Case Study: Propensity Weighting and Extrapolation in Non Small Cell Lung Cancer

Lead Organisations: University of Leicester, Eli Lilly
1. Executive summary

A key objective of Work Package 1 of GetReal is to develop a framework for incorporating real world data (RWD) in decision-making. Case Studies have been constructed to explore different ways by which use of RWD may help demonstrate relative effectiveness of new medicines.

This report relates to a case study examining the use of:

- Propensity Weighting to use observational data (real world data: RWD) to re-weight data available from an RCT to generate estimates of effect generalizable to a target population (for Health Technology Assessment and reimbursement decision making)
- Extrapolation techniques to use RWD to support generation of estimates of treatment effect, especially survival and quality-adjusted survival over time periods beyond those measured in RCTs

The choice of Non Small Cell Lung Cancer (NSCLC: stage IIIb/IV) to illustrate/test these methods was primarily due to the availability of RCT data and corresponding RWD provided by a GetReal partner (Lilly). The methods are of particular value to cancer medicines for which trials are often undertaken in study populations drawn from many centres, satisfying strict inclusion and exclusion criteria. Generalizability of trial results to national or local ‘reimbursable’ populations may be questioned. In oncology there is a large volume of health technology assessment (HTA) assessments to support reimbursement decision-making. The methods tested in this case study have application to a wide variety of medicines in cancer and other disease areas.

The outputs of the analyses relating to effectiveness of pemetrexed in Stage IIIb/VI NSCLC were discussed at a GetReal workshop held in Frankfurt on 10th Sept 2015, during which stakeholder views of the utility and applicability of the methods tested were obtained.

Although overall survival differences between pemetrexed and gemcitabine appear more pronounced after reweighting, the reweighted analysis of the clinical trial yielded a hazard ratio (HR) closer to 1, with greater uncertainty: HR of 0.86 (95% CI: 0.59 to 1.30) compared with 0.81 (95% CI: 0.70 to 0.94) in a similar population in the clinical trial. Sensitivity analyses to both the methods of reweighting and the inclusion of baseline covariates gave broadly similar results.

Although there was substantial interest in this method, feedback from the stakeholder workshop was that it is not ready for use in regulatory or reimbursement decision-making yet. However there may be application for Pharma R&D as a way of understanding or exploring the benefit-risk profile of medicines in development and (thereby) helping design Phase 3 trials. A wider range of case studies, coupled with technical review of the methods, validation of
outputs and communication/education of decision-makers will be required before the method receives wider support. It will most likely have utility in disease areas where there is known to be a large ‘efficacy-effectiveness gap’ and possibly in the context of adaptive pathways or managed access schemes. In addition there was a call for earlier planning of RWE studies and more involvement of RWE experts in trial designs, to enable such analyses to be performed more readily in future.

In the second analysis (parametric) survival functions were fitted to both unweighted and FRAME-reweighted overall survival data from the JMBD trial alone, and then by adjusting these curves to fit survival estimates from the UK cancer registry. Goodness of fit of the different survival curves was assessed using the Deviance Information Criterion (DIC) and the Area Under the Curve (and standard error) calculated over 6 years. When trial data only were used a Weibull distribution appeared to provide the best fit, however when cancer registry data were included use of a lognormal distribution was more supported.

These types of analyses were more familiar to workshop participants, although it was acknowledged that they are of most interest currently to UK NICE (in helping to estimate lifetime QALYs for trial subjects). The method was thought to have most utility for chronic conditions with long-term (unmeasured) outcomes, and possibly in the context of adaptive pathways. The main value may be to help reconcile differences of opinion about which survival model to use for such extrapolations, but attention needs to be paid to the match of the source trial and RWE populations. Similar to the reweighting case study, more evidence of the robustness of the method and case studies are required.

A framework for further development of new analytical techniques (using RWE) was proposed: Understanding of methodology → Validation of methods → Peer-reviewed publications (both methodological and applied/tutorial) → Guidance → Training/education.

This report first (Section 2) provides a summary of NSCLC and its treatment options, together with a brief description of the two data sources used in the analyses: the JMBD trial (RCT) and the FRAME study (RWE). Section 3 gives an overview of the generalizability issues faced by HTA agencies addressed in the case study analyses. Section 4 describes the method of reweighting data using propensity scores and gives summary results from the case study. The extrapolation method is described and summary results provided in Section 5. Sections 6 to 8 give summary feedback from the stakeholder workshop, for which further details are provided in Appendices.

Sections 2-4 correspond to GetReal deliverable 1.5 and Sections 5-8 and the Appendices to GetReal deliverable 1.6 for this case study.
Acknowledgements

Alan Brnabic, Eli Lilly and Company, 112 Wharf Road, West Ryde, Sydney, 2132, Australia
Douglas Faries, Lilly Corporate Center, Eli Lilly and Company, Indianapolis, IN, 46285, USA
Katherine Winfree, Lilly Corporate Center, Eli Lilly and Company, Indianapolis, IN, 46285, USA
Allicia Girvan, Lilly Corporate Center, Eli Lilly and Company, Indianapolis, IN, 46285, USA
Pall Jonsson, National Institute for Health and Care Excellence (NICE), Piccadilly Plaza, Manchester, M1 4BT, UK
Joseph Johnston, Lilly Corporate Center, Eli Lilly and Company, Indianapolis, IN, 46285, USA
Mark Belger, Lilly Research Centre, Eli Lilly and Company, Erl Wood Manor, Sunninghill Road, GU20 6PH, UK
Pascale Dequen, Department of Health Sciences, University of Leicester, UK
# Table of Contents

1. EXECUTIVE SUMMARY ........................................................................................................ 2
2. NSCLC AND PATIENT MANAGEMENT ............................................................................. 6
3. GENERALIZABILITY ISSUES IN HTA ............................................................................... 10
4. GENERALISING RCT EVIDENCE FROM JMDB TRIAL: POPULATION (PROPENSITY SCORE RE-WEIGHTING) .................................................................................................................. 12
5. GENERALISING RCT EVIDENCE FROM JMDB TRIAL: TIME HORIZON (EXTRAPOLATION) ......................................................................................................................... 18
6. STAKEHOLDER WORKSHOP: 10TH SEPT 2015 ............................................................... 24
7. SUMMARY OF WORKSHOP FEEDBACK – PROPENSITY SCORE WEIGHTING (PW) METHOD ..................................................................................................................... 25
8. SUMMARY OF WORKSHOP FEEDBACK – EXTRAPOLATION METHOD ......................... 28
REFERENCES .......................................................................................................................... 31
APPENDIX A. WORKSHOP AGENDA ..................................................................................... 33
APPENDIX B. WORKSHOP PARTICIPANTS ....................................................................... 35
APPENDIX C. DETAILED GROUP FEEDBACK ...................................................................... 36
2. NSCLC and patient management

Lung Cancer

Lung cancer is the most common cancer worldwide, and it is also the most common cause of death from cancer. Approximately 70% of patients with Non Small Cell Lung Cancer (NSCLC) present with locally advanced (Stage IIIB) or metastatic (Stage IV) disease and these patients have a very poor prognosis with a 5-year survival rate of approximately 6%. NSCLC is divided into three primary histologic groups (adenocarcinoma, squamous cell carcinoma, and large cell carcinoma) (Peters et al, 2012). Adenocarcinoma is the most frequent non-squamous histology. The identification of predictive factors associated with the clinical outcome of a specific treatment in patients with specific histology in principle supports a tailored approach to the treatment of patients with advanced NSCLC.

First Line Treatment for Stage IIIB/IV NSCLC

The treatment of locally advanced inoperable or metastatic NSCLC (stage IIIB/IV) hinges mainly on chemotherapy (Molina, 2006). First line treatment has evolved in the course of the last decade. The majority of therapeutic options consist of chemotherapy combining a platinum salt with a third generation cytotoxic agent (gemcitabine, vinorelbine, paclitaxel, docetaxel), which has led to the definition of a therapeutic standard for patients who are still in good general health (Eastern Cooperative Oncology Group - ECOG performance status score ≤ 1) (Peters, 2012).

Pemetrexed (Alimta® Lilly) is a multi-targeted anti-cancer antifolate agent that disrupts crucial folate-dependent metabolic processes essential for cell replication. Pemetrexed, combined with cisplatin as a doublet, has obtained marketing authorisation (MA) as a first line treatment for patients with predominantly non-squamous NSCLC, at a locally advanced or metastatic stage (Scagliotti et al, 2008).

The JMDB trial

JMDB was a non-inferiority, phase III, randomized study that enrolled 1,725 chemotherapy-naive patients of any histology with stage IIIB or IV NSCLC and an ECOG performance status of 0 to 1 (Scagliotti et al, 2008). Patients received cisplatin 75 mg/m² on day 1 and gemcitabine 1,250 mg/m² on days 1 and 8 (n = 863) or cisplatin 75 mg/m² and pemetrexed 500 mg/m² on day 1 every 3 weeks for up to six cycles (n = 862). The primary objective of JMDB was to compare overall survival of patients treated with pemetrexed plus cisplatin vs gemcitabine plus cisplatin.
Figure 1 displays the results of JMDB for both OS and PFS in patients with non-squamous histology. For OS, pemetrexed was associated with a statistically significant improvement in overall survival (HR 0.81, 95% CI: 0.70 to 0.94), with an increase of 1.4 months in median overall survival.

**FRAME RWE Study**

FRAME was a prospective, non-interventional, multicentre observational study conducted in 11 European countries (Schnabel et al, 2012). 1567 patients were observed, including adult patients receiving a platinum-based doublet as first-line therapy (+/- targeted therapy) for Stage IIIB or IV NSCLC of any histology, according to staging guidelines at the time of study development. Patient visits and evaluations were completed within the physician’s routine clinical practice; choice of therapy, timing and dosage were at the physician’s discretion.

Figure 2 shows the actual treatments received by patients in the FRAME study. Patients were followed for at least 18 months or until discontinuation for any reason (including death) or until the end of the study. The primary objective of FRAME was to evaluate overall survival among different first-line treatment cohorts with and without additional targeted therapy, and the initial results are presented in Figure 3 for those patients receiving pemetrexed plus platinum (n=569) or gemcitabine plus platinum (n=361).
Figure 2. Treatments received in FRAME study

![Treatments Diagram](image)

Table 1 displays the baseline characteristics for study subjects in JMDB and FRAME. With the exception of demographic characteristics, there are statistically significant differences in the patient populations, these being most marked in the presence and number of sites of metastases.

Figure 3. FRAME results: OS and PFS for Pemetrexed + Platinum and Gemcitabine + Platinum

![Survival Graphs](image)

Table 1 displays the baseline characteristics for study subjects in JMDB and FRAME. With the exception of demographic characteristics, there are statistically significant differences in the patient populations, these being most marked in the presence and number of sites of metastases.
<table>
<thead>
<tr>
<th>Characteristic</th>
<th>FRAME (N=361)</th>
<th>JMDB (N=1216)</th>
<th>p-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age in years mean (sd)</td>
<td>60.4 (8.88)</td>
<td>59.6 (9.39)</td>
<td>0.96</td>
</tr>
<tr>
<td>Time since diagnosis month (sd)</td>
<td>2.7 (11.84)</td>
<td>2.3 (9.40)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Female n (%)</td>
<td>111 (31%)</td>
<td>424 (35%)</td>
<td>0.16</td>
</tr>
<tr>
<td>Non-Asian n (%)</td>
<td>355 (98%)</td>
<td>1006 (83%)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Smoker</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Current smoker</td>
<td>108 (30%)</td>
<td>274 (23%)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Ex smoker</td>
<td>183 (51%)</td>
<td>601 (49%)</td>
<td></td>
</tr>
<tr>
<td>Never smoker</td>
<td>56 (16%)</td>
<td>202 (17%)</td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>14 (4%)</td>
<td>139 (11%)</td>
<td></td>
</tr>
<tr>
<td>Diagnosis n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cytological</td>
<td>84 (23%)</td>
<td>459 (38%)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Histopathological</td>
<td>277 (77%)</td>
<td>757 (62%)</td>
<td></td>
</tr>
<tr>
<td>Diag subtype n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adenocarcinoma</td>
<td>306 (85%)</td>
<td>828 (68%)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Large Cell Carcinoma</td>
<td>24 (7%)</td>
<td>157 (13%)</td>
<td></td>
</tr>
<tr>
<td>other</td>
<td>31 (9%)</td>
<td>231 (19%)</td>
<td></td>
</tr>
<tr>
<td>Stage n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IIIB</td>
<td>55 (15%)</td>
<td>287 (24%)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>IV</td>
<td>306 (85%)</td>
<td>929 (76%)</td>
<td></td>
</tr>
<tr>
<td>ECOG n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>107 (30%)</td>
<td>477 (39%)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>1</td>
<td>254 (70%)</td>
<td>739 (61%)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>295 (82%)</td>
<td>296 (24%)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>3+</td>
<td>56 (16%)</td>
<td>286 (24%)</td>
<td></td>
</tr>
<tr>
<td>N Met. Sites n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>10 (3%)</td>
<td>634 (52%)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>39 (11%)</td>
<td>40 (7%)</td>
<td>0.01</td>
</tr>
<tr>
<td>Prior Rtherapy n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>43 (12%)</td>
<td>73 (6%)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Prior Surgery n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>42 (12%)</td>
<td>96 (8%)</td>
<td>0.03</td>
</tr>
<tr>
<td>CV History n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>126 (35%)</td>
<td>736 (61%)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Diabetes n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>40 (11%)</td>
<td>84 (7%)</td>
<td>0.01</td>
</tr>
<tr>
<td>Lung History n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>134 (37%)</td>
<td>410 (34%)</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>
3. Generalizability Issues in HTA

Population

While the demonstration of improved outcomes within RCTs is widely accepted as foundational evidence of the efficacy of new treatments, stakeholders frequently express concern that RCTs lack “external validity” – i.e., that the results may not apply to the full range of patients who may eventually receive care outside the trial environment, in “real-world” clinical situations. The use of restrictive patient selection criteria, often used to isolate a population in whom benefit can most clearly be attributed to treatment (to maximize “internal validity”), can exacerbate these concerns. Ultimately, a combination of evidence from RCTs and observational studies may be required to definitively overcome these concerns. However at the time of product marketing authorization, such an array of studies is unlikely to be available.

Post-stratification is a common way of generalizing results to a target population. It reweights effects based on population distributions (Stuart 2001). Post-stratification can be very effective when there are only a small number of variables to control for, but it is infeasible when there are many (or continuous) variables, leading to a very large number of post-stratification cells. Alternative strategies include meta-analysis (Sutton 2008), cross-design synthesis (Prevost 2000) and the confidence profile method (Eddy 1992). Many of these approaches aim to model treatment effects as a function of study parameters, such as randomized vs non-randomized, or explicit inclusion/exclusion criteria, and they generally rely on having a relatively large set of studies to include in the analysis. Unfortunately at the time of marketing authorisation and reimbursement of a new medicine there may be only one or two studies from which conclusions can be drawn. In addition, when applying these approaches limited attention is usually paid to the types of participants who are enrolled in the various source studies, and how variation in their characteristics may affect the results.

One approach to addressing concerns about external validity of RCTs is to use existing observational data from the target population to “re-weight” available RCT data. For example, efficacy results for patient subgroups under- or over-represented in RCTs can be over- or under-weighted (respectively) to generate a weighted estimate more reflective of the make-up of the target population. In this way, one can generate an estimate of the expected treatment benefit had the trial been run in this broader “real-world” population.

An ideal circumstance in which to use such an approach would be one where comparable baseline variables are available from both a clinical trial cohort and a (less restrictive) observational study cohort, but outcomes data are available only from the clinical trial. Caution is warranted in cases where, for example:

- Variable definitions vary
- Only a limited number of RCT variables are available in the RWE data source
- Matching is insufficient due to diverging inclusion/exclusion criteria
• Patients willingness to participate depends on study form
• Unobserved confounding still exists

**Time Horizon**

RCTs evaluating different treatments in terms of overall survival are often relatively short-term in nature compared to the natural history of the disease. Survival data are often censored, i.e. death is not observed in all patients, meaning that extrapolation techniques need to be used to obtain estimates of the entire survival benefit that would be experienced in the ‘real-world’ by a cohort of patients treated with a new medicine of interest.

Where such analyses are not undertaken estimates of the survival benefit will be restricted to that observed directly in the relevant clinical trial(s) and this is likely to be a poor estimate of the true survival gain. There are a number of methods available for performing extrapolation. Exponential, Weibull, Gompertz, log-logistic or log normal parametric models can be used, as well as more complex and flexible models (Royston 2011). The extrapolated survival curve is sometimes then summarized by the Area Under the Curve (AUC) which yields (gains in) mean overall survival and aggregate life years gained which is likely to be required for an economic evaluation to be submitted to HTA (Latimer 2013). The choice of which parametric distribution to adopt may be based on the statistical fit to the RCT data, commonly using metrics such as the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) or Deviance Information Criterion (DIC). However, this approach only considers the fit to the available RCT data, and not how well each parametric distribution predict future survival. If individual patient data (IPD) are available from real world evidence (RWE) studies, for example cancer registries, then approaches incorporating both RCT and RWE data can be employed, and indeed have been used in HTA submissions. A common approach is to use the observed Kaplan-Meier curve for the RCT data and to then splice this with the registry data beyond the follow-up of the RCT. A key consideration for this approach is at what time point to splice the RCT and RWE datasets (Bagust 2014).
4. Generalising RCT Evidence from JMDB trial: Population (Propensity Score Re-Weighting)

Method

Generalizability was initially assessed by summarizing differences in the individual demographic and baseline clinical characteristics as well as in assessing propensity score distributions between patients in the JMDB and FRAME studies. The propensity score was used in this context to denote the approximate probability of a patient being enrolled in either the JMDB trial or the FRAME observational study.

In principle, the results derived from JMDB were considered generalizable to the more general FRAME population if the propensity score distributions in these two studies were not too dissimilar. The similarity was evaluated in two ways. The first approach was to plot histograms of the propensity scores from the two studies and compare the portion of their overlapping region. The larger the overlapping region, the more evidence there was to support the generalizability of the JMDB trial. The second approach was to quantify the similarity in terms of the difference in mean propensity scores between JMDB or FRAME patients. Differences were standardised by dividing them by their standard deviation. Although there was no absolute rule to determine the magnitude of the standardized difference, generalization was considered to be unreliable if the standardised difference was large (i.e. greater than a threshold between 0.10 and 0.25) (Rubin, 2001).

Efficacy in the JMDB trial was re-calibrated relative to the FRAME population using the general approach proposed by Cole (2010). The calibration procedure was carried out in two steps. In the first step, each patient in the JMDB trial was assigned a weight based on the inverse of his or her propensity score. In the second step, overall and progression-free survival was computed using weighted values for each patient in JMDB.

As a sensitivity analysis, a more refined weighting algorithm was also applied. Entropy weights (Heinmuller 2012) were computed for each patient in the JMDB trial, using the FRAME study as the target population. Further details on this method are provided below.

Assessing Generalizability

Generalizability was initially assessed by summarizing differences in the individual demographic and baseline clinical characteristics as well as in propensity scores between patients in the JMDB and FRAME studies. Comparison between studies was formally assessed using t-tests or analysis of variance (ANOVA) or the median test for continuous variables, and the Fisher’s exact test for categorical variables. Individual variables provided detailed insight at a univariate level into the differences in patient composition in the JMDB and FRAME studies.
**Propensity score weighting**

The propensity score provided a singular numeric metric to summarize simultaneously all patient characteristics. In the present analysis ‘propensity score’ denotes the approximate probability of a patient being enrolled in the JMDB trial (NB this differs from other more well-recognized definitions of propensity scores). Given that the propensity score represents the probability of an individual patient being assigned to the JMDB or FRAME studies, it served as the vehicle to demonstrate the similarity (or dissimilarity) between the trial participants and the general population, i.e. represented by the FRAME study. In this context, the inverse propensity score can be considered as analogous to weighting used in probability sampling for surveys.

The computation of propensity scores was carried out via the following steps:

1. Patient data from JMDB or FRAME studies were combined using their common demographic and baseline clinical characteristics, creating an indicator to denote the membership of individual patients in each study (using 1 and 0 to denote JMDB or FRAME, respectively).

2. A logistic regression model was constructed using the membership indicator as the dependent variable and the other covariates as the independent variables. Before moving forward with the analysis, the propensity-adjusted balance, based on the covariates used, between the studies (JMDB vs FRAME) and between the weighted treatment cohorts within JMDB were examined using standardized differences. If successful, propensity adjustment should produce standardized differences for almost all covariates of 0.1 or less. The propensity model may be re-evaluated with different factors or interactions to achieve balance.

3. Each patient’s propensity score was calculated by applying the logistic model to the given covariates for that patient.

Results derived from JMDB may be considered generalisable to the more general FRAME population if the propensity score distributions in these two studies are not too dissimilar. The similarity was evaluated in two ways. The first approach was to plot histograms of the propensity scores from the two studies and compare the degree of overlap, a larger overlapping region being supportive of the generalisability of the JMDB trial. Although there is no absolute rule to assess the standardized difference, results derived from the trial (JMDB) population are in principle considered generalisable to the more general FRAME population if the propensity score distributions in these two studies are not too dissimilar. Similarity was evaluated firstly by plotting histograms of the propensity scores from the two studies and comparing the portion of their overlapping region. The larger the overlapping region, the more evidence there is to support the generalizability of the JMDB trial.

The second approach was to quantify the similarity in terms of the difference in mean propensity scores between JMDB or FRAME patients. The difference can be standardized by dividing it by its standard deviation. Although there is no absolute rule to determine the magnitude of the standardized difference, it is considered an unreliable generalization if the standardized difference is large (i.e. greater than a threshold between 0.10 and 0.25) (Rubin 2001).

Efficacy in the JMDB trial was re-calibrated relative to the FRAME population using the general approach proposed by Cole (2010). The calibration procedure was carried out in two steps. In the first step, each
patient in the JMDB trial was assigned a weight based on the inverse of his or her propensity score. Percentile or fixed value trimming was also performed to reduce the impact of patients with extreme weight values prior to calculating the final outcome. A successful calibration results in a weighted JMDB population appearing to represent a random sample from the FRAME study.

To examine the quality of the calibration, we compared the weighted demographic characteristics and the balance in baseline covariates after weighting between the pemetrexed and gemcitabine groups in the JMDB trial. In the second step, overall and progression-free survival was computed for each patient in JMDB. The calibrated efficacy was estimated by the hazard ratio from a weighted Cox proportional hazards model comparing the pemetrexed and gemcitabine arms. The variability and confidence interval of the calibrated efficacy was assessed by the non-parametric bootstrap procedure outlined below:

1. A random sample of the same number of patients in the combined JMDB and FRAME dataset was drawn (with replacement). The random sample was performed separately within the pooled JMDB and FRAME populations to maintain their relative sizes.

2. The propensity score was estimated by building a logistic regression model using the randomly sampled data.

3. The procedure to estimate the calibrated efficacy of pemetrexed versus gemcitabine (as described above) was followed, using the randomly-sampled data and a Cox proportional hazards regression model.

4. Steps 1 to 3 were repeated 1000 times.

5. Using the sample variance as an estimate of the variability of the calibrated efficacy (hazard ratio for OS), 1000 bootstrap samples with replacement were created for each of the datasets, using the same sample size as in the original dataset. Weights described above were calculated for each dataset and applied to each patient in the JMDB dataset (using the weight statement in PROC PHREG). This created a distribution of Hazard Ratios, from which a 95% confidence interval was obtained using the percentile approach: 2.5th and 97.5% percentiles of the bootstrap distribution.

In order to compare adjusted outcomes of individual treatments in JMDB against FRAME outcomes for the same treatments, the same procedure was applied to patients on pemetrexed or gemcitabine only. Patients on pemetrexed (or gemcitabine) in JMDB were reweighted based on characteristics of pemetrexed (or gemcitabine) patients in FRAME. Cox proportional hazard models and Kaplan-Meier methods were then be used to compare overall and progression-free survival between the FRAME pemetrexed (or gemcitabine) patients and reweighted JMDB pemetrexed (or gemcitabine) patient treatment arms.

As a sensitivity analysis, a more refined weighting algorithm was also applied. Entropy weights (Heinmuller, 2012) were computed for each patient in the JMDB trial – using the FRAME study as the target population. Calculation of weights was performed separately for each cohort in the JMDB trial (pemetrexed and gemcitabine) to ensure that the balance between cohorts within JMDB would not be distorted. Weights were also trimmed and then re-calibrated to avoid any undue influence of specific individuals, which would inflate of the variance in the analysis. To examine the quality of the calibration, we again compared the weighted demographic characteristics. In the second step, efficacy measurements, overall and progression-free survival, were computed for each patient in the JMDB.
dataset and the calibrated efficacy estimated as the hazard ratio from a weighted Cox proportional hazards model comparing the pemetrexed and gemcitabine arms. The variability and confidence interval of the calibrated efficacy was assessed by the non-parametric bootstrap procedure as described above.

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

Figure 4 shows the distribution of propensity scores for both JMDB and FRAME. Whilst it is evident that the distributions clearly overlap, for JMDB the scores are skewed for the higher propensity scores. This is mostly driven by differences in the distribution of patients between JMDB and FRAME in the number of metastases present (see Table 1, Appendix Figure 4A). This leads to considerable reweighting as shown in figure 5. 80% of patients in JMDB receive a weight (and its respective OS outcome) between 0.0-0.6, while some individual patients receive a weight of up to 10.

Figure 4. Distribution of propensity scores for JMDB (Red) and FRAME (Black).

Figure 5. Distributions of weights
Table 2 displays the original unweighted results for JMDB. Scagliotti 2008 reported a HR for OS of 0.81 (95% CI: 0.70 - 0.94) for pemetrexed compared to gemcitabine. By contrast the analysis using propensity weights produced a slightly smaller effect for pemetrexed (HR 0.859) with greater associated uncertainty, which was not statistically significant. Overall survival differences appear less pronounced.

Table 2. Results of propensity-weighted re-analysis of JMDB RCT using RWE from FRAME

<table>
<thead>
<tr>
<th>Category</th>
<th>Outcome</th>
<th>Treatment</th>
<th>N</th>
<th>Point Estimate</th>
<th>Hazard ratio</th>
<th>Bootstrap 2.5th percentile</th>
<th>Bootstrap 97.5th percentile</th>
</tr>
</thead>
<tbody>
<tr>
<td>No weights</td>
<td>OS</td>
<td>Gemcitabine</td>
<td>624</td>
<td>10.09</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>OS</td>
<td>Pemetrexed</td>
<td>592</td>
<td>11.47</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weighted</td>
<td>OS</td>
<td>Gemcitabine</td>
<td>614</td>
<td>8.57</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>OS</td>
<td>Pemetrexed</td>
<td>608</td>
<td>14.32</td>
<td>0.859</td>
<td>0.591</td>
<td>1.303</td>
</tr>
</tbody>
</table>

Further sensitivity analyses are reported in Table 3, in which the inclusion of the number of metastases as a baseline covariate was removed, Entropy weights were used in place of propensity weights, and the corresponding weights were standardised. Whilst results from the different sensitivity analyses are broadly comparable, there is a slight increase in the effect of pemetrexed (i.e. smaller HR) when the number of metastases at baseline is excluded from the analyses, and that in addition standardization reduces the level of uncertainty such that the results return to being of borderline statistical significance.

Table 3. Results of sensitivity analyses of propensity weighted re-analysis of JMDB RCT using FRAME

<table>
<thead>
<tr>
<th>Category</th>
<th>Outcome</th>
<th>Treatment</th>
<th>N</th>
<th>Point Estimate</th>
<th>Hazard ratio</th>
<th>Bootstrap 2.5th percentile</th>
<th>Bootstrap 97.5th percentile</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary cohort</td>
<td>OS</td>
<td>Gemcitabine</td>
<td>614</td>
<td>8.57</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>OS</td>
<td>Pemetrexed</td>
<td>608</td>
<td>14.32</td>
<td>0.859</td>
<td>0.591</td>
<td>1.303</td>
</tr>
<tr>
<td>Var # **</td>
<td>OS</td>
<td>Gemcitabine</td>
<td>614</td>
<td>9.86</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>OS</td>
<td>Pemetrexed</td>
<td>608</td>
<td>11.24</td>
<td>0.810</td>
<td>0.644</td>
<td>1.017</td>
</tr>
<tr>
<td>Entropy balancing</td>
<td>OS</td>
<td>Gemcitabine</td>
<td>614</td>
<td>9.92</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>OS</td>
<td>Pemetrexed</td>
<td>608</td>
<td>14.32</td>
<td>0.851</td>
<td>0.591</td>
<td>1.269</td>
</tr>
<tr>
<td>Entropy balancing **</td>
<td>OS</td>
<td>Gemcitabine</td>
<td>614</td>
<td>9.89</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>OS</td>
<td>Pemetrexed</td>
<td>608</td>
<td>11.33</td>
<td>0.794</td>
<td>0.632</td>
<td>1.003</td>
</tr>
<tr>
<td>Standardized weights</td>
<td>OS</td>
<td>Gemcitabine</td>
<td>614</td>
<td>9.92</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>OS</td>
<td>Pemetrexed</td>
<td>608</td>
<td>11.89</td>
<td>0.851</td>
<td>0.730</td>
<td>0.999</td>
</tr>
<tr>
<td>Standardized **</td>
<td>OS</td>
<td>Gemcitabine</td>
<td>614</td>
<td>9.99</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>OS</td>
<td>Pemetrexed</td>
<td>608</td>
<td>11.40</td>
<td>0.841</td>
<td>0.733</td>
<td>0.970</td>
</tr>
</tbody>
</table>

** metastases not accounted for
5. Generalising RCT Evidence from JMDB trial: Time horizon (extrapolation)

Method

Extrapolation over time can be used to inform the “best” and most appropriate extrapolation functions, preferably of an untreated (or BSC) cohort to reflect the situation where evidence of effectiveness is being assembled for a new medicine prior to launch. We used the UK cancer registry (UK CR), which provides a fairly comprehensive set of patient-level data and was assumed to be less heavily censored than many RCTs (ONS 2014). Extrapolation functions were fitted to unweighted JMDB data and FRAME-reweighted JMDB data, in each case for the comparator arm (gemcitabine). The extrapolated survival curve was summarized by the Area Under the Curve (AUC), from which mean overall survival was obtained. The choice of which parametric distribution to use (exponential, Weibull & logNormal) was based on the Deviance Information Criterion (DIC). In this analysis differences of 3 or more in DIC values were considered important, with lower DIC values representing a better relative fit to the data.

Further details of method

The situation was considered where RWE individual patient data (IPD) are not available, requiring the analysis to be based on summary RWE, for example overall survival at 5 years from a cancer registry for the target population.

In the case of JMDB the follow-up was 30 months, but in the associated NICE submission for pemetrexed (NICE TA181, 2009) this was extrapolated to 6 years. A number of parametric distributions were fitted to the data from JMDB and the AUC was calculated for each. The AUCs were then averaged using weights (corresponding to how plausible each is) derived from the 5 year overall survival estimates from the UK cancer registry for NSCLC, in order to estimate a single averaged AUC and associated uncertainty, which includes both the uncertainty associated with each parametric distribution and the fact there is often variability between the different distributions (Jackson 2009; ONS 2014).

The UK cancer registry (UK CR) provided a fairly comprehensive set of patient level data, and is considered to less heavily censored than many RCTs (ONS 2014). Extrapolation over time can be used to inform “best” and most appropriate extrapolation functions, preferably of an untreated (or BSC) cohort to reflect the situation of a product prior to launch. Similarly, unweighted and FRAME-reweighted JMDB data can be used to fit extrapolation functions to (BSC treated – here gemcitabine) data.

A variety of methods were used to extrapolate results from the JMDB trial to estimate (mean overall survival, AUC) beyond the 30 months of follow-up for the trial itself. These included extrapolation of trial data that were unweighted and also reweighted according to FRAME for a UK population, and well as using and not using external RWE from UK cancer registry data. In the associated NICE submission (NICE...
TA181, 2009) this extrapolation considered a 6 year time horizon and this is assumed here. Specific extrapolation methods considered were:

1. Use of parametric survival distributions (exponential, Weibull & logNormal) chosen using DIC. Differences of 3 or more in DIC values are considered important, with lower DIC values representing a better relative fit to the data.

2. Use of parametric survival distributions averaged over using; (i) uniform weights (i.e. assuming that each parametric distribution is equally plausible a priori to JMDB, (ii) Gaussian weights derived from UK cancer registry data, i.e. each parametric distribution is assessed against UK CR data at 5 years and weights (assessing the plausibility of each) derived for use in a Bayesian model averaging approach (Jackson et al, 2009).

All analyses were undertaken using Markov Chain Monte Carlo (MCMC) methods implemented in WinBUGS version 1.4.3 and R version 3.1.2.
Results

Table 3 displays the AUC and associated standard error (SE) for the three parametric distributions, together with DICs. It can be seen that in terms of the fit of the three distributions to both the pemetrexed and gemcitabine arms the Weibull distribution appears to provide the better fit. However, all three models indicate that pemetrexed is superior to gemcitabine in terms of the AUC calculated over 72 months (6 years).

Table 4. Area Under Curve (AUC) and Deviance Information Criterion (DIC) for parametric survival models fitted separately to Gemcitabine and Pemetrexed arms of JMDB RCT

<table>
<thead>
<tr>
<th>OS (72 months)</th>
<th>Gemcitabine</th>
<th>Pemetrexed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AUC (SE)</td>
<td>DIC</td>
</tr>
<tr>
<td>Exponential</td>
<td>12.95 (0.62)</td>
<td>3337.93</td>
</tr>
<tr>
<td>Weibull</td>
<td>12.85 (0.48)</td>
<td>3301.66</td>
</tr>
<tr>
<td>Log-Normal</td>
<td>14.98 (0.79)</td>
<td>3369.55</td>
</tr>
<tr>
<td>Uniform weights</td>
<td>12.85 (0.48)</td>
<td>-</td>
</tr>
<tr>
<td>UK CR Weights</td>
<td>12.95 (0.62)</td>
<td>-</td>
</tr>
</tbody>
</table>

*Results are for 72 months, based on original JMDB RCT (unweighted)*

Unsurprisingly using uniform/equal weights the Weibull distribution dominates, and the model-averaged results closely resemble those of the Weibull distribution. As can be seen from Figure 6 the UK CR data is closer to the log-Normal distributions fitted to the two trial arms – in fact the weights based on the UK CR data for the log-Normal distribution are over 99% for both arms. Consequently the model-averaged results using the CR UK data are drawn towards those of the log-Normal distribution, but with substantially increased uncertainty compared to that associated with the Weibull distribution. This illustrates the fact that a model averaging approach not only accounts for the uncertainty associated with each model but also between model uncertainty.
Figure 6. KM survival curves and extrapolated parametric survival models for unweighted JMDB data

Restricted to 72 months
Table 5 and Figure 7 show the corresponding results for the analyses applied to JMDB data re-weighted by FRAME using the propensity weighting approach described above. The results are similar to the unweighted analysis, except that generally the level of uncertainty associated with the AUCs for all analyses are substantially greater than for the unweighted analysis, and whilst the Weibull distribution appears to be the best supported for the gemcitabine arm, the log-Normal does for the pemetrexed arm. Consequently the uncertainty associated with the UK CR model averaged results is now greater than for any of the parametric distributions, reflecting the fact that whilst the data (i.e. reweighted JMDB) best supports the Weibull distribution, the log-Normal distribution is best supported by the UK CR data.

**Table 5. Area Under Curve (AUC) and Deviance Information Criterion (DIC) for parametric survival models fitted separately to Gemcitabine and Pemetrexed arms of JMDB RCT weighted using inverse propensity weights.**

<table>
<thead>
<tr>
<th>OS (72 months)</th>
<th>Gemcitabine</th>
<th>Pemetrexed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AUC (SE)</td>
<td>DIC</td>
</tr>
<tr>
<td>Exponential</td>
<td>15.32 (1.32)</td>
<td>935.29</td>
</tr>
<tr>
<td>Weibull</td>
<td>14.61 (1.14)</td>
<td>933.01</td>
</tr>
<tr>
<td>Log-Normal</td>
<td>17.13 (1.59)</td>
<td>935.92</td>
</tr>
<tr>
<td>Uniform weights</td>
<td>15.13 (1.54)</td>
<td>-</td>
</tr>
<tr>
<td>UK CR Weights</td>
<td>17.08 (1.62)</td>
<td>-</td>
</tr>
</tbody>
</table>

*Results are for 72 months*
Figure 7 (A1). KM survival curves and extrapolated parametric survival models for weighted JMDB data using inverse propensity weights

Restricted to 72 months
6. Stakeholder workshop: 10\textsuperscript{th} Sept 2015

Workshop Objectives

A workshop to discuss the usefulness and acceptability to stakeholders of the analytical methods used in the case study (re-weighting, extrapolation) was held in Frankfurt on 10 September 2015.

Following presentation of the main findings of the analyses and issues raised at the Technical Webinar (2nd September 2015), workshop participants were asked to consider both types of analysis (morning session: re-weighting, afternoon session: extrapolation) and results, using the following questions:

1. Could you envisage using these approaches in your decision making process?
2. What issues might stand in the way of adopting this approach?
3. Are there situations where this approach is particularly useful (or not at all useful)?
4. How can we communicate the implications of this approach to engage a broad range of stakeholders?

A summary of the feedback from the workshop participants is provided in the following sections

**Appendix A** gives the workshop agenda

**Appendix B** lists the workshop participants

**Appendix C** gives detailed feedback transcribed from the group sessions at the workshop.
7. Summary of Workshop Feedback – Propensity Score Weighting (PW) Method

1. Could you envisage using this method in your decision making process?

- Acceptability of this method will differ across decision makers, irrespective of data, populations etc.
- HTA: Maybe. Methods are not yet well understood, a framework is needed
- Regulators: Not for use in a regulatory setting
- Pharma R&D: Yes, but only as part of sensitivity analysis, and not as a base case in submissions. Potential application to understanding the benefit-risk profile earlier in the drug development process in order to refine the design of pivotal Phase III trials
- Acceptable but only after further experience and peer-reviewed publication

2. What issues might stand in the way of adopting this method?

- A better understanding of the method is required, including the results of applying them in a wide range of applications, to better understand when they might have most impact on decision-making. Development of guidance to support appropriate use would also be helpful. Acceptance by decision makers will be required for wider adoption of the methods
- Use of this method may increase the perception (and formal assessment) of methodological uncertainty - the methods could be viewed as something of a “black box”. This may increase the risk of negative determinations by decision makers
- This method needs to be compared/contrasted to other approaches to generalising trial results (using RWD), including those being assessing GetReal WP4
- In each case there needs to be an assessment of the applicability of the RWE study to any target population - the approach might require tailoring for particular jurisdictions
- A key analytical concern is the potential effect and impact of unobserved covariates. A potential solution to this was proposed. The outcome in the RWE dataset should be compared with the outcome in the re-weighted placebo/control arm of the trial: substantial differences between these would challenge the validity of the assumption of no unmeasured confounders
- A challenge to applying the method is where there have been changes in treatment over the time period for which RWD was collected
- Appropriate summary statistics for sub-groups in RWE populations are required in order to apply entropy re-weighting – this is a method which involves a reweighting scheme that directly incorporates covariate balance into the weights subject to pre-specified constraints.
• Consideration needs to be given to the level and timing of investment in RWE generation or collection.

• The implementation of this method/approach may be limited by the quality and diversity of many RWE data sources, but this could be ameliorated with greater co-operation and planning.

• There are privacy concerns related to making RWE available for this type of analysis, especially when RWE are derived from electronic health records.

• This method needs to be compared/contrasted to other approaches to generalising trial results (using RWD).

• There is lack of consideration of RWE and inclusion of RWE experts in the design of Phase III trials. Drawing on their expertise would (a) help in the design of the trials themselves, but also (b) facilitate the combined use of trial and RWE data in the future.

3. Are there situations where this method is particularly useful (or not at all useful)?

• Obviously the method will be more readily applicable when RWE are already available.

• Further analyses (range of case studies) are required in order to demonstrate in what circumstances this method may have an impact. ‘Easy wins’ will build trust in the method.

• As well as further examples, simulation studies are required to understand how well the methods perform, and not just whether they give different results to analyses only considering the RCT data.

• Validation of the method is first required, i.e. first demonstrate that there is not a treatment-covariate interaction. Otherwise the methods may produce erroneous effect estimates.

• Validation exercises could be based on combining (historic) Phase II trial data with RWE and comparing with later reported results (Phase 3 etc.)

• To avoid the problem of lack of comparability in a re-weighting approach (not just in terms of patient characteristics, but for example different health systems), a comprehensive cohort design could be used in which patients who do not consent to randomisation, but who do to follow-up, enable a comparison of treatment effects both within a RCT setting and in a RWE setting. This could overcome many of these issues.

• The method may have greater applicability in managed access schemes, to help mitigate risks to payers. Also in adaptive pathways, to help understand uncertainty at each decision point.

• If a PW analysis is anticipated, prior to starting a RCT the proposed inclusion/exclusion criteria should be checked against the RWD population planned to be used in the PW analysis. Relaxing exclusion criteria may be needed to enable sufficient subjects to be available for PW analysis.

• If you only have a single arm trial, eg rare diseases, you can use the approach to create comparisons with other treatments.

• The method may be useful where it shows that expected effectiveness deviates from observed clinical trial efficacy, indicating a potential benefit (or not) for a PRCT.
• Method is less useful where RCT treatment arms reflect current practice well (whether an adaptive trial or not)

• Where a RCT shows high efficacy (leap forward) supplemental RWE analyses might be required by HTA (but there is less incentive for Pharma to conduct a costly PRCT). Support for the high level of effectiveness is given by using this method to demonstrate a (likely) small efficacy-effectiveness gap

• If new RWD emerges during Phase 2b RCT, the PW analysis could inform decisions on the subsequent development programme (eg mix of Phase 3 and other studies such as PRCT)

**4. How can we communicate the implications of this method to engage a broad range of stakeholders?**

• Execute a variety of case studies in multiple indications looking at the methodological, clinical, and decision-making perspectives, so that the method can be stress-tested and different stakeholders can appreciate the potential uses of the method from their own perspective. Multiple (published) case studies will allow an assessment of the consistency/robustness (hence credibility) of the method, and help remove the ‘black-box’ perception.

• A suggested framework:
  - Understanding the method
  - Validation of the method
  - Peer-reviewed publication (both methodological and applied/tutorial)
  - Guidelines/guidance
  - Training/education

• Communicate value of method in R&D decision making, especially planning RCTs – selecting more generalisable populations. Issues are data availability and the involvement of RWD experts in design of clinical trial programme.

• Use of RWD will be facilitated if communication with clinicians responsible for the data is clear, where possible using the frame of clinical questions.

• Seek for increased alignment within the statistical community.
8. Summary of Workshop Feedback – Extrapolation Method

1. Could you envisage using this method in your decision making process?
   • Extrapolation is already used extensively in HTA, and therefore use of RWE for this purpose faces fewer barriers than its use in other analyses e.g. PW studies (above).
   • Could be extremely useful in R&D planning, but use in HTA submissions would need to be carefully justified.
   • Potentially when there are differences of opinion as regards which survival model to use (there may be concern about the wide range of possible model-based extrapolations) – RWE could be used to reconcile these.
   • Important to identify as many long-term data sources as possible (to serve as basis of extrapolation), to confirm variability.

2. What issues might stand in the way of adopting this method?
   • When different RWD sources indicate that different survival models are appropriate, how to reconcile results using different RWD sources?
   • The quality and relevance of the RWD is absolutely critical to acceptance
   • A key related question is how far back to go when trying to source RWE – clearly the further back the less validity, due to changing treatment practices etc.
   • If extrapolation effect estimates are uncertain then this could have implications for trial design, e.g. decision to extend follow-up time in RCT rather than accept uncertainty associated with extrapolation.
   • Complexity of the approaches need to be justified in terms of the benefits that these might bring – a single measure of effect with built-in uncertainty or a range of measures of effect which need to be subjectively assessed.
   • Combination of re-weighting and extrapolation is attractive, but the same source of RWE should be considered for both aspects of the modelling.
   • Need to check the fit first (using data from the RCT) to exclude some possible models, before application to extrapolation
   • Use of long-term RWE may help fit extrapolation models when the sequence of outcomes is variable rather than suitable for a single survival curve. Can consider building a variety of extrapolation curves by choosing different time points at which to introduce RWE
   • Early planning is important: If extrapolation is expected to be uncertain, this may have implications for trial design (extend duration, change eligibility criteria). If RWD sources are inadequate, may need to invest to have better RWE by the trial end
• The long-term follow-up of patients in early Phase trials may also enable validation of RWE, albeit with smaller trial populations.

• There are clearly dangers in using RWE to extrapolate the results of RCTs when the underlying populations differ or treatment practices have changed over time. There is also the danger of extrapolating a treatment effect into the future - further sensitivity analyses will be required to explore the impact of this on any decision making process.

3. Are there situations where this method is particularly useful (or not at all useful)?

• When some survival models have already been excluded by examining their fit to the actual RCT data first.

• It may be useful to produce multiple extrapolated estimates using multiple time points at which to start blending RCT and RWE data. Constraints may be introduced at earlier time points, however this may have the impact of implicitly down-weighting the RCT data, which may not be considered acceptable.

• Within adaptive pathways or patient access schemes (multi-stakeholder approach). Probably less applicable in traditional (regulatory) settings

• More suited to chronic conditions (long-term outcomes unmeasured in RCTs)

• May be more feasible where there is access to an independent multi-stakeholder registry, which can be planned to contain the required variables

• This approach is particularly useful for UK NICE (who are interested in extrapolating survival and quality-adjusted survival) – need to discuss the implications for Germany, France, EMA

• Can be used within Pharma R&D to support design of trials, interim analyses, planning long-term follow-up etc.

• At reassessment (for HTA), an assessment can be made of how true were the initial (projected) estimates

4. How can we communicate the implications of this method to engage a broad range of stakeholders?

• Use of more (and different) case studies to educate decision makers, that the methods offer something over and above those which are currently used and that there is a clear need for them

• More evidence of the robustness of the method: consistent and ongoing evaluation through case studies

• Proposed framework: Understanding of methodology -> validation of methods -> peer-reviewed publications (both methodological and applied/tutorial) -> guidance -> training/education.
• Decision makers will need to be educated in these new methods and their rationale: should not be put off by current HTA views, keep pushing

• The case study demonstrates the value of combining two approaches/methods (weighting and extrapolation) potentially using the same RWD, both of which are required – with implications (challenges) for communication
References


Royston P, Lambert PC. Flexible Parametric Survival Analysis Using Stata: Beyond the Cox


## APPENDIX A. Workshop Agenda

<table>
<thead>
<tr>
<th>Time</th>
<th>Activity</th>
<th>Presenter</th>
</tr>
</thead>
<tbody>
<tr>
<td>09:30</td>
<td>Registration &amp; Coffee</td>
<td>NA</td>
</tr>
<tr>
<td>10:00</td>
<td>Introductions</td>
<td>Simone Thomsen &amp; Michael Lees</td>
</tr>
<tr>
<td>10:15</td>
<td>Presentation – Generalizability of RCT Challenges &amp; Reweighting Case study</td>
<td>Keith Abrams, Michael Happich</td>
</tr>
<tr>
<td>11:10</td>
<td>Presentation – Generalizability - Adaptive pathways</td>
<td>Mark Trusheim</td>
</tr>
<tr>
<td>11:30</td>
<td>Breakout session 1 (Reweighting)</td>
<td>Group work Group facilitators: Michael Lees, Rob Thwaites, Pall Jonsson, Stefan Schwoch</td>
</tr>
<tr>
<td></td>
<td>Participants divided into groups to discuss the following questions on feasibility, acceptability and differences to existing approaches:</td>
<td>Group work Group facilitators: Michael Lees, Rob Thwaites, Pall Jonsson, Stefan Schwoch</td>
</tr>
<tr>
<td></td>
<td>1.) Could you envisage using this approach in your decision making process?</td>
<td>Group work Group facilitators: Michael Lees, Rob Thwaites, Pall Jonsson, Stefan Schwoch</td>
</tr>
<tr>
<td></td>
<td>2.) What issues might stand in the way of adopting this approach?</td>
<td>Group work Group facilitators: Michael Lees, Rob Thwaites, Pall Jonsson, Stefan Schwoch</td>
</tr>
<tr>
<td></td>
<td>3.) Are there situations where this approach is particularly useful (or not at all useful)?</td>
<td>Group work Group facilitators: Michael Lees, Rob Thwaites, Pall Jonsson, Stefan Schwoch</td>
</tr>
<tr>
<td></td>
<td>4.) How can we communicate the implications of this approach to engage a broad range of stakeholders?</td>
<td>Group work Group facilitators: Michael Lees, Rob Thwaites, Pall Jonsson, Stefan Schwoch</td>
</tr>
<tr>
<td>12:45</td>
<td>Lunch</td>
<td></td>
</tr>
<tr>
<td>13:45</td>
<td>Plenary: feedback from breakout session</td>
<td>Facilitators</td>
</tr>
<tr>
<td>14:15</td>
<td>Presentation – Generalizability of extrapolation findings</td>
<td>Keith Abrams</td>
</tr>
<tr>
<td>14:40</td>
<td>Breakout session 2 (Extrapolation)</td>
<td>Group work Group facilitators: See above</td>
</tr>
<tr>
<td></td>
<td>“World café” format, break out into four groups to discuss feasibility, acceptability (see above)</td>
<td>Group work Group facilitators: See above</td>
</tr>
<tr>
<td>Time</td>
<td>Activity</td>
<td>Facilitator</td>
</tr>
<tr>
<td>-------</td>
<td>-----------------------------------</td>
<td>---------------</td>
</tr>
<tr>
<td>15:40</td>
<td>Tea/Coffee</td>
<td></td>
</tr>
<tr>
<td>16:00</td>
<td><strong>Plenary:</strong> feedback from breakout session</td>
<td>Facilitators</td>
</tr>
<tr>
<td>16.30</td>
<td><strong>Workshop summary</strong></td>
<td>Keith Abrams</td>
</tr>
<tr>
<td>17.00</td>
<td><strong>Workshop end</strong></td>
<td></td>
</tr>
</tbody>
</table>
# APPENDIX B. Workshop Participants

**IMI GetReal WP1 Workshop**  
**Generalizability – Re-Weighting of RCTs**  
**September 10th 2015, Frankfurt, Germany**

<table>
<thead>
<tr>
<th>Name</th>
<th>Role</th>
<th>Organisation / Company</th>
<th>IMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Francesca Cerreta</td>
<td>Regulatory</td>
<td>EMA, UK</td>
<td>YES</td>
</tr>
<tr>
<td>2 David Bowen</td>
<td>Regulatory</td>
<td>EMA, UK</td>
<td>YES</td>
</tr>
<tr>
<td>3 Chris Chinn</td>
<td>Industry</td>
<td>Sanofi, UK</td>
<td>YES</td>
</tr>
<tr>
<td>4 Elaine Irving</td>
<td>Industry</td>
<td>GSK, UK</td>
<td>YES</td>
</tr>
<tr>
<td>5 Houria Mouas</td>
<td>HTA</td>
<td>HAS, France</td>
<td>YES</td>
</tr>
<tr>
<td>6 Alison Lightbourne</td>
<td>Patient Organisation</td>
<td>IAPo</td>
<td>YES</td>
</tr>
<tr>
<td>7 Pall Jonsson</td>
<td>HTA</td>
<td>NICE, UK</td>
<td>YES</td>
</tr>
<tr>
<td>8 Michael Lees</td>
<td>Industry</td>
<td>BMS, France</td>
<td>YES</td>
</tr>
<tr>
<td>9 Michael Happich</td>
<td>Industry</td>
<td>Lilly, Germany</td>
<td>YES</td>
</tr>
<tr>
<td>10 Stefan Schwoch</td>
<td>Industry</td>
<td>Lilly, Germany</td>
<td>YES</td>
</tr>
<tr>
<td>11 Mark Belger</td>
<td>Industry</td>
<td>Lilly, UK</td>
<td>YES</td>
</tr>
<tr>
<td>12 David Jenkins</td>
<td>Academia</td>
<td>University of Leicester, UK</td>
<td>YES</td>
</tr>
<tr>
<td>13 Keith Abrams</td>
<td>Academia/HTA</td>
<td>University of Leicester, UK</td>
<td>YES</td>
</tr>
<tr>
<td>14 Reynaldo Martina</td>
<td>Academia</td>
<td>University of Leicester, UK</td>
<td>YES</td>
</tr>
<tr>
<td>15 Clementine Nordon</td>
<td>Consultancy</td>
<td>Laser, France</td>
<td>YES</td>
</tr>
<tr>
<td>16 Amr Makady</td>
<td>HTA</td>
<td>ZIN, Netherlands</td>
<td>YES</td>
</tr>
<tr>
<td>17 Rachel Kalf</td>
<td>HTA</td>
<td>ZIN, Netherlands</td>
<td>YES</td>
</tr>
<tr>
<td>18 Rob Thwaites</td>
<td>Industry</td>
<td>Takeda, UK</td>
<td>YES</td>
</tr>
<tr>
<td>19 Sandro Gsteiger</td>
<td>Industry</td>
<td>Roche, Switzerland</td>
<td>YES</td>
</tr>
<tr>
<td>20 Jessica Davies</td>
<td>Industry</td>
<td>Roche, Switzerland</td>
<td>YES</td>
</tr>
<tr>
<td>21 Noemi Hummel</td>
<td>Academia</td>
<td>University of Bern, Switzerland</td>
<td>YES</td>
</tr>
<tr>
<td>22 Eva Didden</td>
<td>Academia</td>
<td>University of Bern, Switzerland</td>
<td>YES</td>
</tr>
<tr>
<td>23 Rod Taylor</td>
<td>Academia/HTA</td>
<td>University of Exeter, UK</td>
<td>NO</td>
</tr>
<tr>
<td>24 Richard Grieve</td>
<td>Academia</td>
<td>LSHTM, UK</td>
<td>NO</td>
</tr>
<tr>
<td>25 Aline Gauthier</td>
<td>Consultancy</td>
<td>Amaris, UK</td>
<td>NO</td>
</tr>
<tr>
<td>26 Julie Roiz</td>
<td>Consultancy</td>
<td>Creativ-ceutical, UK</td>
<td>NO</td>
</tr>
<tr>
<td>27 Radek Wasiak</td>
<td>Consultancy</td>
<td>Creativ-ceutical, UK</td>
<td>NO</td>
</tr>
<tr>
<td>28 Ulrich Fricke</td>
<td>Academia/HTA</td>
<td>TH Nürnberg, Germany</td>
<td>NO</td>
</tr>
<tr>
<td>29 Adrian Towse</td>
<td>Consultancy</td>
<td>OHE, UK</td>
<td>NO</td>
</tr>
<tr>
<td>30 Mark Trusheim</td>
<td>Academia</td>
<td>MIT, USA</td>
<td>NO</td>
</tr>
</tbody>
</table>
APPENDIX C. Detailed group feedback

Workshop Morning session – Propensity Score Weighting method

1. Could you envisage using this method in your decision making process?

**Group 1**
- Mixed message
- HTA  Might be stretching methodological acceptability, but depends on reason for use. Methods not well understood, methods framework needed
- R&D  We know the limitations. Use as sensitivity analysis (not base case). Methods have an HTA focus. Might help understand benefit-risk profile. Understanding methodology is key (then acceptance). Yes after consistent application, publications
- R&D  Yes – use early on for informing trial design rather than submission

2. What issues might stand in the way of adopting this method?

**Group 1**
- Current uncertainties (about method)
- Perception of stretching methodology too far – use in more moderate situations
- Method needs to be applied in more cases
- Propose development of guidelines or other parameters to support use
- Black box
- Prior knowledge
- Ensure RWE profile generalizable to local patients/practice
- Adding more uncertainty (i.e. using this method) may increase risk of ‘no’ from decision-makers

**Group 2**
- Consider alternative approaches to assess treatment effect (tested in WP4?)
- Possibility of reweighting inappropriately – e.g. choice of characteristic where there is no reason to believe in difference in biological effect. So need to be more specific in selection of characteristics to reweigh each patient (some differences between RCT and RWD populations may be irrelevant)
- Also relevant to communication
- Acceptability of method will differ across decision makers, irrespective of data, populations etc.
- Changes over time - using analysis of RWD over a period where treatments have improved (eg. breakthrough medicines introduced, or where a new drug is introduced during the clinical trial and RW outcomes are now improving - less relevant if you are aware)

**Group 3**
- Effect modifiers are treated equally regardless of potential effect: key is relative treatment effect
- Interaction between relative treatment effect and characteristics should be the starting point
- Need to understand what are the key characteristics and comorbidities (and measured in the right way)
- Adoption – contrast with current practice in addressing generalizability
- Timing of RWD collection and investment in RWD: fragmentation in available RWD, quality of data, privacy concerns
- Tally investment in RWD vs investment in translational/other research
- Highlight disconnect between emerging registries – chaos coming
- Impact of unobserved covariates - wrong to ignore outcomes data
• Compare outcomes in RWD with reweighted analysis in placebo arm/comparator
• If substantial difference there is unobserved effect, eg by omitting socioeconomic status
• Use what is learned from trials, then use ‘rule-out’ approach to understand how each could affect the results
• Practicality is difficult – desire to exactly replicate RCT population in a RWE study
• Randomised registry?
• Which variables to include? - those that modify effect and explain participation in trial
• What about subgroup level?
• Need to make it clear as to why variables chosen

3. Are there situations where this method is particularly useful (or not at all useful)?

Group 1
• Need examples of analyses to show where it has most impact
• Bridging between study areas
• Method applicable in non-RWE context
• When observational data already available
• Easy wins will build trust in methodology
• Validation of method could come from use in Phase II RCT + RWE
• Link use of method to access schemes, to mitigate risk to payers
• Use method to support regression?
• Relax RCTs and use method

Group 2
• Need both RCT and RWD evidence:
  o RCT data (IPD), but no RWD: issue is how to obtain RWD
  o RCT data (IPD), some RWD: issue is missing data
  o RCT data (PD), but aggregate RWD: issue is to get IPD
• Prior to starting any RCT: recommend checking proposed inclusion/exclusion criteria against RWD population given the PW analysis you plan to do at the end of the clinical trial (relax exclusion criteria so that at least a few subjects are available for the PW analysis, rather than none)
• If you only have a single arm trial, eg rare diseases, you can use the approach to create comparisons with other treatments:
  o Would this method be acceptable if clinical trial subjects were very different from RWD patients?
  o What happens If a key variable/measure is missing? - this is general issue with propensity scoring, or if there are other shortcomings in the trial
• Where the method shows that expected effectiveness is greater than observed clinical trial efficacy, and indicates need or not for a PCT (i.e. where we have both RWD and RCT results)
• Less useful where RCT treatment arms reflect current practice well (whether an adaptive trial or not)
• More useful in an adaptive pathways context, as you wish to understand uncertainty at each decision point
• Where RCT shows great efficacy (leap forward): supplemental RW analyses might be required by HTA, however there is less incentive for Pharma to conduct an expensive PCT. Support for high effectiveness is given by showing small efficacy-effectiveness gap
• If new RWD emerges during Phase 2b RCT, the PW analysis could inform decision on nature of subsequent development programme (e.g. mix of Ph3 and other studies such as PCT)

Group 3
• Commercial issues around trial design may be an ongoing hurdle
4. How to communicate the implications of this method to engage a broad range of stakeholders?

**Group 1**
- Explain clearly why this methodology is needed - facilitates buy-in
- Expand on principles: methodology, situations when method is appropriate
- Adjust RCTs to better reflect health systems
- Focus on niche patient/disease areas
- Compare with mixed treatment comparisons
- Need clear explanation of method
- Support with publications: 1. Issue, 2. Solution
- Need multiple case studies, in multiple indications: show consistency, robustness, stress test
- Suggest a stepwise approach: simulations/scenarios – work out when appropriate, develop guidelines
- Engagement with academic community
- Remove the ‘black box’ perception

**Group 2**
- Understand the methodology, based on independent validation, guidelines
- Effective communication
- [application of the method] needs to be pre-agreed
- Education: what does the method/results give us?
- Pragmatic Clinical trials before submission
- Get ‘better’ RCT populations
- Complimentary
- Everyone would take some risk
- Useful/easy
- Depends on populations/indication

**Group 3**
- One of the (many) pieces of evidence that is supposed to help make decisions
- More case studies needed across different examples
- RWD – need decent summary statistics across sub-groups (before) performing [ ] weighting (?)
- To make these studies credible requires:
  - Multiple examples
  - Translation to decision-makers (including those from other fields)
  - Acceptance by decision-makers (eg via simulation studies, meta-analysis)
- Getting clinicians to agree with use of RWD, need to communicate within the frame of clinical questions
- There is lack of alignment within stats community. However, the data are different, and should be expected to be so
- Data credibility: role for health authorities to provide data (could be aggregated data)
- Use to aid R&D decision-making? Issues are data availability, involvement of RWD experts in design of clinical trial programme – continue ongoing [ ]
Workshop Afternoon session – Extrapolation method

1. Could you envisage using this method in your decision making process?

Group A
- RWE to validate is fine, but what is considered valid?
- Variety of (possible) extrapolations is a real concern, so validation is absolutely necessary
- Use [ ] arm and compare with UK registry? (same distribution?)
- Identify as many long-term data sources as possible to confirm variability
- Potentially useful in adaptive environment

2. What issues might stand in the way of adopting this method?

Group A
- More rigorous research needed
- Within adaptive pathways/PAS – multi-stakeholder
- Conceptually useful – but in practice not there yet
- [ ] What are the implications for Germany, France, EMA? – (need) communication

Group B
- Use of RWD can validate the choice of extrapolation model when opinions differ over choice
- If there is a discrepancy between results from different RWD sources - how to reconcile, and validate choice of model?
- Check the fit first (using data you have from RCT) to exclude some of the possible models, before extrapolation
- Discrepancy in evidence used to extrapolate the two different arms (when using IPD) could introduce a bias?
- Build multiple extrapolation curves by choosing multiple time points to start the bending
- Addition of long-term RWE could help fit extrapolation models in conditions where the sequence of outcomes is variable rather than suitable for a survival curve
- If extrapolation is expected to have uncertainty, this could have implications for trial design: eg need to extend trial by 1 year, change eligibility criteria to pick more events
- Having decided at the outset that we want to use RWD after the clinical trial for extrapolation, then if RWD sources are inadequate we will need to invest to have better RWE by the trial end
- Benchmarking is of limited use where clinical practice in a setting is very different from the design of clinical trial
- Fewer hurdles as this type of analysis is done regularly

Group C
- Uncertainty appears to increase, however:
  - It comes down to preference of decision-makers. One estimate with inbuilt uncertainty or a table with the results of different goodness fit
  - It also comes down to assessment of quality of RWD
- Potential to incorporate constraints into modelling – e.g. curve needs to fit 1yr, 2yr, 3yr survival
- Could cause the registry data to dominate the RCT data – implicitly giving the trial lower weight
- Key barrier is where registry data are very different, but there is need for modelling anyhow
- Is the more complex approach going to add anything extra:
  - Will different country decision-makers prefer local data?
  - May help to reduce uncertainty?
- One potential worry is where (the largest part of) treatment effect is in unobserved period, which is inconsistent with methods guides
- Validate control group only?
- Impact of subsequent therapies?
- Healthy survival
- Unintended consequence where there are multiple sources of RWD (several RWD studies and only one RCT) - which do you choose? Another analysis to add into mix
- How far back would you go [to use] RWD: 5y, 10y or 20y? If you need 10y survival, need to go back at least the equivalent time (10y)
- One reason for weighting rather than constraints

3. Are there situations where this method is particularly useful (or not at all useful)?

**Group A**
- Within adaptive pathways/PAS (multi-stakeholder initiatives) - with later reassessment
- Conceptually useful – but in practice not there yet
- [ ] What are the implications for Germany, France, EMA? – needs communication
- If you can expect your product follows a similar distribution (to SOC)
- Chronic conditions
- Reweighting approach doesn’t add anything, rather introduces further variability
- Note that RCT-RWE bridging is already done
- You don’t have data on your product
- Registry has often a limited number of variables, which leads to a need for independent multi-stakeholder registry – more relevant for me-toos?
- Treatment decision is not captured – even more relevant for extrapolation

**Group B**
- Less applicable in traditional setting - i.e. to regulatory
- Issue with the impact of method
- Can be used within Pharma R&D to support design of trials, interim analyses etc.
- At reassessment, how true were the initial (projected) estimates?
- Could be more relevant for early access situations in future
- Needs to be discussed and agreed before [ ]
- Use the method for projection and validate it
- Difference between curves in same distribution is (more) important

**Group C**
- Less an issue for R&D decision-making, but in justification of submission
- Duration of follow up: planning long-term follow-up in early trials. Can help to plan future extrapolation/value

4. How to communicate the implications of this method to engage a broad range of stakeholders?

**Group B**
- Method is not yet there – more academic work is needed
- More contextualisation is required (to convince payers of value)
- Not clear enough to stakeholders on its application (more work required)
- More evidence on robustness (in order to support communication)

**Group C**
- Now we add the weighting approach to the extrapolation - want both aspects: weighting and extrapolation using the same RWD
- Communication issue/disconnect for case study different sources of RWD
- Clearly define the target population
- Educate decision makers, but don’t be put off by current HTA views – should keep pushing
- Need for consistent and ongoing evaluation of method with different case studies
- Pushing on an open door – is this the case for all EU HTA agencies?