Comparing relapse rates between Gilenya and Tecfidera: a subgroup analysis involving innovative machine learning methods

Concept document

**Table of Contents**

Background 3

Objectives, outcome and data 4

Overview of methodology 5

A. Stage 1: Data transformation 5

B. Stage 2: Model estimation 5

C. Stage 3: Descriptive statistics 6

Timeline and deliverables 7

Background on predictive analytics 8

Brief Biography for John Rigg PhD – Director, Predictive Analytics 9

References 10

1. Background

Gilenya and Tecfidera are oral treatments which are both used extensively as second-line options for relapsing-remitting multiple sclerosis (RRMS) patients after failure of BRACE products. This document proposes a proof of concept study aimed at better understanding patient profiles which would be expected to experience lower relapse rates on Gilenya compared to Tecfidera. This evidence can be used to promote greater awareness of the determinants of treatment response for Gilenya vs. Tecfidera and may be used to support the uptake of Gilenya.

This is a proof of concept study which will use innovative machine learning methods to establish whether predictors of differential treatment response between Gilenya and Tecfidera exist based on data from the Novartis MS VERO platform. Identifying predictors of differential treatment response is highly challenging from a traditional analytical perspective. On the one hand, the analytical framework needs to consider all potential predictors (including combinations of covariates) to minimise false negative findings. On the other hand, models estimated with a large set of covariates are prone to producing false positive findings, a problem known as overfitting, where the results may not be generalisable. Additional challenges include the need for methods which are robust in ‘fat’ data settings, where the number of covariates is relatively high compared to the number of available observations.

Advanced statistical methods from the field of machine learning (Hastie et al., 2009) are increasingly being applied to health care (Tomar and Agarwal, 2013) and can help overcome the analytical challenges mentioned above. Machine learning methods provide a framework where all potentially important covariates can be incorporated into the analysis, whilst overfitting is minimised through techniques such as cross-validation and regularisation. These methods are also performant with ‘fat’ data, including situations where the number of covariates exceeds the number of observations. Examples of recent machine learning applications to model differential treatment response include Tian et al. (2014), Hardin (2013) and Nassif (2012); also see Guelman et al (2014) for a similar application to identify target populations for marketing initiatives in insurance.

The proposed study is designed to provide valuable insight whilst minimising budgetary requirements. Should the study provide interesting evidence, follow-up studies would be necessary to fine-tune the methodology and identify distinct patient subgroups which may benefit from Gilenya vs. Tecfidera.

In light of the continued collaboration between IMS and Novartis as well as the innovative nature of the study, we propose a co-investment approach to this proof of concept study. In particular, we propose to allocate a modest proportion of the outstanding analytics budget already approved by Novartis (~4 weeks; to be used for an R programmer). IMS will provide free of charge machine learning expertise to define the analysis plan, oversee the project and interpret results. The machine learning expert will be John Rigg PhD, Director of Predictive Analytics. John leads development of the predictive analytics within Real-World Evidence Solutions at IMS and has many years of successfully developing and deploying machine learning solutions across industries (a brief biography for John is included at the end of this document).

We believe this is an exciting proof of concept study which could contribute substantially towards the science base and commercial opportunities for Gilenya. We would welcome an opportunity to discuss this proposal further.

1. Objectives, outcome and data

The primary objectives are to:

* + Estimate the proportion of patients initiating on Gilenya or Tecfidera who would be predicted to have a preferable response to Gilenya compared to Tecfidera;
  + Identify the main predictors of positive response to Gilenya compared to Tecfidera.

The outcome variable will be whether a patient experiences a relapse within 6 months of the index date. A ‘preferable’ or ‘positive’ response to Gilenya refers to a lower probability of a relapse for a patient on Gilenya compared to Tecfidera.

Data will be drawn from the PharMetrics PlusTM database using the latest claims data available. Patients will be identified based on first receipt of Gilenya or Tecfidera between October 2010 (Gilenya launch) and June 2013, the latest index month that would allow 6 months follow-up.

1. Overview of methodology

The exact methodology will be determined during the initial phase of the project, culminating in a Statistical Analysis Plan (SAP). The following represents the likely stages to the analysis.

* 1. Stage 1: Data transformation

Prior to model estimation (Stage 2), the data will be transformed as described by Tian et al. (2014), so that predictors of treatment response can be better identified. The loss functions for standard statistical models are not designed to maximize treatment-by-covariate response (Tian et al., 2014, Guelman et al, 2014) which is the focus of this study. The data transformation proposed by Tian et al. (2014) provides a simple yet effective way to identify differential predictors (treatment-by-covariate interactions). In summary, this involves coding the treatment variable to +1 and -1 for treatment and non-treatment respectively and multiplying all (centred) covariates by the treatment variable. As Tian et al. (2014) demonstrate, this facilitates improved identification of treatment-by-covariate interactions in a way that can be applied to any model, such as tree ensemble methods (see below).

Additional interaction (covariate-by-covariate) terms will be created to ensure all potentially important associations are captured. Note this is only necessary for the logistic regressions discussed below.

* 1. Stage 2: Model estimation

The probability of relapse will be estimated using logistic regression (with and without regularisation), Random Forests and Causal Conditional Inference Forests. These methods have been chosen since they are all relatively straightforward to implement (thereby minimising the time and cost for the project) and address the problem at-hand with increasing complexity. The performance of each model will be assessed using the Area Under the Curve (AUC) measure of prediction accuracy computed on out-of-sample data. The preferred model will be selected according to the model which maximises out-of-sample predictive accuracy. In general, more sophisticated modelling approaches would be expected to produce a higher level of accuracy. However, more simple models may perform as well, especially where significant predictors are relatively few and linear. Where more than one model produces similar levels of out-of-sample accuracy, the ‘simpler’ model will be chosen as listed by increasing level of complexity below.

The models to be estimated are:

**Logistic regression (without regularisation)**

Logistic regression is a well-known multivariate statistical technique which will provide an excellent baseline by which to assess subsequent models.

**Logistic regression with regularisation**

A drawback to conventional implementations of logistic regression is that it can struggle to return robust and stable results in the presence of a relatively high number of covariates (recall that additional interaction terms will be created thereby expanding the input variable space). Regularisation introduces a penalty to the loss function to penalise complexity, resulting in the shrinking of ‘less’ important coefficients to zero. The appropriate magnitude of the penalty is assessed using out-of-sample metrics on model accuracy. We will use Lasso regression (Tibsharani, 1996) to impose a penalty for logistic regression.

**Random Forests**

The approach above is based on logistic regression, a highly parametric approach assuming, for instance, that all covariates have a linear association with the outcome. Whilst attempts will be made to capture non-linear and interaction terms through the creation of additional covariates, tree-based methods provide a ready and more natural alternative to capture complex associations. We will use Random Forests, a highly popular machine learning algorithm developed by Breiman and Cutler (Breiman, 2001).

Random Forests involve making predictions on the basis of many decision-trees. A decision-tree is a schematic method used to illustrate the relationship between an outcome variable and different values of input variables. A single tree is prone to overfitting / high variance. Forming predictions on the basis of many trees can reduce variance and produce highly accurate solutions.

**Causal Conditional Inference Forests**

Causal Conditional Inference Forests (Guelman et al, 2014) is an extension of Random Forests specifically designed for prediction of differential treatment response. Compared to Random Forests, the algorithm implements a hypothesis test to determine each split in each tree (Hothorn et al., 2006) which helps minimise overfitting as well as explicit testing for treatment-by-covariate interaction effects. Guelman et al. (2014) demonstrates effectiveness of the methodology on both simulated data and real-world marketing insurance data, though this technique has yet to be applied to healthcare.

**Variable importance**

For the preferred model, predictor importance will be computed and reported in the standard manner. For logistic regressions, coefficients and confidence intervals will be reported. For the tree-based approaches, predictor importance will be computed by calculating the change in out-of-sample accuracy through randomly permuting each predictor in turn and re-estimating the model.

* 1. Stage 3: Descriptive statistics

Descriptive statistics will be reported showing the mean value of covariates for positive and non-positive responders to Gilenya vs. Tecfidera. Moreover, to help demonstrate the importance of key predictors in determining positive response to Gilenya vs. Tecfidera, results for a series of hypothetical / simulated patient profiles will be reported. This will help illuminate how variations in key baseline characteristics impact treatment response.

1. Timeline and deliverables

The table below provides a high-level, illustrative guideline on when Novartis may expect results from the project. The exact dates would be subject to confirmation and revision. The timetable assumes four weeks of a Biostatistics FTE resource over a 12 week (three month) period.

|  |  |
| --- | --- |
| Illustrative timetable | |
| **Date** | **Deliverable** |
| Mid September | Project Kick-off |
| End September | Finalise Statistical Analysis Plan and extract data from VERO |
| End October | Report interim findings based on logistic regression (with and without regularisation) and random forests |
| End November | Report final results for all models (logistic regression, Random Forests and Conditional Causal Inference Forests |
| Mid December | Report detailed findings for preferred model, including simulation results to illustrate predictor importance |

1. Background on predictive analytics

Predictive analytics, encompassing artificial intelligence, machine learning and data mining, spans a wide range of advanced analytical methods from the disciplines of computer science, statistics and mathematics. These highly innovative techniques have been successfully applied across many industries. They underpin applications as diverse as voice recognition systems, credit scoring and high-frequency financial trading algorithms. High profile internet-based companies such as Google, EBay and Amazon rely extensively on predictive analytics. Google, for instance, employ machine learning methods to retrieve and prioritise results in its search engine and to personalize on-line advertisements. These methods are also a pivotal part of the technology that help navigate the Google driverless car.

Predictive analytics is also gaining traction in health care. Automatic extraction of diagnostic information from medical images, such as MRI scans and x-rays, is becoming more commonplace due to the application of machine learning algorithms. Payers and providers are increasingly using a variety of risk stratification tools embedding predictive analytics, such as models to identify patients at high-risk of hospital readmission. Predictive analytics is also contributing to the evolution of personalized medicine through, for instance, identification of important genotype-phenotype relations. Physician decision-support tools based on predictive analytics represents a particularly promising area, where treatment recommendations reflect individual patient profiles.

Compared to traditional statistical approaches, predictive analytics adheres to a data-driven, inductive scientific philosophy (Breiman, 2001). For example, the choice of variables to include in a model is determined empirically rather than being pre-determined. The methods themselves have disparate intellectual origins. For instance, artificial neural networks, a popular class of algorithms, stem from neuro-science. The algorithms attempt to mimic the way the brain processes often vast streams of information.

1. Brief Biography for John Rigg PhD – Director, Predictive Analytics

John Rigg is the Director of Predictive Analytics in IMS’s Real-World Evidence Solutions. This involves application of advanced predictive modelling techniques to leverage rich healthcare datasets, enabling clients to address a range of opportunity areas in decision support and healthcare systems engagement.

John joined IMS from Detica, a market-leading provider of counter-fraud analytics technology. John headed-up the analytics services practice, led analytics thought-leadership initiatives and developed highly successful predictive analytics solutions for several tier 1 financial services clients. The counter-fraud predictive algorithms have produced substantial savings amounting to tens of millions of dollars for several clients, receiving industry recognition as highly innovative, market-leading solutions that have advanced fraud detection capability.

Prior to Detica, John developed statistical arbitrage models based on machine learning techniques for hedge fund trading strategies in financial markets. The models were traded on equities and futures markets and produced returns in the upper quartile compared to peers.

John has a PhD in Economics from Cambridge University and has held post-doctoral research positions at the London School of Economics and the University of Essex. He has peer-review publications in Economics, Social Policy and Health Economics journals.

1. References
   * + Breiman L. Random Forests. Machine Learning, 2001 Oct; 45 (1): 5-32
     + Guelman L, Guillen M, Perez-Marin A.M. A survey of personalized treatment models for pricing strategies in insurance. Insurance: Mathematics and Economics. 2014; 58: 68-76
     + Hardin DS, Rohwer RD, Curtis BH, Zagar A, Chen L, Boye KS, Jiang HH, Lipkovich IA. Understanding heterogeneity in response to antidiabetes treatment: a post hoc analysis using SIDES, a subgroup identification algorithm. J Diabetes Sci Technol. 2013 Mar 1;7(2):420-30
     + Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Series in Statistics, 2009, second edition
     + Hothorn T, Hornik K, Zeileis A. Unbiased Recursive Partitioning: A Conditional Inference Framework. Journal of Computational and Graphical Statistics, 2006; 15(3): 651–674
     + Nassif H, Wu Y, Page D, Burnside E. Logical Differential Prediction Bayes Net, improving breast cancer diagnosis for older women. AMIA Annu Symp Proc. 2012;2012:1330-9.
     + Tian L, Alizaden AA, Gentles AJ, Tibshirani, R. A simple method for detecting interactions between a treatment and a large number of covariates. 2014. http://statweb.stanford.edu/~tibs/ftp/JASA.complete-new.pdf
     + Tibshirani R. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society, 1996; B (58): 267-88
     + Tomar D, Agarwal S. A survey on Data Mining approaches for Healthcare. International Journal of Bio-Science and Bio-Technology, 2013; 5(5): 241-66.