Data from patient registries can provide valuable information on patient outcomes in a real-life setting. Patient registries are organized systems that use observational study methods to collect data for specific purposes, e.g., to observe the natural history of disease, to monitor safety, to measure quality of care, and to assess the clinical and cost burden of illness [1]. Depending on the research question at hand, patient registries or randomized clinical trials may be the best way to get a satisfactory answer; they are complementary in assessing patient outcomes. In contrast to clinical trials, a registry will include a wider population representing patients actually receiving the treatment according to clinical practice, place no restriction on treatments provided, alter assessments and their frequency as little as possible, and follow patients over a longer time frame. These characteristics make results from observational research more generalizable to a wide range of patients.

Traditionally, patient registries have been considered most appropriate for generating hypotheses. However, with increased interest in research and supporting a product through its life cycle, researchers are also looking at using these data to compare treatment effectiveness in real-world practice.

A major challenge in conducting effectiveness comparisons is avoiding bias, i.e., factors that distort the study findings by introducing systematic errors. Bias may be classified based on when it occurs: at patient selection, or at measurement or information collection. Selection bias occurs when patients are preferentially assigned to treatments (e.g., novel therapy used for the most ill patients), or when treatment settings differ (e.g., better-funded centers may use novel therapies and also provide other or better care and so obtain improved outcomes). Measurement bias can arise from inexact or invalid methods of measurement (e.g., assessment instruments may differ across sites or change with time). Information bias (also called ascertainment bias) refers to errors in measuring exposure or disease and addresses the degree to which data are valid (represent what is intended) and accurate (approximate the true value).

The strongest safeguard to reduce bias is in the design of the registry and by using analytical methods addressing the potential sources of bias. Recent initiatives have brought forth guidelines to develop and evaluate observational study data [1, 2]. There is, however, a need to further explore the statistical aspects of the analytical methods to help explain differences in outcomes and obtain unbiased estimates, and to provide practical guidance on suitable statistical analyses, which is the focus of this article.

**Analytical Methods**

Data analysis is a stepwise procedure with the obvious first step being familiarization with the data. One must consider whether data originate from an existing registry or one set up specifically for the research question. Whether the data distribution is normal and symmetrical or skewed, will determine the use of parametric (for a normal distribution) or non-parametric methods. Clinician input is essential as clinical and statistical aspects are equally important in planning analyses and in exploring and understanding results. After these initial investigative steps, it is possible to select an appropriate analytical method.

**Longitudinal Data**

Analysis In an observational study, patients are followed over time, monitoring risk factors and health outcomes. The measurement sequence is the same as the order of cause and effect: first the risk factor, then the outcome. Any baseline imbalances between groups in age, gender, and/or disease severity that may influence outcome can bias statistical tests and must be adjusted.

In the Fabry Outcome Survey (FOS), a multicenter world-wide observational study designed to examine the natural history of Fabry disease and study effects of enzyme replacement therapy with agalsidase alpha [3]. Measurements of heart left ventricular mass index (LVM), a key measure of disease severity, show differences by gender (Figure 1) and by age (Figure 2) at first investigation after inclusion/treatment start. Females have lower LVM than males, and LVM increases with age. If gender and age differences are not controlled for, results may be biased in either direction, positive or negative, and lead to erroneous conclusions.

Methods to control bias during analysis include covariate adjustments, matching, and propensity scoring. Matching can also be applied in the study design phase.

**Covariate Adjustments**

A covariate is a variable that is measured in both treatment and control groups prior to an intervention and that is expected to predict the response. As the response usually is related to the covariates collected at baseline, they can be included in the modeling to control for possible differences in age and gender.

If the relationship between response and covariate is suspected to be modified by another variable, an interaction term (i.e., a score describing variables interacting with one another) is added to the model and retained if statistically significant.
Adjustment using too many covariates can lead to over fitting and spurious or incorrect associations. The “rule of thumb” is no more than one covariate for every ten observations, and less when interaction terms are present. A good model should include as few covariates as possible to prevent increase in variance, but enough covariates to allow for good adjustment.

Matching
Matching to create similar samples, in contrast to covariate adjustment, does not assume any relationship between covariates and the response. The concept is to choose variables that account for the baseline differences and use an algorithm to pair the treatment groups to as much similarity as possible, including only paired cases. There are many different matching methods, all based on case-control techniques. After appropriate matching, comparisons can be performed using standard statistical methods.

The number of baseline covariates that can be matched is limited since for every added matching variable, study groups get smaller until matching is no longer meaningful. The technique requires a large pool of potential controls to match each case, and numbers may decrease considerably over time. For example, the FOS included 1428 patients in October 2007; 692 were treated at the onset of a 5-year observational period. After matching for age, gender, kidney function, and disease severity, 23 matched pairs were available for the outcome variable of interest after 5 years. The concern here is about internal and external validity. Are these patients representative of the study population (internal validity) and can results be generalized to circumstances outside the study (external validity)?

Propensity Scoring
Propensity scoring is a technique developed to adjust treatment comparisons for multiple baseline covariates when treatment assignment is not random, but can be assumed to be unconfounded [4].

A propensity score is the conditional probability of a patient being assigned to a particular treatment given the observed covariates, and is calculated as a single weight summarizing information from all covariates. It can be used for matching or as a single covariate to reduce selection bias, specifically confounding by indication (i.e., when patients who receive a drug are inherently different from those who do not get the drug), by equating groups based on these covariates. A fundamental assumption is that there is an equal probability of receiving treatment. For example, in a registry if there are centers where only one of the treatments being compared is available, this technique cannot be employed.

An advantage of this method is that multiple covariates can be incorporated; drawbacks are the additional step of calculation, and that unobserved covariates may not be balanced.

Cross-sectional Data Analysis for Longitudinal Data
In real life, longitudinal data may not be collected at regular intervals, and in observational registry data, the absence of protocol-defined assessments often results in scattered measurements. In this case, cross-sectional data analysis can provide a “snapshot” of the frequency and characteristics of a population at a single occasion or within a short period of time. In order to study change over time in this setting, it is necessary to create a time frame, a window, around the point of interest. Cross-sectional design may also be used when making comparisons between and within groups. A within-group comparison, with patients as their own controls, is recommended if bias cannot be controlled in a between-group comparison.

Although it is difficult to establish causality since risk factors and outcome are measured simultaneously, there is the advantage of not having to wait for the outcome.

Missing Data
Missing data may bias the conclusions, and it is important to understand why the data are missing to be able to handle the nonresponse correctly. Missing data are divided into three categories: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) [1]. MAR and MNAR data potentially introduce bias, which may be controlled in the analysis of MAR, but not of MNAR data.

Handling Missing Data
Missing data should always be clearly reported and addressed within an analysis. Imputation is a common strategy in which plausible values are substituted for missing data. This approach includes imputation by last observation carried forward (LOCF) and next value carried backwards (NVCB), substituting all missing values with a worst-case value, and using a regression model to predict the missing values with previous variables as covariates. Multiple imputation, recently of greater interest, reflects uncertainty around a missing value by repeating the imputation process several times, thereby generating multiple simulated datasets from which a summary finding is calculated [5]. Although lack of simple software previously limited use of this arduous procedure, such software is now available.

Imputation may be difficult to apply to observational research because measurements can occur within a wide time range. When imputation is used, it is important to perform sensitivity analyses to ensure that the most appropriate method has been utilized. An alternative to imputation is using mixed models that take missing values into account.

Summary
Methodological challenges in observational studies of effectiveness emerge from the lack of randomization to treatment leading to concern about bias. Control of bias may be obtained by design and/or analysis, the latter being discussed with a statistical angle in this article. The three methods suggested for controlling bias are covariate adjustments, matching, and propensity scoring, but there are several other methods available. The importance of correct handling of missing data has been emphasized and several imputation techniques presented.

In the ongoing debate about hierarchy of evidence, randomized clinical trials are the gold standard at the top providing the strongest form of evidence; observational studies are in the middle and opinion at the bottom. Through increased understanding of analysis and interpretation of real-world data, the value of observational studies are increasing and receiving more recognition.

Acknowledgment
We thank the FOS investigators for sharing data and Ingrid Edsman (Edsman Medical Writing) for medical writing support on behalf of Shire Human Genetic Therapies. Authors are members of the ISPOR Patient Registry Special Interest Group’s Data Management & Analysis Working Group.

References