Effectiveness Research in Inflammatory Bowel Disease: A Necessity and a Methodological Challenge

Julia Salleron, Silvio Danese, Laurence d'Agay, Laurent Peyrin-Biroulet

Journal of Crohn's and Colitis, 2016, 1096–1102
doi:10.1093/ecco-jcc/jjw068
Advance Access publication March 3, 2016
Review Article

Review Article

Effectiveness Research in Inflammatory Bowel Disease: A Necessity and a Methodological Challenge

Julia Salleron,[a] Silvio Danese,[b] Laurence D'Agay,[c] Laurent Peyrin-Biroulet[d]

[a]Department of Biostatistics and Data Management, Institut de Cancérologie de Lorraine, Vandoeuvre-Lès-Nancy, France [b]Instituto Clinico Humanitas, Milan, Italy [c]Ferring Pharmaceuticals, Saint Prex, Switzerland [d]University Hospital of Nancy-Brabois, Department of Hepatogastroenterology, Université Henri Poincaré 1, France

*Corresponding author: Prof. Laurent Peyrin-Biroulet, MD, PhD, Inserm U954 and Department of Gastroenterology, Université de Lorraine, Allée du Morvan, 54 511 Vandœuvre-lès-Nancy, France. Tel.: + 33 3 83 15 36 61; fax: + 33 3 83 15 36 33; email: peyrinbiroulet@gmail.com

Abstract

Efficacy, safety and economic issues are the main factors influencing the use of inflammatory bowel disease [IBD]-related medications. The best level of evidence comes from randomised clinical trials. The benefit of the intervention observed in a clinical trial could be reduced once it is implemented in clinical practice: its real-life efficacy, known as effectiveness, could be questioned. That is why effectiveness research based on observational studies is required to obtain-long term data on natural history, including surgery or hospitalisation, and safety. Before starting these real-life studies, it is crucial to be aware of the inherent risks of bias and confounding, to develop a good study plan, and to select the optimal design. Even if the choice of the design is optimal and if the risks of bias and confounding are minimised, the implementation of robust statistical methodology is necessary to increase the validity of the results and allow their dissemination into clinical practice. The objective of this paper is to highlight some inherent methodological problems in effectiveness research and to review some statistical tools with a focus on IBD studies and trials.

Key Words: Effectiveness; methodology; IBD

1. Introduction

Inflammatory bowel disease [IBD] is a chronic disabling condition[1] requiring long-term medical treatment. The treatments of IBD rely on drugs such as 5-aminosalicylic acid, corticosteroids, thiopurines, cyclosporine, anti-integrin therapy and anti-tumor necrosis factor alpha [anti-TNFα] agents.[2] The main factors influencing the use of IBD-related medications are efficacy, safety, and economic issues. Efficacy refers to whether a drug demonstrates a health benefit over placebo or other intervention when tested in an ideal situation.[3] Sometimes the benefits of the intervention observed in a clinical trial reduce once it is implemented in clinical practice: its real-life efficacy, known as effectiveness, could be questioned.

The distinction between efficacy and effectiveness is sometimes not clear, and the terms are sometimes wrongly used.[4,5] Efficacy can be separated from effectiveness since effectiveness relates to how well a treatment works in practice, whereas efficacy measures how well it works in clinical trials or laboratory studies.[6] As demonstrated in asthma, these two concepts can be considered as a continuum[7] rather than a dichotomy since several steps are necessary for an efficacious therapy to be effective in clinical practice. El-Serag *et al.*[8] defined effectiveness through a combination of six factors: efficacy, access, diagnosis, recommendation, acceptance, and adherence.

The best level of evidence is considered to come from randomised controlled trials [RCTs] that are designed to maximise the internal validity[9,10] with no confounding or prescription bias. However, in

the field of IBD research, only a few patients are eligible for RCTs[11] due to inclusion and exclusion criteria aiming to select a well-defined study population. Hence, patients included in RCTs are not representative of a real-life IBD setting[11] and parameters at baseline may be different compared with those for patients in real life. In addition, the management and monitoring of the patients is more intensive and may be responsible for the high placebo effect observed in some clinical trials.[12] Follow-up is usually too short to guide long-term management strategies, especially for chronic diseases such as IBD. The main differences between efficacy and effectiveness studies are listed in Table 1.[10]

For these reasons, generalisation [external validity] from RCTs is complicated, and efficacy trials results may be overgraded by guidelines to make recommendations for everyday clinical practice. In May 2008, the Food and Drug Administration [FDA] suggested guidelines for testing the effectiveness of a new drug.[6] They promote studies 'that are of different design and independent in execution, perhaps evaluating different populations, endpoints, or dosage forms' in order to provide 'support for a conclusion of effectiveness that is as convincing as, or more convincing than, a repetition of the same study'.[6]

The objective of this paper is to highlight some inherent methodological problems in effectiveness research based on prospective observational studies, and to emphasise the importance of using appropriate methodologies for designing the studies. Specific statistical tools are also reviewed with a focus on IBD studies and trials.

2. Threats to Internal Validity: Bias and Confounding

Large computerised databases [including claims databases, patient registries, electronic medical record databases, and other routinely collected health-care data] with millions of observations on the use of drugs and biologicals are useful in assessing which treatments are most effective and safe in routine care. However, these prospective observational studies fall into intermediate levels in commonly used hierarchies of evidence,[9,13] largely because of their heterogeneity and the potential for bias and/or confounding in the results. [14]

The observed statistical association between an outcome and an intervention could be due to the effect of additional variables that might be responsible for the observed association: a confounder.[15,16] Confounding is essentially a mixing of effects[16] that occurs when a factor [confounder] associated with the intervention is also associated with the development of the disease or outcome of interest independently of the intervention. The confounding factor can affect the association between intervention and disease positively or negatively.[16] For example, confounding by indication for treatment occurs when: [1] physicians prescribe one therapy over another depending on the severity of disease; [2] effectiveness measured by the outcome of interest is different from one drug compared with another in population with various severity levels; and [3] the severity of disease is a risk factor for the outcome of interest.[14] In this case, apparent [ie estimated] treatment effects are confounded, that is they are not causal but they may actually be caused by the severity of disease that led to patients being prescribed a given treatment.[14] The observed statistical association between an outcome and an intervention could also be the result of bias,[15,17] ie systematic errors in the collection of data [sampling, disease, and exposure ascertainment] or its interpretation [Table 2].

Selection[18] is one of the most frequently observed biases and can sometimes lead to confounding by indication, since it arises when the decision to administer a certain treatment is influenced by patient characteristics.[17] In a population-based inception cohort, Lakatos *et al*.[19] showed that the time to initiation of azathioprine therapy during follow-up depended on the disease behaviour at diagnosis, ie the severity of the illness. In a population-based cohort during 1979–2011, the results of Rungoe *et al*.[20] showed an increased risk of surgery within the first year after diagnosis in current users of immunosuppressive and biological drugs, suggesting failure to respond to medical treatment. Selection bias also occurs when patients who agree to participate to the study may share a characteristic that makes them different from non-participants. For example, only patients with very severe disease may agree to participate or inversely, only patients with less severe disease. If this was the case, it would not be fair to conclude the effectiveness of the therapy for all

Table 1. Differences between efficacy and effectiveness studies.

	Efficacy studies	Effectiveness studies
Condition of realisation	Clinical trials	Clinical practice
Population	High selected, homogeneous population	Heterogeneous population
Providers	Highly experienced and trained	Representative usual providers
Intervention	Prescribed treatment regimen; strictly enforced and standardised	Variable treatment regimen with optimisation; Applied with flexibility
Follow-up	Regimented with schedule	Not fixed; long-term management
Maximised validity	Internal	External

Table 2. Sources of bias in observational studies and advantages of randomised controlled trials [RCT].

Bias	Definition	Advantages of RCT
Selection	Systematic differences among the groups being compared	Eliminated by randomisation
Attrition	Selective loss to follow-up	Close patient monitoring but not completely avoidable, due to lack of efficacy and/or adverse eventss
Misclassification	Treatment or outcome is incorrect or missing	Minimised by standardised data capture and monitoring of data records
Detection	How outcome is determined	In principle eliminated by trained, blinded study participants and personnel
Performance	Systematic differences in care other than the intervention under study	Eliminated by blinded study personnel

the patients, since the observed treatment effect may be overstated [ie the treatment may be more likely to be given to patients with a good prognosis] or understated [ie the treatment may be more likely to be given to patients with a poor prognosis].

Another bias is attrition,[18] which refers to selective loss to follow-up.[17] Loss to follow-up can affect the validity and overall outcome of a study when one group of the study has a much higher attrition rate than the other group.[21] It may lead to over- or under-estimating a treatment effect.[21] Actually, it is possible that the high attrition rate occurred because the patients were unhappy with their treatment outcomes and sought other care providers. It is also possible that the high attrition rate is due to the success of the treatment, so that patients thought they no longer needed to see a physician.

Another bias is the misclassification which occurs when an exposure or outcome is incorrect or missing.[17] When comparing two treatments [A and B], it is possible that some patients categorised as treatment A had in fact received B in the past. The treatment B may not have been recorded because it was before the time of data collection. This bias was well discussed after an analysis studying the impact of infliximab on serious infection in Crohn's disease.[22] In Lichtenstein *et al.*,[22] a potential limitation of the data is that only medication use during the registry or during the period immediately before enrolment was considered. It is possible that some patients categorised as other-treatments only had in fact received infliximab in the more distant past.[22] This misclassification can also be due to a lack of completeness. For effectiveness studies, no predefined schedule of visits or medical procedures is required and patients are not as rigorously monitored as in an RCT. As such, it is possible that some important data, such as drug exposure, outcome, or clinical data, are not captured.

One of the challenges of the effectiveness study is therefore to distinguish real treatment effects from those caused by bias or confounding. Consequently, effectiveness studies require methodology and statistical analyses to identify, and if possible eliminate, bias and/or confounding in order to not over- or under-estimate the effect of treatment.[17] Government agencies and other funders have partnered with researchers to create guidelines for evaluating the validity of claims for intervention effectiveness; the International Society for Pharmacoeconomics and Outcomes Research[14,23,24,25] has published *Guidelines for Comparative Effectiveness Research Using Nonrandomised Studies in Secondary Data Sources*. As reported by Berger *et al.*,[23] a standardised approach to reporting observational studies should be adopted [Supplementary Table 1, available as Supplementary data at *ECCO-JCC* online], similar to the CONSORT [Consolidated Standards of Reporting Trials] recommendations[26] which have been modified by STROBE [Strengthening the Reporting of Observational Studies in Epidemiology] statement.[27] Another example is the initiative of Good Research for Comparative Effectiveness [GRACE].[17] Its aim is to enhance the quality of observational comparative effectiveness research [CER] and to facilitate its use for decision-making about medical therapies.

3. Methodological Aspects

3.1. Developing a good study plan

The first point of the GRACE principles[17] is the need to establish a specific study plan in advance of conducting the effectiveness research based on observational prospective study. The objectives, the patient population, and outcome measures of interest have to be clearly defined before starting the study. It is the only way to define precisely the data to be collected and the feasibility of the data collection for the study.[17] In particular, the endpoints should be clearly stated, clinically relevant, and appropriate for measuring effectiveness.[17] Due to the lack of guidelines, the selection of IBD endpoints and the timing of their evaluation remains an extensive debate in RCTs.[28,29,30] For effectiveness research, a further constraint is the availability of relevant endpoints, since the chosen outcome must be part of daily practice. After the choice of the primary endpoint, groups of treatments have to be clearly defined and have to reflect the complexities of interventions as used in real practice settings.[17] Comparisons with a number of real-world alternatives are preferable to a single comparator. Each treatment regimen has to be clearly defined by the dosage, the duration of treatment, and its eventual discontinuation.[17]

3.2. Selecting the optimal design

Observational studies of effectiveness should be designed in order to reduce potential bias.[17,18,23] Selection of the design depends on the study question. In the case of effectiveness research, a follow-up of the patient is necessary [from induction of the treatment of interest to the outcome of interest]. The prospective cohort design [Figure 1] meets this requirement since the timing between drug exposure and outcome is well characterised.[23] Each patient cohort corresponds to a drug exposure, and clinical endpoints are compared across the cohorts. A nested case-control study is composed of patient cohorts determined by the disease status [Figure 2]. Nested case-control studies are generally used for rare diseases.[23] Even if disease

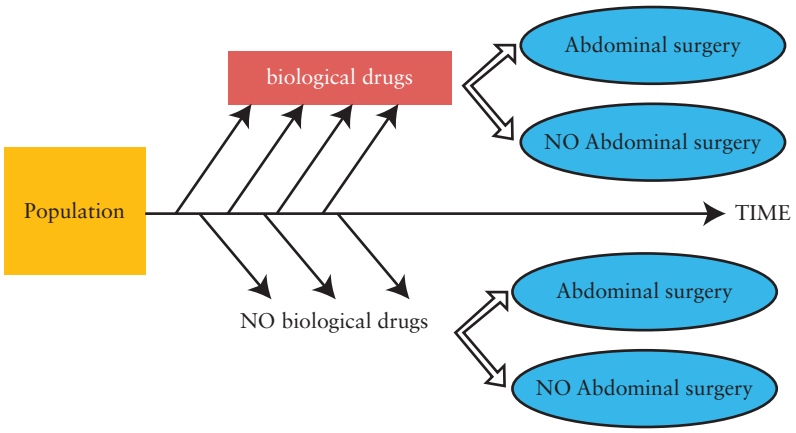

Figure 1. Illustration of cohort study design for responsiveness study to assess the relationship between biological drugs and abdominal surgery.

outcome is known for all cohort subjects, it may be too expensive to process information on covariates of interest for the entire cohort [for example, analysis of all samples of biological collection made at the time of inclusion in the cohort]. At least two groups of patients differing in outcome are then randomly selected and compared on the basis of causal factors, such as drug exposure. In these two previous designs [cohort and case-control], groups of patients may not be comparable at the induction of treatment [selection bias]. To overcome this problem, the matching case-control design [Figure 3] may be used: controls are sampled from the cohort and matched to cases on one or more attributes. The attributes are the main confounding factors.[32,33]

Waterman *et al.*[34] has published a study based on a matched case-control design. The aim was to evaluate whether abdominal surgery performed after exposure to biological therapy results in an increase in postoperative complications. The cases were the patients with exposure to biologicals [infliximab and/or adalimumab] and the controls were patients with non-exposure to biologicals within 180 days before the date of operation. Cases and controls were matched according to four factors: the main operative procedure, the IBD subtype, the exposure to preoperative treatment within 7 days, and patient age at surgery. These factors were chosen since they are well-known to affect the risk of postoperative complications. This method is sometimes difficult to implement: even if a sample size is large, there can be many confounders within many categories and data can be sparse in some strata. The number of matching factors must be reasonable given the potential number of cases and controls. Therefore, this design does not always remove all potential bias.

Another method to address bias due to heterogeneous groups is the use of propensity scores [PS].[35] The PS allows for control of many covariates [and their interactions] to be combined into a single-scale variable, namely the conditional probability of receiving the treatment given the observed covariates. PS are usually computed using logistic regression with treatment as a dependent variable and with all known and measured confounding factors as independent variables.[36] Juillerat *et al.*[37] studied the relationship between the use of proton pump inhibitors or histamine$_2$-receptor antagonists and incidence of 'flares' [hospitalisation/surgery and change in medication]. In an effort to reduce the number of variables that need to be adjusted, they matched exposed and unexposed patients to proton pump inhibitors and histamine$_2$-receptor antagonists on their PS. Propensity scores were based on all available patient characteristics. By using a matching on PS, a 'quasi-randomised' experiment is created. However, the PS only takes into account the observed covariates [cannot balance unmeasured covariates], unlike in an RCT for which randomisation of units of treatments gives a theoretical guarantee that there are no systematic differences between observed and unobserved covariates.[38]

4. Statistical Aspects

Even if the choice of the design is optimal for the research question and if the risk of bias is minimised, the implementation of robust statistical methodology is crucial to increase the validity of the results and allow their dissemination. The statistical analysis is dependent upon the research question and the study design.

4.1. Missing data

A particular challenge for effectiveness studies is the presence of non-responses and missing values, which may cause bias and cast doubt over any conclusions that can be drawn from the results.[31]

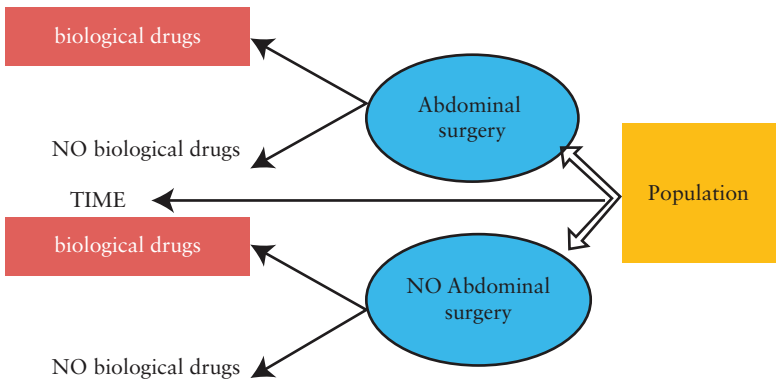

Figure 2. Illustration of nested case-control study design for responsiveness study to assess the relationship between biological drugs and abdominal surgery.

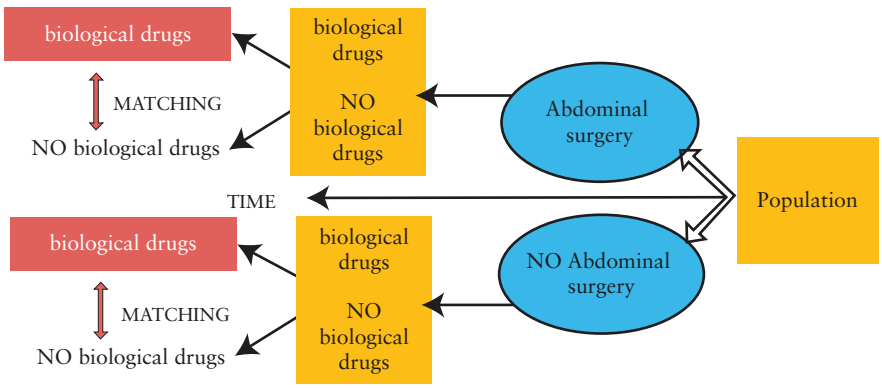

Figure 3. Illustration of matching case-control study design for responsiveness study to assess the relationship between biological drugs and abdominal surgery.

In classical multivariate analyses [linear regression, logistic regression, or survival analyses], patients are excluded if they are missing any values for a variable included in the model.[24] Consequently, few missing observations across a number of variables can lead to an important loss of analysed patients. This kind of naïve analysis without adequate handling of missing data is known to lead to biased and less robust results.[31] The simplest strategy consists in imputing the mean value for each missing observation or imputing the predicted value from a regression model. This approach is not recommended since it reduces the variability in the data.[31] Multiple imputation [MI] has been shown to produce unbiased parameter estimates which reflect the uncertainty associated with estimating missing data.[31] MI procedure replaces each missing value with a set of plausible values that represent the uncertainty about the right value to impute.[31,39] These multiply-imputed data sets are then analysed by using standard procedures for complete data and combining the results from these analyses.[39] Forrest *et al.*[40] implemented MI in the context of a clinical study assessing the effectiveness of anti-TNF therapy. In this study, the data were extracted from the registry ImproveCareNow which comprised longitudinal records collected during outpatient encounters. The Short Pediatric Crohn's Disease Activity Index [sPCDAI] was an eligibility criterion, 8% of the components were missing, and 41% of visits had at least one missing component. MI of components permits computation of the sPCDAI for all patients and without excluding too many patients.

4.2. Multivariate analyses

Multivariate models are the most common statistical method used to identify and adjust for factors predictive of disease course and outcome. Multivariate models can handle large numbers of covariates simultaneously, provided that the concept of parsimony is respected: it is more desirable to explain the outcome of interest with a simple model rather than a complex one.[41] In order to evaluate the number of confounders which can be taken into account in the model, some recommendations have been published. Peduzzi and Concato[42] evaluate the effect of the number of events per variable [EPV] analysed in proportional hazards regression analysis. For EPV values of 10 or greater, no major problems occurred. For EPV values less than 10, regression coefficients are biased in both positive and negative directions. Consequently, if 50 patients undergo surgery among a population of 500 patients, five factors can be included in multivariate analysis.

The use of the propensity score can provide clearly improved estimates of drug effects when one has relatively few events compared with the number of potentially important confounders.[43] Use of the propensity score provides an effective way to reduce the dimensionality of the covariates before modelling. Peyrin-Biroulet *et al.*[44] identified independent predictors of surgery using a multivariate Cox proportional hazard model. Adjustment to the propensity score was performed to take into account the likelihood of being treated with a Crohn's disease-specific medication according to patient characteristics. They adjusted the treatment's effect by computing the propensity score, reflecting the probability of patients with Crohn's disease being treated with disease-specific medications.

Time-dependent analysis[45] is a particular case of multivariate analyses. Risk factors included in multivariate analysis are typically fixed covariates [their value will not change over the time] and mostly known at the start of the study. Fixed covariates include demographic information and baseline characteristics. Contrary toa fixed covariate, a time-dependent covariate refers to a covariate that is not necessarily constant through the whole study and has different values at different time points. This is the case with treatment exposure. The naïve possibility to study the effect of a treatment is to perform a traditional Cox model by considering the treatment as a fixed factor, ie considering that the patient is treated whatever the duration of treatment and time of induction. It may seem more appropriate to use a time-dependent Cox regression model that takes into account that treatment may change over time.[46] With time-dependent covariates, the ability to predict is usually lost because the values are unknown at baseline.

4.3. Stratified analyses

Stratified analysis is a fundamental methodology used in observational studies, such as cohort studies. It involves allocating data into subcategories, called strata, so that each subcategory can be observed separately.[24] Subsequently, confounding factors are categorised and the sample is divided into strata, according to the number of categories of confounders.[24] Interaction testing is then performed to assess the homogeneity of treatment effect between the strata. If the treatment effect is comparable within each stratum [no interaction detected], pooled effect estimates can be calculated. If treatment effect is different across the strata, results cannot be pooled and have to be interpreted separately. In a cohort study where patients are not always treated at the same period of their disease courses, the time of prescription has to be taken into account since the effect of biologicals may vary to the time of prescription. Rungoe *et al.*,[20] in a population-based cohort during 1979–2011, analysed two subgroups of patients based on when the drug was initiated [< 1 or > 1 year after diagnosis] and compared the rate of surgery among current drug users according to duration of use. In the subgroup of patients initiating treatment on either 5-amino-salacylic acid, topical or oral corticosteroids, or azathioprine, within the first year after diagnosis, the risk of surgery was significantly higher for patients treated for 3–11 months than those treated for less than 3 months. However, in patients initiating treatment more than 1 year after diagnosis, only the use of oral corticosteroids for > 3 months reduced the risk of surgery. The treatment effects were different across the subgroups corresponding to a time-stratified effect.[47]

It should be noted that stratified analysis is feasible if there are not a lot of strata and if a only few confounders have to be controlled. When the number of potential confounders is large, stratification on the propensity score can offer a solution. Subjects are ranked according to their estimated propensity score and then stratified into subsets based on previously defined thresholds. In 1968, Cochran demonstrated that stratifying on the quintiles of a continuous confounding variable eliminated approximately 90% of the bias due to that variable.[48] Within each propensity score stratum, treated and untreated subjects will have roughly similar values for the propensity score. Therefore, when the propensity score has been correctly specified, the distribution of potential confounding factors will be similar between treated and untreated subjects within the same PS stratum.[48]

Nevertheless, the efficiency of a study can be affected dramatically by stratifying the data.[49,50] The smaller the numbers within each stratum, the more extreme the variation in the results across the strata is likely to be. The rationale for stratified analyses should be defined before the study in order to strengthen the credibility of any findings and to have the statistical power to detect clinically significant differences.[51,52] It should also be borne in mind that the false-positive rate increases as the number of comparisons increases.[52] For this reason, interpretation of findings should always be made with great caution.

5. Conclusion

Efficacy evidence is generated from interventional studies, whereas effectiveness evidence is generated from real-life studies. The two are not mutually exclusive, as they provide different types of information for the clinicians. To generate robust and clinically relevant data, effectiveness studies have to be designed as prospective, with robust and thorough methodology. This should be predefined in order to minimise the risk of bias and confounders. In this regard, both the CESAME and I-CARE[53] cohort studies have generated data that provide unique information in the field of IBD. Further effectiveness studies exploring newly approved IBD drugs, for example budesonide MMX,[54] should be encouraged, to provide further evidence on effectiveness to support clinicians in their daily practice and inform therapeutic guidelines.

Funding

This work was supported by Ferring Pharmaceuticals, St Prex, Switzerland, for the medical writing assistance.

Conflict of Interest

JS received consulting fees for writing this article; LA is an employee of Ferring; SD has served as a speaker, consultant, and advisory board member for Schering-Plough, Abbott Laboratories, Merck & Co, UCB Pharma, Ferring, Cellerix, Millenium Takeda, Nycomed, Pharmacosmos, Actelion, Alpha Wasserman, Genentech, Grunenthal, Pfizer, Astra Zeneca, Novo Nordisk, Cosmo Pharmaceuticals, Vifor, and Johnson & Johnson; LPB has receivedonsulting fees from Merck, Abbvie, Janssen, Genentech, Mitsubishi, Ferring, Norgine, Tillots, Vifor, Therakos, Pharmacosmos, Pilège, BMS, UCB-Pharma, Hospira, Celltrion, Takeda, Biogaran, Boerhinger-Ingelheim, Lilly, Pfizer, HAC-Pharma, Index Pharmaceuticals, Amgen, and Sandoz, and lecture fees from Merck, Abbvie, Takeda, Janssen, Takeda, Ferring, Norgine, Tillots, Vifor, Therakos, Mitsubishi, and HAC-Pharma.

Acknowledgments

We would like to acknowledge Anders Nordlund, senior statistician, Ferring Pharmaceuticals, Copenhagen, Denmark, for his contribution to the review of all the drafts of the manuscript.

Author Contributions

JS, writing of the manuscript; SD, final approval of the manuscript; LA, critical review of the manuscript, final approval of the manuscript; LPB, writing of the manuscript and study supervision.

Supplementary Data

Supplementary data are available at *ECCO-JCC* online.

References

1. Peyrin-Biroulet L, Cieza A, Sandborn WJ, *et al*. Development of the first disability index for inflammatory bowel disease based on the international classification of functioning, disability and health. *Gut* 2012;**61**:241–7.
2. Watson RR, Preedy VR. *Bioactive Food as Dietary Interventions for Liver and Gastrointestinal Disease*. San Diego, CA: Academic Press, 2013.
3. Thaul S. *How FDA Approves Drugs and Regulates Their Safety and Effectiveness*. CreateSpace Independent Publishing Platform, createspace.com, 2012.
4. Sands BE, Katz S, Wolf DC, Feagan BG, *et al*. A randomised, double-blind, sham-controlled study of granulocyte/monocyte apheresis for moderate to severe Crohn's disease. *Gut* 2013;**62**:1288–94.
5. Stack WA, Jenkins D, Vivet P, *et al*. Lack of effectiveness of the platelet-activating factor antagonist SR27417A in patients with active ulcerative colitis: a randomized controlled trial. The Platelet Activating Factor Antagonist Study Group in Ulcerative Colitis. *Gastroenterology* 1998;**115**:1340–5.
6. U.S. Department of Health and Human Services Food and Drug Administration Center for Drug Evaluation and Research [CDER]. *Guidance for Industry Providing Clinical Evidence of Effectiveness for Human Drug and Biological Products*. Rockville, MD: U.S. Dept. of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research : Center for Biologics Evaluation and Research, 1998.
7. Price D, Hillyer EV, van der Molen T. Efficacy versus effectiveness trials: informing guidelines for asthma management. *Curr Opin Allergy Clin Immunol* 2013;**13**:50–7.
8. El-Serag HB, Talwalkar J, Kim WR. Efficacy, effectiveness, and comparative effectiveness in liver disease. *Hepatology* 2010;**52**:403–7.
9. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N Engl J Med* 2000;**342**:1887–92.
10. Singal AG, Higgins PDR, Waljee AK. A primer on effectiveness and efficacy trials. *Clin Transl Gastroenterol* 2014;**5**:e45.
11. Ha C, Ullman TA, Siegel CA, *et al*. Patients enrolled in randomized controlled trials do not represent the inflammatory bowel disease patient population. *Clin Gastroenterol Hepatol* 2012;**10**:1002–7.
12. Akehurst R, Kaltenthaler E. Treatment of irritable bowel syndrome: a review of randomised controlled trials. *Gut* 2001;**48**:272–82.
13. Concato J. Observational versus experimental studies: what's the evidence for a hierarchy? *NeuroRx* 2004;**1**:341–7.
14. Cox E, Martin BC, Van Staa T, *et al*. Good research practices for comparative effectiveness research: approaches to mitigate bias and confounding in the design of nonrandomized studies of treatment effects using secondary data sources: the International Society for Pharmacoeconomics and Outcomes Research Good Research Practices for Retrospective Database Analysis Task Force Report - Part II. *Value Health* 2009;**12**:1053–61.
15. Viswanathan M, Berkman ND, Dryden DM, Hartling L. *Assessing risk of bias and confounding in observational studies of interventions or exposures: further development of the RTI Item Bank*. Rockville [MD]: Agency for Healthcare Research and Quality [US]; 2013.
16. Pearce N, Greenland S. Confounding and interaction. In: Ahrens W, Pigeot I, editors. *Handbook of Epidemiology*. Berlin Heidelberg: Springer, 2005.
17. Dreyer NA, Schneeweiss S, McNeil BJ, *et al*. GRACE principles: recognizing high-quality observational studies of comparative effectiveness. *Am J Manag Care* 2010;**16**:467–71.
18. Ellenberg JH. Selection bias in observational and experimental studies. *Stat Med* 1994;**13**:557–67.
19. Lakatos PL, Golovics PA, David G, *et al*. Has there been a change in the natural history of Crohn's disease? Surgical rates and medical management in a population-based inception cohort from Western Hungary between 1977–2009. *Am J Gastroenterol* 2012;**107**:579–88.
20. Rungoe C, Langholz E, Andersson M, *et al*. Changes in medical treatment and surgery rates in inflammatory bowel disease: a nationwide cohort study 1979–2011. *Gut* 2014;**63**:1607–16.
21. Krishnan E, Murtagh K, Bruce B, Cline D, Singh G, Fries JF. Attrition bias in rheumatoid arthritis databanks: a case study of 6346 patients in 11 databanks and 65,649 administrations of the Health Assessment Questionnaire. *J Rheumatol* 2004;**31**:1320–6.
22. Lichtenstein GR, Feagan BG, Cohen RD, *et al*. Serious infections and mortality in association with therapies for Crohn's disease: TREAT registry. *Clin Gastroenterol Hepatol* 2006;**4**:621–30.
23. Berger ML, Mamdani M, Atkins D, *et al*. Good research practices for comparative effectiveness research: defining, reporting and interpreting nonrandomized studies of treatment effects using secondary data sources: the ISPOR Good Research Practices for Retrospective Database Analysis Task Force Report-art I. *Value Health* 2009;**12**: 1044–1052.
24. Johnson ML, Crown W, Martin BC, *et al*. Good research practices for comparative effectiveness research: analytic methods to improve causal inference from nonrandomized studies of treatment effects using sec-

ondary data sources: the ISPOR Good Research Practices for Retrospective Database Analysis Task Force Report - Part III. *Value Health* 2009;**12**:1062–1073.
25. Schneeweiss S. On Guidelines for Comparative Effectiveness Research Using Nonrandomized Studies in Secondary Data Sources. *Value Health* 2009; **12**:1041.
26. Moher D, Schulz KF, Altman D. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *JAMA* 2001; **285**:1987–91.
27. Von Elm E, Altman DG, Egger M, *et al*. Strengthening the Reporting of Observational Studies in Epidemiology [STROBE] statement: guidelines for reporting observational studies. *BMJ* 2007;**335**:806–8.
28. Sandborn WJ, Feagan BG, Hanauer SB, *et al*. A review of activity indices and efficacy endpoints for clinical trials of medical therapy in adults with Crohn's disease. *Gastroenterology* 2002;**122**:512–30.
29. Mazzuoli S, Guglielmi FW, Antonelli E, Salemme M, Bassotti G, Villanacci V. Definition and evaluation of mucosal healing in clinical practice. *Dig Liver Dis* 2013;**45**:969–77.
30. Annaházi A, Molnár T. Optimal endpoint of therapy in IBD: An update on factors determining a successful drug withdrawal. *Gastroenterol Res Pract* 2015;**2015**:832395.
31. Rubin DB, Little RJ. *Statistical Analysis with Missing Data*. 2nd edn. New York, NY: Wiley, 2002.
32. Fisher L, Patil K. Matching and unrelatedness. *Am J Epidemiol* 1974;**100**:347–9.
33. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology* 1999;**10**:37–48.
34. Waterman M, Xu W, Dinani A, *et al*. Preoperative biological therapy and short-term outcomes of abdominal surgery in patients with inflammatory bowel disease. *Gut* 2013; **62**:387–94.
35. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983;**70**:41–55.
36. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. *Am J Epidemiol* 2006;**163**:1149–56.
37. Juillerat P, Schneeweiss S, Cook EF, Ananthakrishnan AN, Mogun H, Korzenik JR. Drugs that inhibit gastric acid secretion may alter the course of inflammatory bowel disease. *Aliment Pharmacol Ther* 2012;**36**:239–47.
38. Rubin DB. Estimating causal effects from large data sets using propensity scores. *Ann Intern Med* 1997;**127**:757–63.
39. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. New York, NY:Wiley,1987.
40. Forrest CB, Crandall WV, Bailey LC, *et al*. Effectiveness of anti-TNFα for Crohn disease: research in a pediatric learning health system. *Pediatrics* 2014;**134**:37–44.
41. Rencher AC, Christensen WF. *Methods of Multivariate Analysis*. 3rd edn. Hoboken, NJ: Wiley, 2012.
42. Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. *J Clin Epidemiol* 1995;**48**:1503–10.
43. D'Agostino RB. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med* 1998;**17**:2265–81.
44. Peyrin-Biroulet L, Oussalah A, Williet N, Pillot C, Bresler L, Bigard M-A. Impact of azathioprine and tumour necrosis factor antagonists on the need for surgery in newly diagnosed Crohn's disease. *Gut* 2011;**60**:930–6.
45. Therneau TM, Grambsch PM. *Modeling Survival Data: Extending the Cox Model*. New York, NY: Springer, 2000.
46. Vernier-Massouille G, Balde M, Salleron J. *et al*. Natural history of pediatric Crohn's disease: a population-based cohort study. *Gastroenterology* 2008;**135**:1106–13.
47. Scheike TH, Martinussen T. On estimation and tests of time-varying effects in the proportional hazards model. *Scand J Stat* 2004;**31**:51–62.
48. Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* 1968;**24**:295–313.
49. Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *JAMA* 1991;**266**:93–8.
50. Dijkman B, Kooistra B, Bhandari M. How to work with a subgroup analysis. *Can J Surg* 2009;**52**:515–22.
51. Brookes ST, Whitely E, Egger M, Davey Smith G, Mulheran PA, Peters TJ. Subgroup analyses in randomized trials: risks of subgroup-specific analyses; power and sample size for the interaction test. *J Clin Epidemiol* 2004;**57**:229–36.
52. Lord SJ, Gebski VJ, Keech AC. Multiple analyses in clinical trials: sound science or data dredging? *Med J Aust* 2004;**181**:452–4.
53. Beaugerie L, Brousse N, Bouvier AM, *et al*. Lymphoproliferative disorders in patients receiving thiopurines for inflammatory bowel disease: a prospective observational cohort study. *Lancet* 2009;**374**:1617–25.
54. Travis SPL, Danese S, Kupcinskas L, *et al*. Once-daily budesonide MMX in active, mild-to-moderate ulcerative colitis: results from the randomised CORE II study. *Gut* 2014;**63**:433–41.