥ 18 years with stage IIIB or IV NSCLC of any histology and an Eastern Cooperative Oncology Group (ECOG) performance status of 0 (fully active) or 1 (restricted in physically strenuous activity).[23] The primary objective was to compare the median overall survival (OS) of patients treated with pemetrexed plus cisplatin vs. gemcitabine plus cisplatin. Pemetrexed in combination with cisplatin is indicated for the first-line treatment of patients with predominantly nonsquamous NSCLC at a locally advanced or metastatic stage.[26] Therefore, for this illustration to be reflective of the licensed indication for pemetrexed, the included JMDB trial population was restricted by an additional eligibility criterion for this analysis: histologic or cytologic diagnosis of nonsquamous histology.

Target population

Observational baseline patient characteristic data for this illustrative case example were derived from the prospective, noninterventional, multicenter, observational FRAME study in patients aged ≥ 18 years.[24,25] The FRAME study was selected as the authors had access to patient-level data from the JMDB RCT and the real-world FRAME observational study, and both were Lilly-sponsored studies with consistent baseline patient characteristic variables available and defined. The FRAME target population included in this analysis was restricted by additional eligibility criteria to match the JMDB trial population in order for this illustration to be reflective of real-world patients with NSCLC to be treated with pemetrexed within the licensed indication: histologic or cytologic diagnosis of nonsquamous histology and ECOG performance status of 0 or 1.

Statistical methods

All variables common to and measured in the same way in both the JMDB RCT and the FRAME observational study data sets were included in the propensity score model. The list of included variables was verified by expert clinical opinion to be clinically relevant and to include important treatment effect modifiers. Interaction testing assessed differences between data sets using the *F*-test or median test for continuous variables and Fisher's exact test for categorical variables.

A singular numeric metric to simultaneously summarize all patient characteristics included in this analysis (i.e., a propensity score) was then used to denote the approximate probability of a patient being enrolled in the JMDB trial. First, patient data from the JMDB and FRAME studies were combined using identified common variables, creating an indicator to denote the membership of individual patients in each study (using 1 and 0 to denote JMDB or FRAME, respectively). Then a main effects logistic regression model was constructed using the membership indicator as the dependent variable and the other identified covariates as independent variables. Each patient's propensity score was calculated by applying the logistic model to the given covariates for the patient.

The propensity score distributions between patients in the JMDB and FRAME studies were then assessed. Histograms of the propensity scores from the two studies were plotted and the degree of overlap compared: the larger the overlapping region, the more evidence there was to support the generalizability of the JMDB trial. The propensity-adjusted balance between the two studies (JMDB and FRAME) was examined using standardized differences[27] (i.e., putting on a 0–1 scale to determine whether

the covariates in the JMDB RCT were similar to those in the FRAME observational study after reweighting). Differences were standardized by dividing them by their SD. Although there was no absolute rule to determine the magnitude of the standardized difference, differences between the cohorts were assumed if the standardized difference was large (i.e., > 0.1).

Each patient in the JMDB trial was then assigned a weight to attempt to match RCT patients with patients in the observational study. The weight for the ith patient was calculated using the formula $w_i = (1 - \hat{p}_i / \hat{p}_i)$, where $\hat{p}_i$ is the propensity score for patient i in the RCT, as described above. This is known as the IPS-weighting approach. Randomization was not affected by the weighting as the weights were applied independent of treatment assignment. The hazard ratio (HR) of OS was computed based on the weighted values for each patient in JMDB using a Cox proportional hazards regression model comparing the pemetrexed and gemcitabine arms. Variability and CIs were assessed by a nonparametric bootstrap procedure. One thousand bootstrap samples with replacement were created for each data set (using sample size from original data sets).[28] A distribution of HRs and associated CIs was obtained using the percentile approach (i.e., selecting the 2.5 and 97.5 percentiles of the bootstrap distribution). The effective sample size (ESS) after reweighting, which accounts for the number of patients actually contributing to the analysis based on the size of the weights, was also reported. The ESS was computed as the square of the summed weights divided by the sum of the squared weights to gauge the impact of reweighting on the available statistical information.[29] For the ith patient this can be represented as follows:

$$\mathrm{ESS} = \frac{\left(\sum_i w_i\right)^2}{\sum_i w_i^2}.$$

All analyses were conducted using SAS version 9.4 (SAS Institute, Cary, NC).

Post hoc sensitivity analyses

Three *post hoc* sensitivity analyses were undertaken: (i) fixed value trimming was performed to reduce the impact of patients with extreme weight values ≥ 4 (e.g., patients who would inflate the variance in the analysis) on the final outcome, prior to computing the HR of OS as described in Statistical methods; (ii) the critical covariate (i.e., main driver of difference between the JMDB RCT and FRAME observational study patient populations) was excluded as a baseline covariate from the original logistic regression model to allow for the potential lack of complete overlap between the JMDB trial and FRAME target patient populations; (iii) an alternative entropy balancing weighting algorithm[30] was applied to assess the robustness of the IPS-weighting results.

RESULTS

The number of patients in the JMDB trial (n = 1,209) and the FRAME study (n = 948) target populations following application of the eligibility criteria for this case study illustration are outlined in **Figure 1**.

In total, 15 variables were available, considered informative for balancing patient characteristics between the JMDB trial and FRAME target populations, and included in the propensity score weighting: (i) age (mean (SD) years and ≥ 70 years vs. < 70 years); (ii) sex (female vs. male); (iii) race (non-Asian vs. Asian); (iv) smoking status (current smoker vs. ex-smoker, current smoker vs. never smoker); (v) basis for diagnosis (cytologic or histopathologic); (vi) time since diagnosis of NSCLC at study entry (mean (SD) months and ≤ 1 month vs. > 1 month); (vii) diagnosis subtype (adenocarcinoma vs. large cell carcinoma); (viii) stage of disease at study entry (IIIB vs. IV); (ix) ECOG performance status (0 vs. 1); (x) number of metastatic sites (0–1, 2, or ≥ 3); (xi) prior surgery (yes or no); (xii) prior radiotherapy (yes or no); (xiii) presence of cardiovascular condition (yes or no); (xiv) presence of lung condition (yes or no); and (xv) diabetes (yes or no). However, definitions of "number of metastatic sites" differed between the two target populations: the JMDB RCT categorized patients as having 1, 2, 3, 4, or ≥ 5 metastatic sites, whereas the FRAME observational study categorized patients as having 0, 1, 2, or 3 metastatic sites. To maximize the comparability on this variable between JMDB RCT and FRAME observational study patients, data were organized into three categories for each population: 0–1, 2, and ≥ 3 metastatic sites.

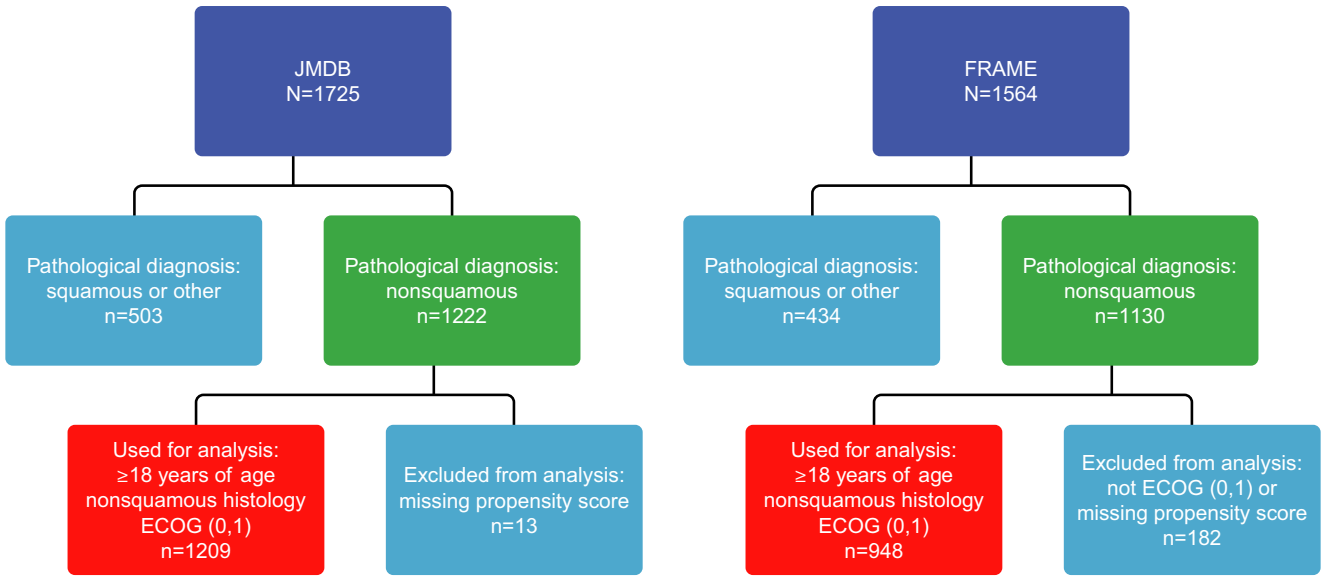

Figure 1 Flowcharts for JMDB randomized controlled trial and FRAME observational study patient populations following application of eligibility criteria for this case study illustration. ECOG, Eastern Cooperative Oncology Group.

Table 1 compares the identified baseline characteristics of JMDB RCT and FRAME observational study patients included in this case study illustration. Baseline characteristics were statistically significantly different ($P < 0.05$) between studies except for sex, time since diagnosis of NSCLC at study entry > 1 month, stage of disease at study entry, and prior surgery. The majority of patients in the FRAME observational study had 0–1 metastatic sites; whereas > 50% of patients in the JMDB RCT were sicker, with ≥ 3 metastatic sites.

Figure 2 shows the distribution of propensity scores for patients in both the JMDB RCT and the FRAME observational study. Although the JMDB and FRAME propensity score distributions clearly overlapped, it is apparent the populations were substantially different (i.e., many patients in the RCT with few equivalents in the real-world observational study and *vice versa*), with scores for JMDB trial patients skewed toward higher propensity scores, likely driven by pronounced differences in the number of metastatic sites between studies (**Table 1**).

Figure 3 shows the propensity-adjusted balance between the two studies (JMDB and FRAME) in a standardized difference plot for differences between studies before and after IPS weighting. Following reweighting, standardized differences of the baseline characteristics are below or close to 0.1, indicating that the weights allow for balance between the baseline characteristics of patients

Table 1 Baseline characteristics of JMDB randomized controlled trial (*N* = 1,209) and FRAME observational study (*N* = 948) patients included in this case study illustration

	JMDB (*N* = 1209)	FRAME (*N* = 948)	*P* value[a]
Age in years, mean (SD)	59.7 (9.3)	62.3 (9.9)	< 0.001
Age ≥ 70 years, *n* (%)	166 (14)	243 (26)	< 0.001
Female, *n* (%)	405 (34)	293 (31)	0.211
Non-Asian, *n* (%)	997 (83)	930 (98)	< 0.001
Smoking status, *n* (%)			< 0.001
Current smoker	277 (23)	297 (31)	
Ex-smoker	585 (48)	484 (51)	
Never smoker	195 (16)	121 (13)	
Unknown	152 (13)	46 (5)	
Basis for diagnosis, *n* (%)			< 0.001
Cytologic	453 (38)	242 (26)	
Histopathologic	756 (63)	706 (75)	
Time since diagnosis of NSCLC at study entry in months, mean (SD)	1.9 (7.8)	2.8 (12.6)	< 0.001
Time since diagnosis of NSCLC at study entry, > 1 month, *n* (%)	403 (33)	342 (36)	0.186
Diagnosis subtype, *n* (%)			0.005
Adenocarcinoma	861 (71)	725 (77)	
Large-cell carcinoma	145 (12)	77 (8)	
Other	203 (17)	146 (15)	
Stage of disease at study entry, *n* (%)			0.676
IIIB	272 (23)	206 (22)	
IV	937 (78)	742 (78)	
ECOG performance status, *n* (%)			< 0.001
0	446 (37)	275 (29)	
1	763 (63)	673 (71)	
Number of metastatic sites, *n* (%)			< 0.001
0–1	288 (24)	771 (81)	
2	296 (25)	157 (17)	
≥ 3	625 (52)	20 (2)	
Prior surgery, *n* (%), yes	94 (8)	92 (10)	0.122
Prior radiotherapy, *n* (%), yes	63 (5)	111 (12)	< 0.001
Presence of cardiovascular condition, *n* (%), yes	723 (60)	390 (41)	< 0.001
Presence of lung condition, *n* (%), yes	738 (61)	117 (12)	< 0.001
Diabetes, *n* (%), yes	78 (7)	107 (11)	< 0.001

ECOG, European Cooperative Oncology Group; NSCLC, non-small cell lung cancer.
[a]*F*-test or median test for continuous variables; Fisher's exact test for categorical variables.

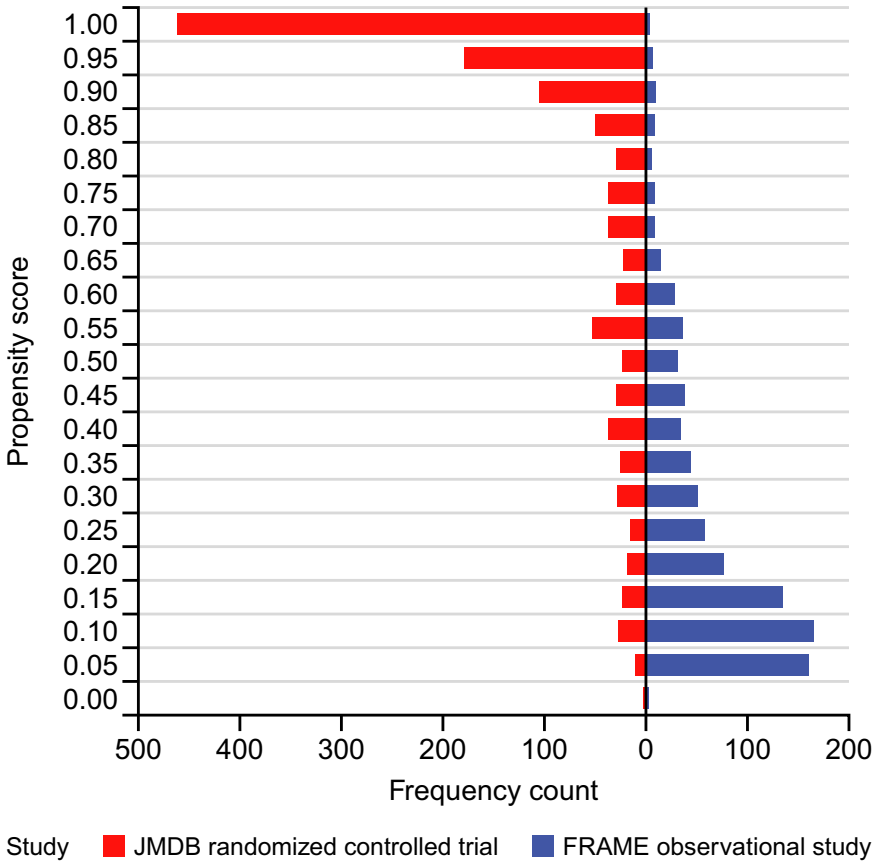

Figure 2 Distribution of propensity scores for patients in JMDB randomized controlled trial and FRAME observational study (unstandardized untrimmed primary analysis).

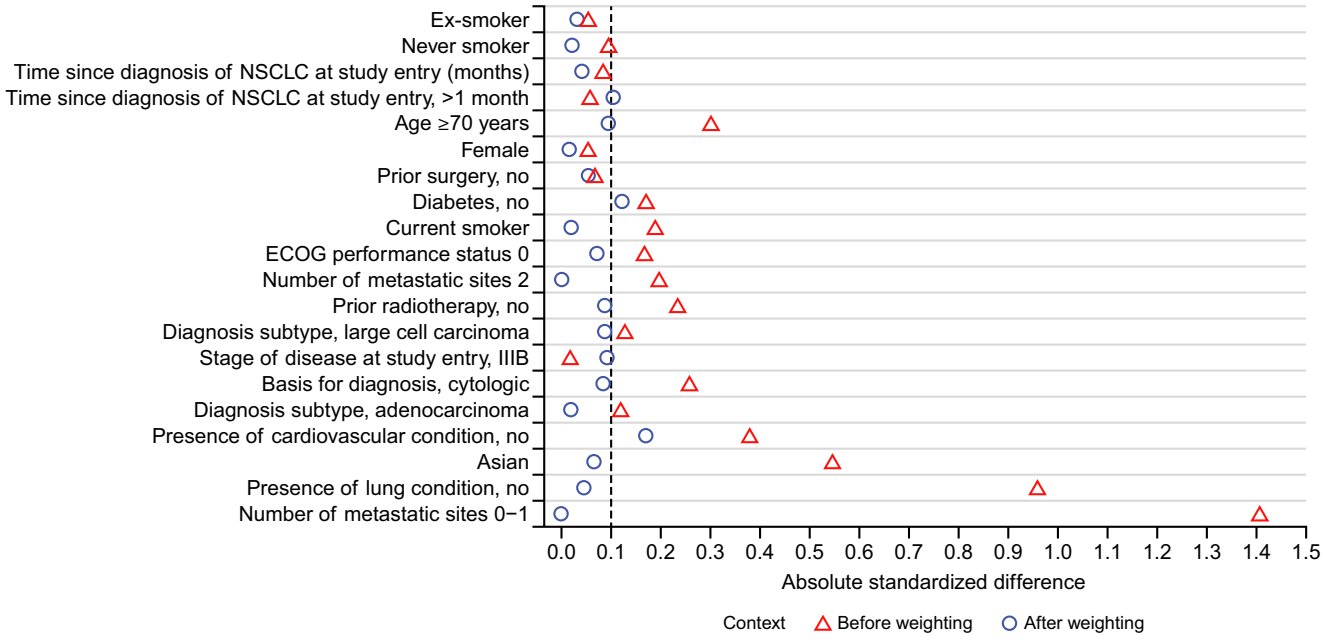

Figure 3 Standardized difference plot for original unweighted and inverse propensity score-weighted differences between studies (JMDB randomized controlled trial vs. FRAME observational study). Standardized difference plot ordered by magnitude of difference between JMDB randomized controlled trial vs. FRAME observational study before and after inverse propensity score weighting. ECOG, Eastern Cooperative Oncology Group; NSCLC, non-small cell lung cancer.

Table 2 Original unweighted and corresponding inverse propensity score-weighted HR of overall survival results for pemetrexed arm (relative to gemcitabine arm) of JMDB randomized controlled trial population included in this case study illustration

Primary analysis	HR	Bootstrap 2.5 percentile	Bootstrap 97.5 percentile
Original unweighted analysis[a] (n = 1,222)	0.85	0.75	0.97
Inverse propensity score-weighted analysis (ESS = 126)	0.91	0.62	1.31

ESS, effective sample size; HR, hazard ratio (pemetrexed [n = 608] vs. gemcitabine [n = 614]).
[a]Differences compared with Scagliotti *et al.*[23] are because a slightly different patient population was included in this case study illustration as a result of eligibility criteria being applied.

in the two studies. IPS-weighted differences between studies were vastly improved for some variables, such as number of metastatic sites, presence of lung condition, non-Asian, and presence of cardiovascular condition, compared with original unweighted differences between studies.

Table 2 displays the original unweighted HR of OS results for the pemetrexed arm of the JMDB trial population included in this case study illustration compared with the gemcitabine arm, alongside the corresponding IPS-weighted results. The IPS-weighting method resulted in a slightly smaller effect for pemetrexed compared with gemcitabine (HR of OS 0.91; 95% CI 0.62–1.31) than the original unweighted analysis (HR of OS 0.85; 95% CI 0.75–0.97) with greater associated uncertainty (indicated by a wider CI), which was not statistically significant. The IPS-weighted ESS was only 10% of the original unweighted sample size, which explains why the HR estimate was no longer statistically significant after reweighting.

Figure 4 shows the original unweighted Kaplan–Meier curves for the pemetrexed and gemcitabine arms of the JMDB trial population included in this case study illustration (**Figure 4a**) alongside the original unweighted and corresponding IPS-weighted predicted Cox proportional hazards model survivor functions (**Figure 4b,c**, respectively).

Post hoc sensitivity analyses

Results of the first two *post hoc* sensitivity analyses are shown in **Table 3**. In the first sensitivity analysis, trimming weights ≥ 4 to avoid giving large weight to one or two outlying individuals resulted in a slightly less uncertain IPS-weighted HR of OS than for the primary analysis (indicated by a narrower CI), as trimming increased the ESS by removing extreme weights.

In the second sensitivity analysis, number of metastatic sites was identified as the main driver of difference between the JMDB RCT and FRAME observational study patient populations on the basis of differences in baseline characteristic definitions between the two data sets (as described above); and limited overlap on this variable between the JMDB RCT and FRAME observational study patients (as observed in **Figure 3**). Excluding the number of

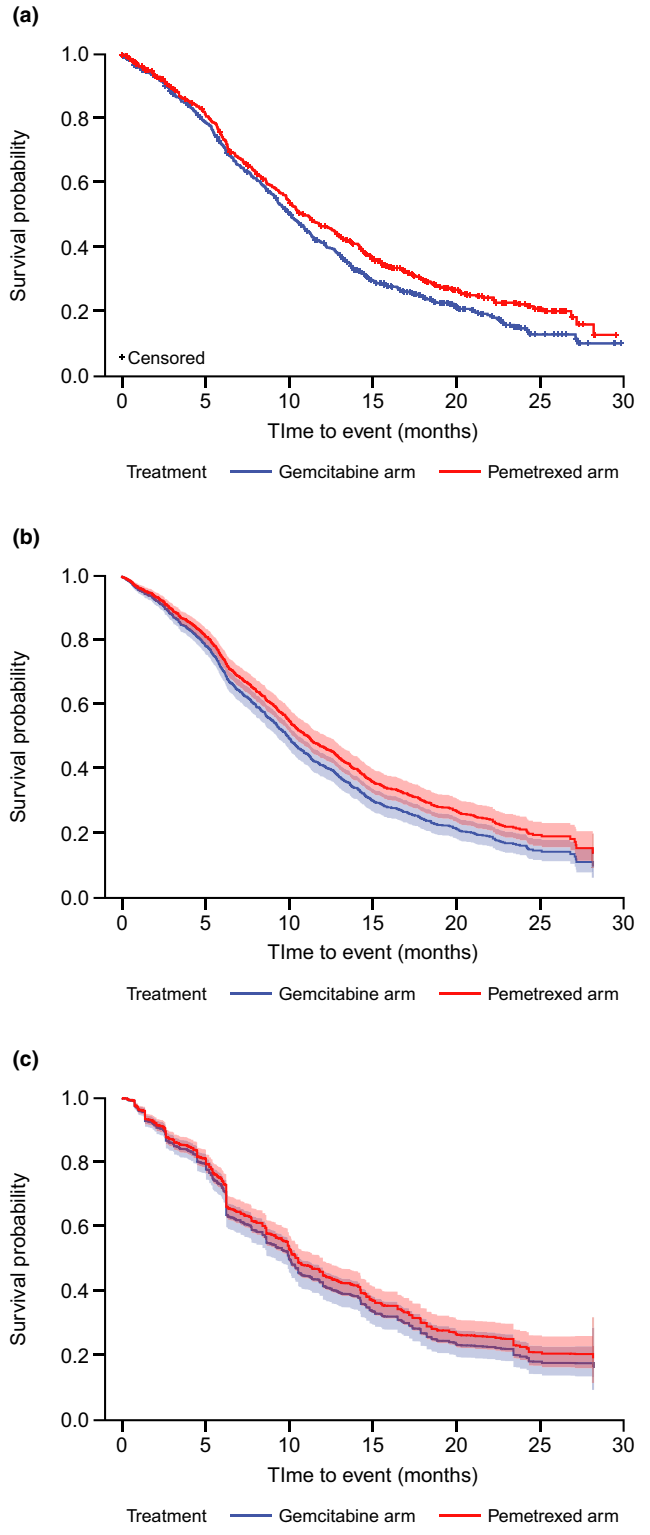

Figure 4 Survival curves for pemetrexed and gemcitabine arms of JMDB randomized controlled trial population included in this case study illustration: (**a**) original unweighted Kaplan–Meier curves; (**b**) original unweighted predicted Cox proportional hazards model survivor functions; and (**c**) corresponding inverse propensity score-weighted predicted Cox proportional hazards model survivor functions (unstandardized untrimmed primary analysis). Panels **b** and **c** show survivor functions with 95% confidence limits.

Table 3 HR of overall survival sensitivity analysis results for pemetrexed arm (relative to gemcitabine arm) of JMDB randomized controlled trial population included in this case study illustration: (1) trimmed weights ≥ 4; (2) excluding number of metastatic sites as a baseline covariate

Sensitivity analysis	HR	Bootstrap 2.5 percentile	Bootstrap 97.5 percentile
Inverse propensity score-weighted analysis – trimmed weights ≥ 4 (ESS = 265)	1.00	0.74	1.34
Inverse propensity score-weighted analysis – excluding number of metastatic sites (ESS = 384)	0.80	0.62	1.02

ESS, effective sample size; HR, hazard ratio (pemetrexed vs. gemcitabine).

metastatic sites as a baseline covariate from the logistic regression model resulted in a slight increase in the effect of pemetrexed vs. gemcitabine (indicated by a smaller IPS-weighted HR of OS) compared with the primary analysis.

The results of the third sensitivity analysis using an entropy balancing reweighting approach were broadly similar to the results of the primary IPS-weighted analysis (data not shown). Considering the three primary and sensitivity IPS-weighted analyses reported here, closer matching to the target FRAME population seemed to come at the expense of higher variability.

DISCUSSION

We presented a case study as part of the IMI GetReal program, applying and illustrating a novel exploratory approach to generalizing RCT data to real-world clinical practice in the context of HTA and regulatory decision making. This approach reweights RCT data based on real-world observational study baseline data to attempt to mirror real-life patient characteristics in a clinical trial setting. It offers a way to address questions commonly raised by healthcare decision makers regarding the applicability of RCT results to the full range of real-world patients who may receive care in their locality postmarketing authorization. However, it is important to note that this method provides an indication of what RCT results would look like if a trial were to be carried out in a real-world target population but does not suggest that these are the results that would be observed outside the RCT setting (i.e., in the real world). Therefore, this approach is proposed only as a decision-making tool and not as an alternative to gathering important RWE once it becomes feasible to do so or to conducting further RCTs.

The main benefit of IPS weighting, as explored in this case study illustration, is that it only requires a cross-sectional observational sample of an indicated target population to simulate clinical trial outcomes for real-world patients. Therefore, this method can address generalizability concerns at the time of regulatory approval/marketing authorization. If it is necessary to assess and reweight data from multiple RCTs (as may be required for multiple RCT-based marketing authorization applications), this reweighting method could be applied to each RCT individually, against the same target population. A meta-analysis of the RCTs could then be conducted using the reweighted results, rather than the original unweighted RCT results. Furthermore, IPS weighting has the flexibility to be applied to multiple different real-world target patient populations depending on the requirements of the decision maker, if data sets for analysis are available. This could also be important for application areas, such as payer access schemes, which link outcomes to payments. IPS weighting is an intuitive approach to tackling the issue of limited external validity associated with RCTs in that it mimics real-world target patient populations while operating within the framework of gold standard RCTs and without necessarily undermining the important randomization concept required for causal inference. In case of a possible reduction in ESS, closer matching to a real-world population needs to be balanced against increased uncertainty. Early scientific advice consultation with regulatory or HTA authorities could help to determine *a priori* what an acceptable threshold of uncertainty is.

As with all approaches to generalizing evidence from RCTs to target populations, the IPS-weighting method is associated with some limitations, such as availability of data, consistent definitions of variables between data sets, overlap in patient characteristics between data sets, balancing the closeness of RCT and target patient populations with ESS, and the potential for breaching underlying model assumptions.

The IPS-weighting method requires patient-level RCT and baseline target population data, although other reweighting approaches, such as entropy balancing,[30,31] can be used if only summary data are available. There are also alternative methods of generating propensity scores to the one used in this analysis (e.g., penalized logistic regression).[32] We would suggest a sensitivity analysis using an alternative weighting method to assess the robustness of any results. Furthermore, the availability of clinically relevant variables may constrain IPS-weighting analyses as clinical relevance should guide the inclusion of variables. In the current analysis, a relatively large number of clinically relevant variables (15) were available and identified for inclusion as a result of the similarities between the Lilly-sponsored FRAME observational study and the JMDB RCT; however, existing registries may not have sufficient variable overlap to cover and control for critical treatment effect-modifying variables. Moreover, caution is warranted in cases where variable definitions vary between data sets. Nonoverlapping definitions of variables to be included in an IPS-weighting analysis should be considered and managed on a case-by-case basis, and the influence of the individual covariates on the results of the overall analysis should be assessed. Although this method requires at least some overlap in terms of available defined variables between RCT and observational data sets, it is not an all or nothing approach. If only a handful of clinically relevant variables are available with overlapping definitions in both the RCT and observational data sets, a population that is more like the target observational study population than the starting RCT population can still be achieved. Generally, justification of an appropriate population is necessary in any submission to decision makers.

The IPS-weighting method assumes there is enough overlap between the RCT and the observational study populations to make inference (i.e., real-world patients who are not represented in the RCT at all cannot be mapped) and that the only differences that matter are those measured and adjusted for. In the current analysis, the JMDB RCT did not include patients with zero metastatic sites, whereas the FRAME observational study did allow for them. Therefore, we categorized these patients together with patients with one metastatic site. Sensitivity analysis showed that excluding the number of metastatic sites as a covariate did not alter the overall results. Although patients' number of metastatic sites is certainly associated with clinical outcomes in advanced stage NSCLC, it did not seem to considerably impact the reweighted comparative outcomes between the treatment arms (i.e., the IPS-weighted HR of OS remained fairly stable). With this method, extreme weight is reflective of significant difference between two populations in a specific variable (i.e., selection bias). However, extreme weight without an effect on the HR estimate is reflective of no treatment effect of that variable (i.e., no treatment heterogeneity). Those variables most strongly associated with both population differences (i.e., selection bias) and treatment effects (i.e., treatment heterogeneity) would be expected to have the greatest effect on the HR estimate.

Integral to the IPS-weighting approach is balancing the closeness of the RCT and target patient populations with the associated impact on the ESS available for the analysis. After reweighting, if the ESS is smaller than the original sample size, then the standard error of the estimate (and subsequent 95% CIs) will be larger in the reweighted analysis. Therefore, it is important that the ESS, a key metric in this approach, is assessed and a balance is achieved between the number of covariates included and the number of patients contributing to the analysis. It is recommended to only include covariates that are considered as potential treatment effect modifiers when generating the weights. Conversely, if an influential covariate is not included then the reweighting will be biased. Furthermore, if there is minimal overlap in the propensity scores, resulting in some extreme weights, then trimming is recommended as a sensitivity analysis (i.e., excess weights are down-weighted to avoid a minority of patients contributing to most of the weights). Although we applied only one illustrative example, the decision to trim the weights is related to the ESS and trimming in itself can introduce bias, thus varying trimming cutoff points can serve as additional sensitivity analyses to demonstrate the robustness of the results. Finally, as with any reweighting approach, caution is warranted with the IPS-weighting approach if the underlying assumptions of the model are not met.

Whereas more case studies in different disease areas are needed to assess the acceptability of this approach as well as its limitations, the IPS-weighting methodology provides an innovative approach to bridging real-world and clinical trial evidence in the context of HTA and regulatory decision making. Insights from such exploratory analyses (e.g., increased/decreased efficacy in under-represented subgroups) could highlight the importance of additional RCTs or of generating RWE sooner rather than later to complement findings from existing RCTs.

ACKNOWLEDGMENTS

The authors would like to acknowledge Sue Williamson and Greg Plosker (Rx Communications, Mold, UK) for medical writing assistance with the preparation of this manuscript, funded by Eli Lilly and Company.

FUNDING

The work leading to these results has received support from the Innovative Medicines Initiative Joint Undertaking under grant agreement number (115546), resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies' in-kind contribution. It was conducted as part of the GetReal consortium. For further information, please refer to www.imi-getreal.eu. This paper only reflects the personal views of the stated authors.

CONFLICT OF INTEREST

M.H., A.B., D.F., K.B.W., A.G., J.J., and M.B. are employees and shareholders of Eli Lilly and Company. K.A., P.J., and Eli Lilly and Company are part of the Innovative Medicines Initiative GetReal consortium.

AUTHOR CONTRIBUTIONS

M.H., A.B., D.F., K.B.W., P.J., J.J., and M.B. wrote the manuscript. M.H., A.B., D.F., K.A., J.J., and M.B. designed the research. K.B.W., A.G., and M.B. performed the research. M.H., A.B., D.F., K.A., A.G., P.J., J.J., and M.B. analyzed the data.

1. UMC Utrecht. Welcome to GetReal. IMI GetReal <http://www.imi-getreal.eu/> (2019). Accessed September 17, 2019.
2. Kennedy-Martin, T., Curtis, S., Faries, D., Robinson, S. & Johnston, J. A literature review on the representativeness of randomized controlled trial samples and implications for the external validity of clinical trial results. *Trials* **16**, 495 (2015).
3. Pressler, T.R., Kaizar, E.E. The use of propensity scores and observational data to estimate randomized controlled trial generalizability bias. *Stat. Med.* **32**, 3552–3568 (2013).
4. United Kingdom National Institute for Health and Care Excellence (NICE). Guide to the processes of technology appraisal. Section 2.4.32 <https://www.nice.org.uk/Media/Default/About/what-we-do/NICE-guidance/NICE-technology-appraisals/technology-appraisal-processes-guide-apr-2018.pdf> (April 2018). Accessed September 17, 2019.
5. United Kingdom National Institute for Health and Care Excellence (NICE). Guide to methods of technology appraisal 2013: process and methods. Sections 3.2.2 and 3.3.3 <https://www.nice.org.uk/process/pmg9/resources/guide-to-the-methods-of-technology-appraisal-2013-pdf-2007975843781> (April 2013). Accessed September 17, 2019.
6. American Academy of Managed Care Pharmacy (AMCP). Format for formulary submission. Version 4.0 <http://www.amcp.org/sites/default/files/2019-03/AMCP-Format-V4.pdf> (April 2016). Accessed September 17, 2019.
7. Australian Pharmaceutical Benefits Advisory Committee (PBAC). Guidelines for preparing a submission to the Pharmaceutical Benefits Advisory Committee. Version 5.0. Section 2.7 <https://pbac.pbs.gov.au/content/information/files/pbac-guidelines-version-5.pdf> (September 2016). Accessed September 17, 2019.
8. Canadian Agency for Drugs and Technologies in Health (CADTH). Guide to providing clinician input and feedback with the pan-Canadian Oncology Drug Review (pCODR) program <https://cadth.ca/sites/default/files/pcodr/pCODR%27s%20Drug%20

Review%20Process/pCODRClinicianInput_FeedbackGuide.pdf> (2019). Accessed September 17, 2019.
9. European Medicines Agency (EMA), Committee for Medicinal Products for Human Use (CHMP). Guideline on the investigation of subgroups in confirmatory clinical trials. EMA/CHMP/539146/2013 <https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-investigation-subgroups-confirmatory-clinical-trials_en.pdf> (March 31, 2019). Accessed September 17, 2019.
10. US Food and Drug Administration (FDA). Enrichment strategies for clinical trials to support determination of effectiveness of human drugs and biological products: guidance for industry <https://www.fda.gov/media/121320/download> (March 2019). Accessed September 17, 2019.
11. Eddy, D.M. The confidence profile method: a Bayesian method for assessing health technologies. *Oper. Res.* **37**, 210–228 (1989).
12. Shih, Y.C. & Kauf, T.L. Reconciling decision models with the real world. An application on anaemia of renal failure. *Pharmacoeconomics* **15**, 481–493 (1999).
13. Baker, S.G. & Kramer, B.S. Randomized trials, generalizability, and meta-analysis: graphical insights for binary outcomes. *BMC Med. Res. Methodol.* **16**, 10 (2003).
14. Frangakis, C. The calibration of treatment effects from clinical trials to target populations. *Clin. Trials* **6**, 136–140 (2009).
15. Pastopoulos, N.A. A pragmatic view on pragmatic trials. *Dialogues Clin. Neurosci.* **13**, 217–224 (2011).
16. Loudon, K., Treweek, S., Sullivan, F., Donnan, P. & Thorpe, K.E. The PRECIS-2 tool: designing tools that are fit for purpose. *BMJ* **350**, h2147 (2015).
17. Cole, S.R. & Stuart, E.A. Generalizing evidence from randomized clinical trials to target populations: the ACTG 320 trial. *Am. J. Epidemiol.* **172**, 107–115 (2010).
18. Stuart, E.A., Cole, S.R., Bradshaw, C.P. & Leaf, P.J. The use of propensity scores to assess the generalizability of results from randomized trials. *J. R. Stat. Soc. Ser. A Stat. Soc.* **174**, 369–386 (2011).
19. Lesko, C.R., Buchanan, A.L., Westreich, D., Edwards, J.K., Hudgens, M.G. & Cole, S.R. Generalizing study results: a potential outcomes perspective. *Epidemiology* **28**, 553–561 (2017).
20. Hong, J.L. *et al*. Generalizing randomized clinical trial results: Implementation and challenges related to missing data in the target population. *Am. J. Epidemiol.* **187**, 817–827 (2018).
21. Webster-Clark, M.A., Sanoff, H.K., Stürmer, T., Peacock Hinton, S. & Lund, J.L. Diagnostic assessment of assumptions for external validity: an example using data in metastatic colorectal cancer. *Epidemiology* **30**, 103–111 (2019).
22. Rosanbaum, P.R. & Rubin, D.B. The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55 (1983).
23. Scagliotti, G.V. *et al*. Phase III study comparing cisplatin plus gemcitabine with cisplatin plus pemetrexed in chemotherapy-naive patients with advanced-stage non-small-cell lung cancer. *J. Clin. Oncol.* **26**, 3543–3551 (2008).
24. Moro-Sibilot, D. *et al*. Outcomes and resource use of non-small cell lung cancer (NSCLC) patients treated with first-line platinum-based chemotherapy across Europe: FRAME prospective observational study. *Lung Cancer* **88**, 215–222 (2015).
25. Schnabel, P.A. *et al*. Influence of histology and biomarkers on first-line treatment of advanced non-small cell lung cancer in routine care setting: baseline results of an observational study (FRAME). *Lung Cancer* **78**, 263–269 (2012).
26. Eli Lilly and Company. Alimta (pemetrexed) 100mg powder for concentrate for solution for infusion. UK Summary of Product Characteristics. January 17, 2019.
27. Rubin, D.B. Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Serv. Outcomes Res. Methodol.* **2**, 169–188 (2001).
28. Efron, B. & Gong, G. A leisurely look at the bootstrap, the jackknife, and cross-validation. *Am. Stat.* **37**, 36–48 (1983).
29. Signorovitch, J.E. *et al*. Comparative effectiveness without head-to-head trials: a method for matching-adjusted indirect comparisons applied to psoriasis treatment with adalimumab or etanercept. *Pharmacoeconomics*. **28**, 935–945 (2010).
30. Hainmueller, J. Entropy balancing for causal effects: a multivariate reweighting method to produce balanced samples in observational studies. *Polit. Anal.* **20**, 25–46 (2012).
31. Hong, J.L. *et al*. Comparison of methods to generalize randomized clinical trial results without individual level data for the target population. *Am. J. Epidemiol.* **188**, 426–437 (2019).
32. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B* **58**, 267–288 (1996).