

Supplementary Online Content

Rudrapatna VA, Glicksberg BS, Butte AJ. Utility of routinely collected electronic health records data to support effectiveness evaluations in inflammatory bowel disease: a pilot study of tofacitinib.

RECORD Statement Checklist

Pre-Registered Protocol and Statistical Analysis Plan

(reproduced from <https://protocols.io/view/robust-measurement-of-the-real-world-effectiveness-2bqgamw>)

eTable 1. Examples of Data Abstraction from the EHR

eTable 2. Most common reasons disqualifying the real-world UC cohort from meeting the OCTAVE trial eligibility criteria

eFigure 1. Durability of tofacitinib treatment

eFigure 2. Reasons for treatment discontinuation

Supplemental References

The RECORD statement – checklist of items, extended from the STROBE statement, that should be reported in observational studies using routinely collected health data.

	Item No.	STROBE items	Location in manuscript where items are reported	RECORD items	Location in manuscript where items are reported
Title and abstract					
	1	(a) Indicate the study's design with a commonly used term in the title or the abstract (b) Provide in the abstract an informative and balanced summary of what was done and what was found	Title	<p>RECORD 1.1: The type of data used should be specified in the title or abstract. When possible, the name of the databases used should be included.</p> <p>RECORD 1.2: If applicable, the geographic region and timeframe within which the study took place should be reported in the title or abstract.</p> <p>RECORD 1.3: If linkage between databases was conducted for the study, this should be clearly stated in the title or abstract.</p>	<p>Title</p> <p>Abstract</p> <p>N/A</p>
Introduction					
Background rationale	2	Explain the scientific background and rationale for the investigation being reported	Introduction		
Objectives	3	State specific objectives, including any prespecified hypotheses	Last paragraph of Introduction		
Methods					
Study Design	4	Present key elements of study design early in the paper	Methods		
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	Methods (Patient Identification)		
Participants	6	(a) <i>Cohort study</i> - Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up	Methods	RECORD 6.1: The methods of study population selection (such as codes or algorithms used to identify subjects) should be listed in detail. If this is not possible, an explanation should be provided.	Methods (Patient Identification)

		<p><i>Case-control study</i> - Give the eligibility criteria, and the sources and methods of case ascertainment and control selection. Give the rationale for the choice of cases and controls</p> <p><i>Cross-sectional study</i> - Give the eligibility criteria, and the sources and methods of selection of participants</p> <p><i>(b) Cohort study</i> - For matched studies, give matching criteria and number of exposed and unexposed</p> <p><i>Case-control study</i> - For matched studies, give matching criteria and the number of controls per case</p>		<p>RECORD 6.2: Any validation studies of the codes or algorithms used to select the population should be referenced. If validation was conducted for this study and not published elsewhere, detailed methods and results should be provided.</p> <p>RECORD 6.3: If the study involved linkage of databases, consider use of a flow diagram or other graphical display to demonstrate the data linkage process, including the number of individuals with linked data at each stage.</p>	<p>Methods (Patient Identification)</p> <p>N/A</p>
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable.	Methods	RECORD 7.1: A complete list of codes and algorithms used to classify exposures, outcomes, confounders, and effect modifiers should be provided. If these cannot be reported, an explanation should be provided.	Methods (Covariate abstraction), Supplement (Pre-Registered Protocol and Statistical Analysis Plan)
Data sources/ measurement	8	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group	Methods (Covariate abstraction), Supplement (Pre-Registered Protocol and Statistical Analysis Plan)		
Bias	9	Describe any efforts to address potential sources of bias	Methods, Discussion, Table 2		
Study size	10	Explain how the study size was arrived at	Methods (Patient Identification)		
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If	Methods, Supplement (Pre-Registered Protocol and		

		applicable, describe which groupings were chosen, and why	Statistical Analysis Plan)		
Statistical methods	12	(a) Describe all statistical methods, including those used to control for confounding (b) Describe any methods used to examine subgroups and interactions (c) Explain how missing data were addressed (d) <i>Cohort study</i> - If applicable, explain how loss to follow-up was addressed <i>Case-control study</i> - If applicable, explain how matching of cases and controls was addressed <i>Cross-sectional study</i> - If applicable, describe analytical methods taking account of sampling strategy (e) Describe any sensitivity analyses	Methods (Data quality, completeness, handling of missing data; Statistics/computing); Supplement (Pre-Registered Protocol and Statistical Analysis Plan)		
Data access and cleaning methods		..		RECORD 12.1: Authors should describe the extent to which the investigators had access to the database population used to create the study population. RECORD 12.2: Authors should provide information on the data cleaning methods used in the study.	Methods (patient identification); Methods (Data quality, completeness, handling of missing data)
Linkage		..		RECORD 12.3: State whether the study included person-level, institutional-level, or other data linkage across two or more databases. The methods of linkage and methods of linkage quality evaluation should be provided.	N/A
Results					
Participants	13	(a) Report the numbers of individuals at each stage of the	Figure 1	RECORD 13.1: Describe in detail the selection of the persons included in the	Methods (Patient Identification),

		<p>study (<i>e.g.</i>, numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed)</p> <p>(b) Give reasons for non-participation at each stage.</p> <p>(c) Consider use of a flow diagram</p>		<p>study (<i>i.e.</i>, study population selection) including filtering based on data quality, data availability and linkage. The selection of included persons can be described in the text and/or by means of the study flow diagram.</p>	Results (Cohort Identification)
Descriptive data	14	<p>(a) Give characteristics of study participants (<i>e.g.</i>, demographic, clinical, social) and information on exposures and potential confounders</p> <p>(b) Indicate the number of participants with missing data for each variable of interest</p> <p>(c) <i>Cohort study</i> - summarise follow-up time (<i>e.g.</i>, average and total amount)</p>	Table 1		
Outcome data	15	<p><i>Cohort study</i> - Report numbers of outcome events or summary measures over time</p> <p><i>Case-control study</i> - Report numbers in each exposure category, or summary measures of exposure</p> <p><i>Cross-sectional study</i> - Report numbers of outcome events or summary measures</p>	Results, Figure 4		
Main results	16	<p>(a) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (<i>e.g.</i>, 95% confidence interval). Make clear which confounders were adjusted for and why they were included</p> <p>(b) Report category boundaries when continuous variables were categorized</p>	Results		

		(c) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period			
Other analyses	17	Report other analyses done—e.g., analyses of subgroups and interactions, and sensitivity analyses	N/A		
Discussion					
Key results	18	Summarise key results with reference to study objectives	Discussion		
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias	Discussion	RECORD 19.1: Discuss the implications of using data that were not created or collected to answer the specific research question(s). Include discussion of misclassification bias, unmeasured confounding, missing data, and changing eligibility over time, as they pertain to the study being reported.	Discussion
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence	Discussion		
Generalisability	21	Discuss the generalisability (external validity) of the study results	Discussion		
Other Information					
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based	Funding Information		
Accessibility of protocol, raw data, and programming code		..		RECORD 22.1: Authors should provide information on how to access any supplemental information such as the study protocol, raw data, or programming code.	Data Sharing Plan

*Reference: Benchimol EI, Smeeth L, Guttman A, Harron K, Moher D, Petersen I, Sørensen HT, von Elm E, Langan SM, the RECORD Working Committee. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. *PLoS Medicine* 2015; in press.

*Checklist is protected under Creative Commons Attribution ([CC BY](https://creativecommons.org/licenses/by/4.0/)) license.



May 23,
2019

Robust measurement of the real world effectiveness of Tofacitinib for the treatment of Ulcerative Colitis using electronic health records: a protocol and statistical analysis plan

Vivek A. Rudrapatna¹, Atul J. Butte¹

¹University of California, San Francisco

Other dx.doi.org/10.17504/protocols.io.2bqgamw

Vivek A. Rudrapatna
University of California, San Francisco

ABSTRACT

The "efficacy-effectiveness gap" refers to the difference between the treatment efficacy as measured by randomized controlled trials (RCTs) and treatment effectiveness as measured in "real world" clinical settings. Prior studies have documented the existence of this gap in a variety of clinical contexts and attributed its existence to a number of factors, including overly restrictive subject inclusion/exclusion criteria in RCTs and differences in treatment efficacy ascertainment.

In this protocol and statistical analysis plan, we document a protocol for a forthcoming study relevant to assessing the efficacy-effectiveness gap of the medication *Tofacitinib* as used to treat Ulcerative Colitis. Specifically, we detail the following:

- 1) What covariates are needed to compare the real-world effectiveness of Tofacitinib with its efficacy as measured by pivotal clinical trials
- 2) What approach we propose to extract these covariates from the electronic health records at our institution in a relatively reproducible fashion, balancing accuracy with practicality.
- 3) How we propose to analyze the data

GUIDELINES

The following criteria is meant to provide guidance to the chart abstractor to facilitate consistent coding. However, these criteria are not absolute, comprehensive or complete; some judgement is required. We advise that this chart extraction be done by a gastroenterologist with IBD experience, ideally also with prior clinical research experience.

- In general, you may have to employ some clinical judgment and your knowledge of the objectives of the study in order to make the proper assessments. For instance, if a patient has an elevated follow-up c-reactive protein but was found at that time to have a concomitant infection, that data must be either be excluded and another datapoint selected (as specified by the protocol to follow) or the column should be annotated as 'NA.' Similarly for situations in which the patient's inflammatory markers rise in the setting of lost insurance. The idea is to capture the data *to the extent that it exists* for the purposes of the study, and if not find a more representative alternative or annotate it as NA. Remember, the statistical models commonly used to analyze the data do not incorporate the *essential*/clinical context that only you have access to during the covariate extraction phase.
- All annotations correspond to the period of time when the drug is under consideration or actively being given, not the time of performing the chart review!
 - For instance, age and duration of disease corresponds to the patient's status at the time of initiating treatment
 - If an annotation calls for a follow-up Mayo endoscopic subscore in 2-8 months after treatment initiation and the patient has already discontinued the treatment by month 1, this variable should be annotated as 'NA' rather than by a different value that may correspond to a different treatment regimen.
- Occasionally patients are co-managed between your home institution (i.e. UCSF) and other sites whose clinic notes/records may be accessible via *care everywhere*. Use *request updates* when examining data using this portal and try to take advantage of the availability of multiple independent observations within the electronic health record to maximize the precision of our estimates.
- Using the "magnifying glass" functionality to globally search the medical record can be helpful to find evidence quickly
- Unselect the "default filter" section in order to maximize the records visible to you. Every time you switch between tabs, recheck this section to see that it remains unchecked.
- Note that the *Encounters* and *Notes* tabs contain slightly different elements; both may need to be reviewed for any given purpose.
- When reviewing patient emails, remember to look at the date above the patient "message bubble" as well as the date last read by the



patient (below the provider message), and not the date of the encounter associated with the email.

- In general, if the precise day is not known, round down to the first of the month. e.g 2016-10-?? becomes 2016-10-01. If only year known, report January 1st of that year.
- When checking dates, try to use supporting evidence (primary care notes, external GI notes available through *care everywhere* around the time of diagnosis) to confirm the date.

MATERIALS TEXT

- Electronic health record system: this protocol has been tested in the ambulatory context of *Epic*TM software (2017).
- A list of prospective (but not necessarily validated – this part is embedded into this protocol) medical record numbers corresponding to Inflammatory Bowel Disease patients prescribed Tofacitinib
- HIPAA-compliant server to store all PHI
- IRB approval to identify patient records and handle PHI

SAFETY WARNINGS

In addition to the above relevant to IRB approval and HIPAA compliance, investigators should all be trained or certified in human subjects protection (www.CITIprogram.org)

BEFORE STARTING

- We obtained approval from our institution's IRB (UCSF #18-24588) to enable reidentification of medical records corresponding to IBD patients receiving Tofacitinib. This is strongly recommended for other investigators seeking to use this protocol.
- These steps are somewhat customized for the *Epic* electronic health records at our institution. They may be applicable outside of this system but have not been formally tested and customized as such.
- We obtained access to a HIPAA-compliant server to store protected health information (PHI). This will be necessary for others seeking to use this protocol since the protocol calls for extraction, storage, and analysis of PHI elements.

Validating prospective MRNs using the 'Received Tofacitinib' spreadsheet

1 **Diagnosis:** {UC, CD, NA}

The best place to unambiguously identify a working diagnosis is a recent colonoscopy done near the time of treatment initiation. In particular, the indication, impression, and recommendations section are often quite useful.

Otherwise, use the diagnosis made at the GI encounter just preceding the prescription and initiation of the drug. This is often helpful because insurance prior authorization of a given medication often requires a clear diagnosis, which is documented in the clinic visit and sent to the payor.

Be mindful of the facts that:

- 1) the EHR can auto-populate the (sometimes incorrect) diagnosis at the very top and at the first line of the assessment and plan, and
- 2) copy-forwarded notes can obscure cases of a revised diagnosis (e.g. top of HPI states Ulcerative Colitis, and mid-way indicates Crohn's).

In the latter case, rely on what has been documented in the assessment and plan.

2 Received Tofacitinib:

{Y, N, NA}

In general, the presence of a medication order or renewal should be considered insufficient evidence that a drug has actually been taken by the patient.

The best evidence for this is a message from a patient indicating that they've started the drug, or a telephone encounter documenting as such. These often yield precise dates of initiation. Reports from colonoscopies performed to assess treatment response can indicate the date or approximate date of start.

A sequence of clinic notes indicating plans to start the drug followed by a post-initiation visit or endoscopy assessing treatment response can be helpful too. Multiple independent notes from GIs and non-GIs suggesting that the patient started a treatment is good evidence supporting this.

If the above is unavailable, historical clinical documentation (notes from treating physicians, Indications/Impression/Recommendations sections from endoscopy reports) can be quite helpful.

3 Initiated to treat IBD:

{Y, N, NA}

Sometimes patients are initiated on Tofacitinib for reasons that are unrelated to Ulcerative Colitis (for instance, rheumatological or dermatological disorders).

Score as Y if the patient was given the drug for the purposes of treating Ulcerative Colitis (this would most often be decided and ordered by a gastroenterology provider)

Score as N if not. In the setting of UC, this should also include uncommon settings where patients had subtotal colectomies and are trying a steroid-sparing agent before proceeding with completion proctectomy (e.g. these are also scored as an N). Although this logically might be a scenario where one could still assess whether or not it is effective in treating inflammation, monitoring response to treatment is more complex and may not reflect the typical performance of this drug in the setting where it has been formally studied.

If this is the case, these patients can be treated the same as those who did **not** receive the medication; their MRNs should not be transferred to the Table 1 and Efficacy sheets for further analysis (see subsequent sections).

For patients with active symptoms and minimal endoscopic IBD activity, if they were given IBD treatment in the hopes that their symptoms would improve (vs not improve because the symptoms were actually due to something else such as IBS/SIBO), code this as a Y.

If the patient did not even receive the medication (as annotated by the prior variable), code as 'NA.'

4 Inconsistent Use:

{Y, N, NA}

Annotate as Y if there is evidence on chart review that the patient was not taking the medication regularly, for example due to personal preferences. Annotate N if no chart evidence to support this. Minimize the use of NA (e.g. if there is no evidence that the patient was the drug inconsistently, annotate as N rather than NA).

Ulcerative Colitis Table 1 Spreadsheet

5 Gender:

{M, F, NA}

As ascertained from the clinic note and the demographic face sheet. If there is any evidence of a mismatch (e.g. indicating error or sex-change), score as NA

6 Date of Disease Onset:

{YYYY-MM-DD, NA}

Use a combination of the earliest notes in the system (use the *Notes* tab, not the *Encounter* tab), the *Scanned Clinical Documents* tab corresponding to the original clinic referral, and *Care Everywhere* in order to determine as precisely as possible when the patient's disease began. As above, round down for situations where only the month (e.g. 4/1/2018) or the year (1/1/2018) is known. When notes document age of onset rather than date of onset, can select the first day of the next month following the patient's birthdate. For instance if the patient is born 2/19/84 and developed the diagnosis at age 10, annotate this as 3/1/94.

For Ulcerative Colitis, the timing of clearly attributable disease onset (rectal bleeding with tenesmus/nocturnal symptoms/weight loss) might have occurred prior to the time of colonoscopic diagnosis. Patients often recall a clear timepoint as to when their symptoms began but sometimes their symptom onset is unclear or not classic for IBD (increased stool frequency without other hallmarks of inflammatory diarrhea as above). In the event of the latter use the date of colonoscopic diagnosis.

7 Age at Tofacitinib initiation (in years, rounded down):

{Integer, NA}

To be calculated based on the patient's birthdate and the date of Tofacitinib initiation. Be wary of using the age reported in clinic notes as the basis of this – these notes are often copy-forwarded and incorrect. Use the birthdate and date of drug initiation.

8 Disease duration at time of Tofacitinib initiation (# months/12):

{Rational Number, NA}

This field requires a greater degree of precision than the prior field assessing age. If chart review does not yield an exact day, use the earliest compatible day at which treatment could have been started. For instance if the patient writes an email indicating that he started treatment about 2 weeks prior to that date of communication, use 14d prior to that date.

Then, calculate disease duration in fractional units of years based on dividing the number of months with disease by 12. In other words, if the patient had the disease for 8 years and 3 months, reports $8 + 3/12$ here. This should be reported as a rational number (if using Excel, it should calculate this for you if you use the '=' prefix).

9 Disease extent at time of Tofacitinib initiation:

{1, 2, 3, NA}

1 = Proctitis or Proctosigmoiditis

2 = Left-sided colitis (extends to but not beyond the splenic flexure)

3 = Extensive or pan-ulcerative colitis (anything extending beyond the splenic flexure)

Prioritize Sigmoidoscopic/Colonoscopic data prior to Tofacitinib initiation as the gold standard to ascertain this (e.g. how far the disease had spread by the time Tofa had started). This should reflect the extent of underlying anatomical territory that has ever been endoscopically involved up to this timepoint, not just what is currently involved (e.g. if the rectum appears spared on a follow-up exam and the patient is on treatments including steroid enemas this should not affect the categorization of the disease extent).

If Sigmoidoscopy reveals disease through the extent of the exam, can also use supportive imaging data as needed to judge extent. If this is unavailable, acceptable to use note-based documentation. Be aware that imaging-based extent can often overcall extent compared to endoscopy and histology.

Under scenarios of discrepant endoscopic findings (endoscopic extent involving the transverse colon in one and not in another), annotate this field by the maximum documented extent.

It is acceptable to use data from within 6 months of the start to retroactively assess disease extent. For instance, if a clinic note indicates left-sided colitis but imaging and surgical pathology within 6 months indicate extensive/pancolitis, annotate as the latter. The assumption is that sampling error and chronic disease progression is a more likely explanation for prior misclassification than the possibility that the drug caused an acute and significant extension of anatomic involvement).

Score this based on endoscopic rather than histologic disease extent. The rationale for this is that the disease extent classification in the trials was more likely to use gross endoscopic extent rather than histologic extent. This need not fall strictly within the same strict time windows as the Mayo Score data.

10 Baseline Mayo stool frequency subscore:

{0, 1, 2, 3, NA}

We are attempting to capture this Mayo subscore within 6 months of the date of Tofacitinib initiation (e.g. **Month -6 to Month 0**). The idea here is to attempt to match our patient cohort with that of the clinical trials.

The rationale for using such a broad time window is that endoscopic scheduling and insurance approval for a planned drug, among other factors, can create a lag between the time of Mayo endoscopic subscore data availability and treatment initiation. So we are allowing the other Mayo subscore variables to capture data from the same time window.

Per the OCTAVE supplemental data protocol (Appendix 1, page 80):

0: Normal number of bowel movements for this patient

1: 1 to 2 bowel movements more than normal

2: 3 to 4 bowel movements more than normal

3: 5 or more bowel movements more than normal

The above data is sometimes documented in the chart by absolute frequency of bowel movements. Per FDA guidance a bowel movement “is defined as a trip to the toilet when the patient has either a bowel movement, or passes blood alone, blood and mucus, or mucus only.” Therefore, stool frequency should incorporate toilet trips when there is passage of any liquid or solid content through the anus, not limited only to feces.

If the stool frequency is reported as a relative increase over baseline, no ascertainment of absolute baseline frequency is required. However, if stool frequency is reported as an absolute frequency, the chart reviewer must make every effort to identify a baseline stool frequency from the chart. This can be identified from anywhere within full record, including timepoints before or after Tofacitinib use. We recommend using the global search functionality to look for keywords like: “baseline” “normal” “1-2” “2-3” “stools” “bowel movements” etc.

If the notes do not explicitly mention “baseline” but do indicate a stool frequency of between 1-3 bowel movements daily in any context (e.g. other treatments such as steroids), treat this number as the baseline. The justification for this is that population-based studies have suggested that up to 3 bowel movements daily can be considered normal.

If the patient has baseline constipation during periods of quiescence and returns to use of stool softeners/laxatives, can impute this field to 0 (e.g. assume use of stool softeners to achieve 1 bowel movement daily).

If no “baseline” nor approximate baseline is mentioned, a baseline of 3 bowel movements daily can be imputed at the chart reviewers discretion. This selection would not be expected to significantly affect the study results since the analysis will involve the paired differences in Mayo scores. If this is done an additional indicator variable denoting that this was performed should be added to the data collection spreadsheet.

When dealing with numeric ranges of bowel movements, calculate the score based on the minimum increase in stool frequency compatible with the range. For instance, if a patient has 1-2 bowel movements at baseline, and is currently having 5-6 during the morning and 2-3 over the rest of the day, this would be a difference between (1 to 2) and (7 to 9), corresponding to a minimum increase of 6 bowel movements above baseline (e.g. 1 to 7 is 6 above baseline, 2 to 9 is 7 above baseline, hence take the minimum). Therefore, this is scored as a ‘3’ based on the above rubric.

In the event that multiple such subscores exist within this window, we use the following algorithm:

1. Select the stool frequency subscore occurring closest in time to the date of treatment failure
2. If no treatment failure and a colonoscopy was performed within this period, select the stool frequency subscore occurring closest in time to the date of the colonoscopy
3. If no colonoscopy was performed in this period, select the highest available subscore within this 6 month window.

The rationale for picking the highest score within the window is that some assessments may be masked by concomitant therapy (e.g. steroids) which are somewhat harder to ascertain from the EHR.

As mentioned in the top section ‘Guidelines’: exclude any stool frequency confounded by the presence of a concomitant infection or treatment interruption. The exception to this rule is the presence of concomitant steroids. In this case, report the stool frequency; we will also capture the presence of steroid use in a separate indicator variable.

Score this field as NA if it is not satisfactorily identified.

11 Baseline Mayo rectal bleeding subscore:

{0, 1, 2, 3, NA}

We are attempting to capture this Mayo subscore within 6 months of the date of Tofacitinib initiation (e.g. **Month -6 to Month 0**). The idea here is to attempt to match our patient cohort with that of the clinical trials.

The rationale for using such a broad time window is that endoscopic scheduling and insurance approval for a planned drug, among other factors, can create a lag between the time of Mayo endoscopic subscore data availability and treatment initiation. So we are allowing the other Mayo subscore variables to capture data from the same time window.

Per the OCTAVE supplemental data protocol (Appendix 1, page 80):

0: No blood seen

1: Streaks of blood with stool less than half the time

2: Obvious blood with stool most of the time

3: Blood alone passes

This part of the score has been recognized by the FDA to be suboptimal (FDA CDER Draft Guidance, Aug 2016) for multiple reasons including the fact that it is a “double-barreled” question (e.g. asks about both amount and frequency of blood).

Presumably as a measure to correct the frequency problem, the Pfizer Protocol indicates that this should be scored “the most severe bleeding of the day.”

We have adopted the following pragmatic strategy:

1) If data of bleeding severity is available, then score based on this as follows:

o If there is clear documentation of no bleeding, score as 0.

o If there is clear documentation of episodes where only blood passes, score this as a 3.

o If there is bleeding but the amount is vague (e.g. “some”) then code this as a 1.5.

o Otherwise if there is a clear indication of degree of bleeding, use 1 or 2 as appropriate.

2) If severity data is not available but the frequency data is, then scoring using this data and the OCTAVE metric above can be performed at the discretion of the chart abstractor.

If there exist multiple datapoints within this 6-month time window, use the approach as outlined in the stool frequency section. If multiple datapoints around the same time suggest slightly different degrees of bleeding, select the most severe degree of bleeding. If the note indicates blood 50% of the time, score as 1.5.

12 **Baseline Mayo PGA subscore:**

{0, 1, 2, 3, NA}

We are attempting to capture this Mayo subscore within 6 months of the date of Tofacitinib initiation (e.g. **Month -6 to Month 0**). The idea here is to attempt to match our patient cohort with that of the clinical trials.

The rationale for using such a broad time window is that endoscopic scheduling and insurance approval for a planned drug, among other factors, can create a lag between the time of Mayo endoscopic subscore data availability and treatment initiation. So we are allowing the other Mayo subscore variables to capture data from the same time window.

Per the supplemental data:

- 0: Normal
- 1: Mild disease
- 2: Moderate disease
- 3: Severe disease

Per the Pfizer protocol used for the OCTAVE trials, this score "acknowledges the three other criteria, the patient's recollection of abdominal discomfort and general sense of wellbeing, and other observations, such as physical findings and the patient's performance status."

This score ideally should reflect a **dynamic**, "real-time" assessment of disease activity (e.g. evidence of ongoing inflammation causing symptoms that is in principle modifiable by medication). This is in contrast to annotations of severe disease based on, for example, the history of multiple prior treatment failures.

A good place to identify the physician's global assessment (PGA) is the end of colonoscopy reports, where patients are commonly characterized as "mild" or "severe" independent of scoring using the Mayo endoscopic score. Other sources of the PGA include the Assessment and Plan of clinic notes, or appeal letters to insurance; the former is preferable because it 1) is written by the attending physician, rather than fellows/residents in clinic, and 2) appeal letters to insurance may be at risk of bias.

In the event that multiple such subscores exist within this window, use the same approach as detailed in the stool frequency section.

As mentioned in the top section 'Guidelines': exclude any stool frequency confounded by the presence of a concomitant infection or treatment interruption. The exception to this rule is the presence of concomitant steroids. In this case, report the stool frequency; we will also capture the presence of steroid use in a separate indicator variable.

13 Baseline Mayo endoscopic subscore:

{0, 1, 2, 3, NA}

We are attempting to capture this Mayo subscore within 6 months of the date of Tofacitinib initiation (e.g. **Month -6 to Month 0**). The idea here is to capture the severity of disease as well as the fact that this data was actively used to guide the decision to initiate Tofacitinib. Unlike the OCTAVE trials which required a staging colonoscopy to be performed within 1 week of randomization, we allow a 6 month lag here to account for delays in insurance approval and initiation.

Use the endoscopic impression of disease – histological data can be supportive but in the trials the primary ascertainment was by centrally-read endoscopy alone. If the endoscopist explicitly reports a Mayo endoscopic subscore than report this in the field. If he or she does not, score it using the descriptive language from the report as they map to the fields below, and annotate based on the most severely affected segment. If the colon was mostly quiet but there was an ulcer in the rectum, this would be scored as a 3.

Per the OCTAVE supplemental data protocol (Appendix 1, page 80), the modified Mayo endoscopic subscore as defined below. Underlined items and in parenthesis are my emphasis to facilitate distinguishing categories.

0: Normal or inactive disease (this includes chronic scarring, pseudopolyps)

1: Mild disease – erythema, decreased vascular pattern

2: Moderate disease – marked erythema, lack of vascularity, any friability, erosions

3: Severe disease – spontaneous bleeding, ulceration(ulcers are defined by the presence of granulation tissue)

If there exist multiple datapoints within this time window, select the latest one up to the time of treatment failure, if it occurred.

If the patient is subsequently found to have superinfection including CMV that may influence the results of the endoscopy, consider excluding these results from the data used to annotate this field.

If the patient is on any glucocorticoid at the time this assessment is done, it's okay to report whatever was found here (as opposed to mark as 'NA'); we are separately assessing corticosteroid use in a different variable.

14 Baseline c-reactive protein:

{Rational number, NA}

As with the Mayo Endoscopic subscore, use the latest available C-Reactive Protein within the 6 months preceding the date of Tofacitinib initiation. Please use the date of the lab draw to make your assessment. If it is unavailable strictly within this 6 month antecedent window, score as NA. If the lab is reported under the quantifiable limit, report '0'. If it is over the quantifiable limit, report the threshold of quantification.

Be careful that different labs use different units! All values should be reported in mg/L, not mg/dL.

15 Baseline fecal calprotectin:

{Rational number, NA}

As with the Mayo Endoscopic subscore, use the latest available Fecal Calprotectin within the 6 months preceding the date of Tofacitinib initiation. Please use the date of the lab draw to make your assessment. If it is unavailable strictly within this 6 month antecedent window, score as NA. If the lab is reported under the quantifiable limit, report '0'. If it is over the quantifiable limit, report the threshold of quantification.

16 Baseline glucocorticoid use:

{Y, N, NA}

The rationale underlying this variable is to give some indication as to if patients who were on steroids are systematically different in terms of response (or any other variable of interest) compared to those who were not.

The OCTAVE trials employed essentially stable glucocorticoid dosing during the induction phase (Table 3, page 40 of the supplemental protocol) at a max entry dose of 25mg prednisone/9mg budesonide, and mandatory tapering (Table 2, page 646) during the maintenance phase. IV and rectal corticosteroids were prohibited from use over the course of the trial.

However, since patients in routine clinical practice are regularly on these co-therapies and do not follow protocolized tapering schedules, this variable will simply capture all patients who received any systemic glucocorticoid (including Prednisone, Prednisolone, Budesonide) either intravenously or orally at the time of Tofacitinib initiation. We are not including rectal steroids here because of generally low systemic absorption unlikely to significantly confound the likelihood of declaring treatment failure or follow-up Mayo score.

Use the clinic or consult note just before and after initiation on the drug as well as communication via phone or email during this period as the source of your data. If the data does not exist in a satisfactory way, record NA.

17 Previous treatment with TNF antagonist:

{Y, N, NA}

Use the clinic notes, *Care Everywhere* GI notes if they exist, and the “magnifying glass” global search functionality to determine if the patient has previously received any of the following TNF inhibitors -- Infliximab, Adalimumab, Golimumab, Certolizumab -- specifically for the purposes of treating IBD (e.g. not for a rheumatological or dermatological disorder).

Score as Y if the patient has previously been treated with a TNF antagonist
Score as N if patient has never previously been treated with a TNF antagonist

18 History of TNF antagonist failure:

{Y, N, NA}

Use the clinic notes, *Care Everywhere* GI notes if they exist, and the “magnifying glass” global search functionality to determine if the patient has failed any of the following TNF inhibitors -- Infliximab, Adalimumab, Golimumab, Certolizumab -- specifically for the purposes of treating IBD (e.g. not for a rheumatological or dermatological disorder).

As in the OCTAVE protocol and accompanying NEJM publication, the history of treatment failure is as determined by the treating clinician. For Tofacitinib, the trial withdrawal criteria included serious infections, significant abnormal labs, adverse events, surgery, new therapy for UC, or at the patient’s request.

Score as Y if the patient has previously failed treatment with a TNF antagonist
Score as N if patient has never previously failed treatment with a TNF antagonist

19 History of oral glucocorticoid failure:

{Y, N, NA}

Use the clinic notes, *Care Everywhere* GI notes if they exist, and the “magnifying glass” global search functionality to determine if the patient has failed any oral glucocorticoid (e.g. budesonide, prednisone, prednisolone). If a patient has been hospitalized for an IBD flare and receives IV steroids (e.g. methylprednisolone) then score this as 1.

Refer to the section History of TNF antagonist failure for additional guidance.

Score as 0 if patient has never previously failed treatment with a Glucocorticoid
Score as 1 if the patient has previously failed treatment with a Glucocorticoid

There may be a questions raised in this section related to patients receiving lower than typical doses of prednisone, having a partial response to therapy, or responding but in the setting of simultaneous immunosuppressant co-therapy (so difficult to tell which was responsible for the improvement). In general would tend to err on the side of no failure rather than coding this as NA (esp if not unambiguous documented as such).

20 History of immunosuppressant failure:

{Y, N, NA}

Use the clinic notes, *Care Everywhere* GI notes if they exist, the presence of thiopurine metabolite laboratories (e.g. 6-Thioguanine), and the “magnifying glass” global search functionality to determine if the patient has failed any ‘immunosuppressant’ (e.g. Azathioprine, 6-Mercaptopurine, Methotrexate, Cyclosporine A, Tacrolimus).

Refer to the section History of TNF antagonist failure for additional guidance.

If the patient was previously on this drug in combination with another (e.g. biologic) and was deemed to have failed that regimen, score as Y.

Score as Y if the patient has previously failed treatment with an Immunosuppressant
Score as N if patient has never previously failed treatment with an Immunosuppressant

Ulcerative Colitis Effectiveness Spreadsheet

21 Date of Tofacitinib initiation:

{YYYY-MM-DD, NA}

As above, the best evidence to ascertain this is a message from a patient indicating that they’ve started the drug, or a telephone encounter documenting as such. These often give relatively precise dates of initiation.

If there is a sequence of clinical notes (before and after) that are consistent with initiation, these can be used in combination with the date of medication orders (listed within a medication order) to find a more precise date of initiation.

If dealing with older notes where supporting data (e.g. patient messages, medication orders) are unavailable, then document use the month (or less commonly, the year) and report a “rounded down” date (e.g. 2018-05-01).

Rarely, there are patients who have been on a drug for multiple continuous periods of time (e.g. initially on a clinical trial then again later, or patients who stop due to possible drug reaction and subsequently restart). In these scenarios, prioritize the longest period of sustained use. If there are multiple, equally long on-off periods, select the most recent period with well-defined start and end dates. If this is not possible because the patient starts and stops a drug erratically or does not use it daily, score this as NA and annotate the “inconsistent use” column accordingly.

22 Date of last known use of Tofacitinib (date of censorship)

{YYYY-MM-DD, NA}

This date corresponds to the last observed date that the drug was observed being used (from the standpoint of survival analysis). This could be the date of discontinuation, the last date of any available evidence (e.g. contact with our clinic or other accessible gastroenterologists), date of surgery, or date of death. If patients are still on the drug as of the date of the record review, this would be the recorded date (and they would be considered administratively censored).

As before, the best evidence to get this comes from a patient email indicating that they’ve stopped the drug, a colonoscopy report with the recommendation to stop the treatment given futility, or a clinician writing to a patient indicating to a patient that they should stop the drug.

23 Status at date of last use (censorship status)

{0, 1, 2, 3, NA}

Score as 1 if the treatment has been deemed futile. This could be for any reason – lack of effectiveness or adverse outcome/intolerance. This could be decided by either the patient or the physician (or both).

Score as 0 (administratively censored) if the 'implicit/true time of treatment failure' is longer than the time under observation. For instance, if the patient is still on a drug as of the date of a chart review, or the patient discontinues the drug due to loss of insurance coverage, their latent 'survival' time is longer than the time under observation (e.g. under different circumstances with greater observation time and guaranteed insurance, the measured time to treatment failure would be longer than what was measured here). Here, the loss of insurance coverage is being treated as MCAR (missing completely at random).

Score as a 2 if the patient was lost to follow-up defined as the absence of any data relevant to IBD status and continued treatment anywhere in the chart (including *Care Everywhere*) for 1 year. This category is a stand in for any status where the investigator must assess on a case by case basis whether the event can be justifiably be understood as resulting in data MCAR vs other types of missingness.

Score as 3 if the patient underwent a true competing event (e.g. all-cause mortality) that fundamentally precludes the observation of interest (e.g. treatment survival time).

If a patient undergoes surgery because a treatment is deemed a failure, score this as a 1. If the patient undergoes surgery due to longstanding fibrostenotic strictures and the drug is not resumed after surgery (with notes suggesting the absence of evidence that the treatment was a failure), score this as a 0.

24 Follow-up Mayo stool frequency subscore:

{0, 1, 2, 3, NA}

We are collecting the components of the Mayo score in order to allow us to closely mimic the primary endpoints of the OCTAVE trials. These endpoints are intended to measure the proportion of patients who respond to maintenance therapy at the 1 year time point. At our institution we do not protocolize the timing of clinic follow-up and different clinicians have different practices. However, many clinicians tend to formal assess response to therapy by month ~3 for Ulcerative Colitis, and occasionally month ~4-6 especially in individuals who have shown partial benefit within the first few months. However, scheduling in routine clinical contexts can be imprecise. Many patients live far away, sometimes in other states. Due to these considerations, including personal and administrative delays in scheduling, insurance approval etc., we are collecting the stool frequency subscore *within the window from month 2 through to month 8 after treatment initiation*.

We considered a longer follow-up time to more closely mimic the 1 year time point studied in OCTAVE Sustain but pilot studies suggested that endoscopic and clinical assessment at that time was less common than at earlier time points (and therefore might result in more missing data).

Per the OCTAVE supplemental data protocol (Appendix 1, page 80), this subscore is to be scored using the following ordinal scale:

- 0: Normal number of bowel movements for this patient
- 1: 1 to 2 bowel movements more than normal
- 2: 3 to 4 bowel movements more than normal
- 3: 5 or more bowel movements more than normal

Per FDA guidance a bowel movement "is defined as a trip to the toilet when the patient has either a bowel movement, or passes blood alone, blood and mucus, or mucus only." Therefore, stool frequency should incorporate toilet trips when there is passage of any liquid or solid content through the anus, not limited only to feces.

If the stool frequency is reported as a relative increase over baseline, no ascertainment of absolute baseline frequency is required. However, if stool frequency is reported as an absolute frequency, the chart reviewer must make every effort to identify a baseline stool frequency from the chart. This can be identified from anywhere within full record, including timepoints before or after Tofacitinib use. We recommend using the global search functionality to look for keywords like: "baseline" "normal" "1-2" "2-3" "stools" "bowel movements" etc.

If the notes do not explicitly mention "baseline" but do indicate a stool frequency of between 1-3 bowel movements daily in any context (e.g. other treatments such as steroids), treat this number as the baseline. The justification for this is that population-based studies have suggested that up to 3 bowel movements daily can be considered normal.

If the patient has baseline constipation during periods of quiescence and returns to use of stool softeners/laxatives, can impute this field to 0 (e.g. assume use of stool softeners to achieve 1 bowel movement daily).

If no "baseline" nor approximate baseline is mentioned, a baseline of 3 bowel movements daily can be imputed at the chart reviewers discretion. This selection would not be expected to significantly affect the study results since the analysis will involve the paired differences in Mayo scores. If this is done an additional indicator variable denoting that this was performed should be added to the data collection spreadsheet.

When dealing with numeric ranges of bowel movements, calculate the score based on the minimum increase in stool frequency compatible with the range. For instance, if a patient has 1-2 bowel movements at baseline, and is currently having 5-6 during the morning and 2-3 over the rest of the day, this would be a difference between (1 to 2) and (7 to 9), corresponding to a minimum increase of 6 bowel movements above baseline (e.g. 1 to 7 is 6 above baseline, 2 to 9 is 7 above baseline, hence take the minimum). Therefore, this is scored as a '3' based on the above rubric.

In the event that multiple such subscores exist within this window, we use the following algorithm:

1. Select the stool frequency subscore occurring closest in time to the date of treatment failure
2. If no treatment failure and a colonoscopy was performed within this period, select the stool frequency subscore occurring closest in time to the date of the colonoscopy
3. If no colonoscopy was performed in this period, select the highest available subscore within this 6 month window.

The rationale for picking the highest score within the window is that some assessments may be masked by concomitant therapy (e.g. steroids) which are somewhat harder to ascertain from the EHR.

As mentioned in the top section 'Guidelines': exclude any stool frequency confounded by the presence of a concomitant infection or treatment interruption. The exception to this rule is the presence of concomitant steroids. In this case, report the stool frequency; we will also capture the presence of steroid use in a separate indicator variable.

Score this field as NA if it is not satisfactorily identified.

25 Follow-up Mayo rectal bleeding subscore:

{0, 1, 2, 3, NA}

We are collecting the components of the Mayo score in order to allow us to closely mimic the primary endpoints of the OCTAVE trials. These endpoints are intended to measure the proportion of patients who respond to maintenance therapy at the 1 year time point. At our institution we do not protocolize the timing of clinic follow-up and different clinicians have different practices. However, many clinicians tend to formal assess response to therapy by month ~3 for Ulcerative Colitis, and occasionally month ~4-6 especially in individuals who have shown partial benefit within the first few months. However, scheduling in routine clinical contexts can be imprecise. Many patients live far away, sometimes in other states. Due to these considerations, including personal and administrative delays in scheduling, insurance approval etc., we are collecting the rectal bleeding subscore *within the window from month 2 through to month 8 after treatment initiation.*

We considered a longer follow-up time to more closely mimic the 1 year time point studied in OCTAVE Sustain but pilot studies suggested that endoscopic and clinical assessment at that time was less common than at earlier time points (and therefore might result in more missing data).

Per the OCTAVE supplemental data protocol (Appendix 1, page 80):

- 0: No blood seen
- 1: Streaks of blood with stool less than half the time
- 2: Obvious blood with stool most of the time
- 3: Blood alone passes

This subscore has been acknowledged by the FDA to be imperfect (FDA CDER Draft Guidance, Aug 2016) for multiple reasons including the fact that it is a “double-barreled” question (e.g. simultaneously assesses both amount and frequency of blood). Presumably as a measure to correct the frequency problem, the Pfizer protocol indicates that this should be scored based on “the most severe bleeding of the day.”

We have adopted the following pragmatic strategy:

- 1) If data of bleeding severity is available, then score based on this as follows:
 - o If there is clear documentation of no bleeding, score as 0.
 - o If there is clear documentation of episodes where only blood passes, score this as a 3.
 - o If there is bleeding but the amount is vague (e.g. “some”) then code this as a 1.5.
 - o Otherwise if there is a clear indication of degree of bleeding, use 1 or 2 as appropriate.
- 2) If severity data is not available but the frequency data is, then scoring using this data and the OCTAVE metric above can be performed at the discretion of the chart abstractor.

If there exist multiple datapoints within this 6-month time window, use the approach as outlined in the stool frequency section. If multiple datapoints around the same time suggest slightly different degrees of bleeding, select the most severe degree of bleeding. If the note indicates blood 50% of the time, score as 1.5.

26 **Follow-up Mayo PGA subscore:**

{0, 1, 2, 3, NA}

We are collecting the components of the Mayo score in order to allow us to closely mimic the primary endpoints of the OCTAVE trials. These endpoints are intended to measure the proportion of patients who respond to maintenance therapy at the 1 year time point. At our institution we do not protocolize the timing of clinic follow-up and different clinicians have different practices. However, many clinicians tend to formal assess response to therapy by month ~3 for Ulcerative Colitis, and occasionally month ~4-6 especially in individuals who have shown partial benefit within the first few months. However, scheduling in routine clinical contexts can be imprecise. Many patients live far away, sometimes in other states. Due to these considerations, including personal and administrative delays in scheduling, insurance approval etc., we are collecting the physician's global assessment within the window *from month 2 through to month 8 after treatment initiation*.

We considered a longer follow-up time to more closely mimic the 1 year time point studied in OCTAVE Sustain but pilot studies suggested that endoscopic and clinical assessment at that time was less common than at earlier time points (and therefore might result in more missing data).

The scoring system is as follows (with acceptable phrasing in parenthesis):

0: Normal ("in remission", "in clinical remission")

1: Mild disease

2: Moderate disease

3: Severe disease

Per the Pfizer protocol used for the OCTAVE trials, this score "acknowledges the three other criteria, the patient's recollection of abdominal discomfort and general sense of wellbeing, and other observations, such as physical findings and the patient's performance status."

This score ideally should reflect a **dynamic**, "real-time" assessment of disease activity (e.g. evidence of ongoing inflammation causing symptoms that is in principle modifiable by medication). This is in contrast to annotations of severe disease based on, for example, the history of multiple prior treatment failures.

A good place to identify the physician's global assessment (PGA) is the end of colonoscopy reports, where patients are commonly characterized as "mild" or "severe" independent of scoring using the Mayo endoscopic score. Other sources of the PGA include the Assessment and Plan of clinic notes, or appeal letters to insurance; the former is preferable because it 1) is written by the attending physician, rather than fellows/residents in clinic, and 2) appeal letters to insurance may be at risk of bias.

In the event that multiple such subscores exist within this window, use the same approach as detailed in the stool frequency section.

As mentioned in the top section 'Guidelines': exclude any stool frequency confounded by the presence of a concomitant infection or treatment interruption. The exception to this rule is the presence of concomitant steroids. In this case, report the stool frequency; we will also capture the presence of steroid use in a separate indicator variable.

27 **Follow-up Mayo endoscopic subcore**

{0, 1, 2, 3, NA}

We are collecting the components of the Mayo score in order to allow us to closely mimic the primary endpoints of the OCTAVE trials. These endpoints are intended to measure the proportion of patients who respond to maintenance therapy at the 1 year time point. At our institution we do not protocolize the timing of clinic follow-up and different clinicians have different practices. However, many clinicians tend to formal assess response to therapy by month ~3 for Ulcerative Colitis, and occasionally month ~4-6 especially in individuals who have shown partial benefit within the first few months. However, scheduling in routine clinical contexts can be imprecise. Many patients live far away, sometimes in other states. Due to these considerations, including personal and administrative delays in scheduling, insurance approval etc., we are collecting the endoscopic subscore within the window *from month 2 through to month 8 after treatment initiation*.

We considered a longer follow-up time to more closely mimic the 1 year time point studied in OCTAVE Sustain but pilot studies suggested that endoscopic and clinical assessment at that time was less common than at earlier time points (and therefore might result in more missing data).

Use the endoscopic impression of disease, and do not use any histological data to make this annotation. It is important to try to score this as the endoscopist would have scored it without introducing your own interpretation. Some endoscopists use a formal scoring system built into Endopro, whereas others use descriptive language to characterize the severity of the disease. In the event of the latter, use the words mapping to the **most** severe endoscopic category to make your annotation.

Per the Supplemental Data form, use the modified Mayo endoscopic subscore as defined below. The unique terms that distinguish one category from another are underlined.

0: Normal or inactive disease (this includes chronic scarring, pseudopolyps)

1: Mild disease -- erythema, decreased vascular pattern

2: Moderate disease -- marked erythema, lack of vascularity, any friability, erosions

3: Severe disease -- spontaneous bleeding, ulceration (ulcers are defined by the presence of granulation tissue)

If there exist multiple datapoints within this time window, select the latest one up to the time of treatment failure, if it occurred.

If the patient is subsequently found to have superinfection including CMV that may influence the results of the endoscopy, consider excluding these results from the data used to annotate this field.

If the patient is on any glucocorticoid at the time this assessment is done, it's okay to report whatever was found here (as opposed to mark as 'NA'); we are separately assessing corticosteroid use in a different variable.

28 **Follow-up c-reactive protein:**

{Rational number, NA}

As with the Mayo Endoscopic subscore, use the latest available C-Reactive Protein within the 6-12 month window following Tofacitinib initiation. If it is unavailable strictly within this window, score as NA. If the lab result is reported under the quantifiable limit, report '0'. If it is over the quantifiable limit, report the threshold of quantification.

Be careful that different labs use different units! We want mg/L, not mg/dL.

If the patient is on any glucocorticoid at the time this assessment is done, it's okay to report whatever was found here (as opposed to mark as 'NA'); we are separately assessing corticosteroid use with a different variable.

29 **Follow-up fecal calprotectin:**

{Rational number, NA}

As with the Mayo Endoscopic subscore, use the latest available Fecal Calprotectin within the 6-12 month window following Tofacitinib initiation. If it is unavailable strictly within this window, score as NA. If the lab result is reported under the quantifiable limit, report '0'. If it is over the quantifiable limit, report the threshold of quantification.

If the patient is on any glucocorticoid at the time this assessment is done, it's okay to report whatever was found here (as opposed to mark as 'NA'); we are separately assessing corticosteroid use with a different variable.

30 **Follow-up glucocorticoid use:**
{0, 1, NA}

Even though this data point does not exist within the clinical trials, it is an important measurement to capture in order to properly interpret the follow-up subscore data.

The rationale underlying this variable is to give some indication as to if patients who were on steroids are systematically different in terms of response (or any other variable of interest) compared to those who were not.

The OCTAVE trials employed essentially stable glucocorticoid dosing during the induction phase (Table 3, page 40 of the supplemental protocol) at a max entry dose of 25mg prednisone/9mg budesonide, and mandatory tapering (Table 2, page 646) during the maintenance phase. IV and rectal corticosteroids were prohibited from use over the course of the trial.

However, since patients in routine clinical practice are regularly on these co-therapies and do not follow protocolized tapering schedules, this variable will simply capture all patients who received any systemic glucocorticoid (including Prednisone, Prednisolone, Budesonide) either intravenously or orally at the time of follow-up. We are not including rectal steroids here because of generally low systemic absorption unlikely to significantly confound the likelihood of declaring treatment failure or follow-up Mayo score.

Use the clinic or consult note just before and after initiation on the drug as well as communication via phone or email during this period as the source of your data. If the data does not exist in a satisfactory way, record NA.

31 **Other co-therapy:**
{0, 1, NA}

This is intended to be an indicator variable for whether or not the patient is on any other concurrent therapy that would be expected to positively confound the treatment effect of tofacitinib (e.g. golimumab, vedolizumab, steroid enemas, 5-ASA, hyperbaric oxygen, VSL #3, curcumin).

Mark this as 1 if the patient is on any other IBD treatments that might be expected to mask the isolated effect of Tofacitinib from a treatment standpoint. Do not include treatments that treat other non-IBD conditions (e.g. PPI). If the patient is on an immunomodulator with systemic action based on prescription by another provider (e.g. rheumatologist) mark this field as '1' only if this agent has been shown in prospective clinical trials to reduce gastrointestinal inflammation, 0 otherwise.

Statistical Analysis Plan

32 **Primary endpoints:**

- 1) Time to treatment failure
- 2) Proportion of patients without treatment failure at 1 year (*real-world effectiveness rate*)

Null hypothesis: The probability of incident users of Tofacitinib without treatment failure at 1 year in the 'real-world' setting is equal to the probability of remission (total Mayo score of ≤ 2 , with no subscore >1 and a rectal bleeding subscore of 0) at week 52 as reported by the OCTAVE investigators (Sandborn et al. 2017)

Alternative hypothesis: The probability of incident users of Tofacitinib without treatment failure at 1 year in the 'real-world' setting is **not equal** to the probability of remission (total Mayo score of ≤ 2 , with no subscore >1 and a rectal bleeding subscore of 0) at week 52 as reported by the OCTAVE investigators (Sandborn et al. 2017)

Sample size considerations: Because this represents a uncontrolled study of observational data, no power calculation was performed. Preliminary data at our institution (UCSF) as of May 2019 suggested up to ~120 IBD patients who had been prescribed Tofacitinib.

Analysis of primary endpoints:

The time to treatment failure will be estimated using the Kaplan-Meier estimator; survival times corresponding to the 25th, 50th, and 75th percentiles of treatment failure will be calculated.

The analysis set will be restricted to individuals with observation time ≥ 1 year, and the counts of individuals with and without treatment failure at 1 year will be tabulated.

Counts corresponding to Sandborn et al. 2017 will be computed as follows: The total proportion of responders will be calculated

as the product of the probability of achieving *clinical response* (a decrease from baseline in Mayo score of at least 3 points and at least 30 percent, with an accompanying decrease in the rectal bleeding subscore of at least 1 point or an absolute rectal bleeding subscore of 0 or 1) at week 8 by the probability of achieving *remission* at week 52. The total number of patients assigned to the active arm of OCTAVE Sustain who achieved remission at week 52 will be divided by the aforementioned product to determine an effective sample size for the week 52 assessment of clinical remission.

These counts will be used to perform a Fisher's exact test.

Analysis of patient demographics:

Patients with greater with observation time ≥ 1 year will be tabulated by baseline demographic characteristics measured here and compared to the summary statistics reported in Table 1 of Sandborn et al. 2017. Proportions and categorical variables will be compared by Fisher's exact test, means will be compared by 1-sample t-test, and medians compared by 1-sample Wilcoxon test. No corrections for multiple testing will be performed unless otherwise specified.

Quality Control: Exploratory data analysis will be performed to confirm accuracy of data entry and identify outlying observations.

Missing data: Missing data will be handled by Multiple Imputation using Chained Equations using the Fully Conditional Specification using at least 20 chains and 10 iterations. The quality of the imputations will be assessed via assessment of Markov Chain Monte Carlo trace plots and univariate and bivariate distributions of imputed values. Results will be assessed using different imputation models.

Software: All analyses will be performed in the R programming environment with the following packages representing a non-exclusive list: 'data.table', 'survival', 'survminer', 'readxl', 'tidyverse', 'scales', 'binom', 'ggplot2', 'lubridate', 'RMarkdown', 'MICE', 'caret'

Sensitivity analysis:

- 1) The hypothesis test comparing the real-world effectiveness rate with that of the OCTAVE Sustain trial will be repeated using less stringent W52 endpoints to account for the possibility that the real-world determination of treatment failure may be less stringent than that of the OCTAVE primary endpoint.
- 2) The data will be analyzed in a Bayesian framework with posterior probability distributions calculated using the following prior Beta distributions: Jeffrey's Prior, prior distribution corresponding to the patient counts from the OCTAVE Sustain trial, and a beta distribution with median equal to 1/3 and 0.05 quantile equal to 0.1.

Exploratory analyses/endpoints: An exploratory analysis will include the following:

- 1) Proportion of real-world patients who would have qualified for the randomized clinical trial based on all Inclusion/Exclusion criteria
- 2) Proportion of real-world patients who would have qualified for the randomized clinical trial based on baseline total Mayo Score
- 3) Proportion of real-world patients meeting the primary endpoint of clinical remission and selected secondary endpoints as defined and published by the OCTAVE investigators
- 4) Other machine learning algorithms (Random Forests, Support Vector Machines, Quadratic Discriminant Analysis, Ensemble Models, penalized logistic regression) will be explored to determine their accuracy in predicting treatment response using baseline and early follow-up covariates



This is an open access protocol distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

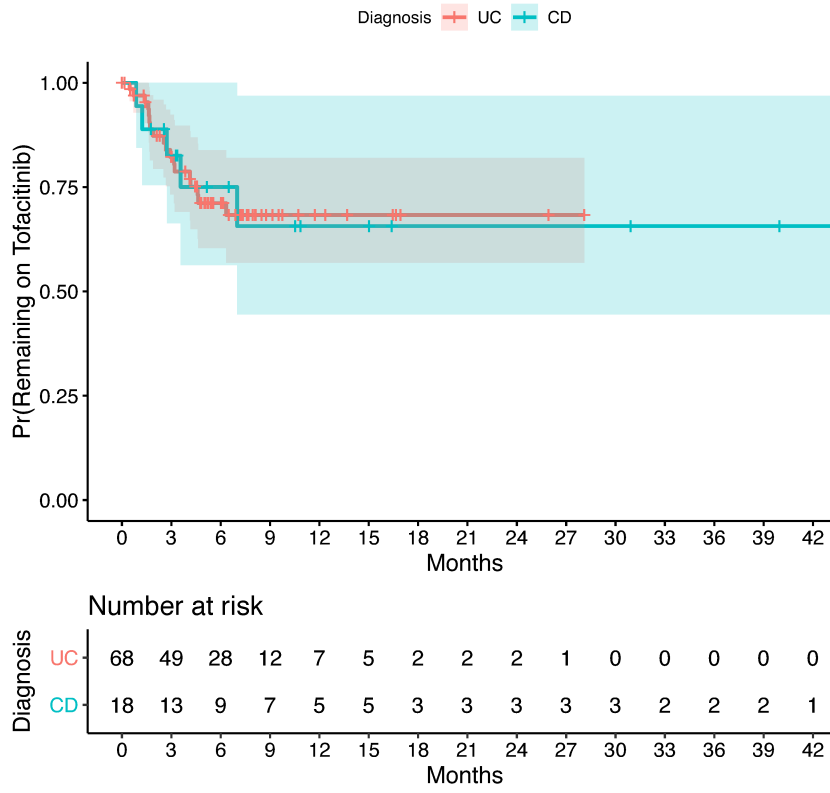
eTable 1. Examples of Data Abstraction from the EHR. The following table illustrates the covariate abstraction process using representative, synthetic, free-text data elements from the EHR. Relevant free-text elements have been colored and underlined for emphasis. Of note, data elements were only abstracted if they were internally consistent across multiple locations within the EHR; otherwise, they were annotated as NA. For example, a mention of a medication start by a patient had to be consistent with accompanying medication prescription orders and follow-up notes consistent with the purported start date. This full process is detailed in the record review protocol that is reproduced elsewhere in this supplemental file.

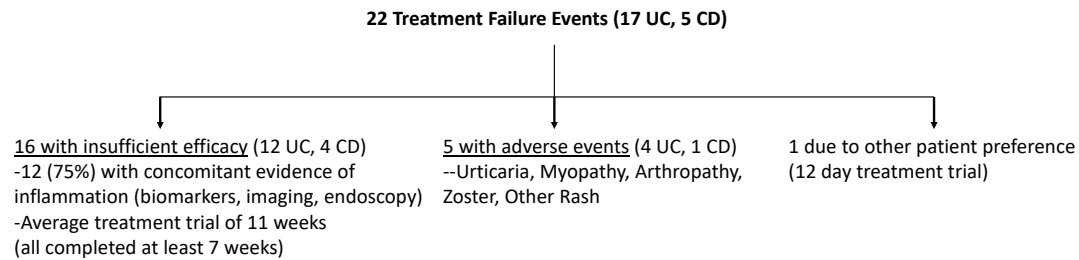
EHR Data Element:	Covariates Extracted:
(January 9 th 2020, MyChart Message) "Hi Dr. Rudrapatna, just wanted to share with you some great news. I <u>just started the new medication on Monday</u> and I already feel SOO much better!"	Start Date: <u>January 5th 2020</u>
(December 8 th 2018, Colonoscopy Report) "Impression: <u>Severe (Mayo 3) UC despite 8 weeks of Tofa 10 BID.</u> Recommendations: 1) <u>Stop Tofa</u> , 2) Surgical Consultation"	Stop Date: <u>December 8th 2018</u> Diagnosis: <u>Ulcerative Colitis</u> Treatment Failure: <u>Yes</u> Mayo endoscopic subscore: <u>3</u>
(April 13 th , 2019, Clinic Note, Assessment/Plan) "A/P: 52yo M with <u>moderate-severe panulcerative colitis refractory to Infliximab</u> and Vedolizumab"	Diagnosis: <u>Ulcerative Colitis</u> Mayo Physician Global Assessment: <u>2.5</u> History of TNFi failure: <u>Yes</u>

eTable 2. Most common reasons disqualifying the real-world UC cohort from meeting the OCTAVE trial eligibility criteria

Recent/Current use of other prohibited immunosuppressives: --Vedolizumab use within the past 1 year --Prednisone at doses exceeding 25mg/d (or equivalent) --Anti-Tumor Necrosis Factor-Alpha use within the last 8 weeks --Immunomodulator use within the past 2 weeks --Any corticosteroid or 5-aminosalicylate per rectum within past 2 weeks --Intravenous corticosteroid use within the past 2 weeks
Likely to require surgery during treatment period
Fluctuating dose of allowable concomitant medications (e.g. oral corticosteroid) during induction period
History of prior surgery potentially affecting drug absorption

eFigure 1. Durability of tofacitinib treatment



eFigure 2. Reasons for treatment discontinuation

Supplemental References:

1. R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
2. Rinker, T. W. & Kurkiewicz, D. (2017). pacman: Package Management for R. version 0.5.0. Buffalo, New York. <http://github.com/trinker/pacman>
3. Matt Dowle and Arun Srinivasan (2019). data.table: Extension of `data.frame`. R package version 1.12.2. <https://CRAN.R-project.org/package=data.table>
4. Alboukadel Kassambara and Marcin Kosinski (2019). survminer: Drawing Survival Curves using 'ggplot2'. R package version 0.4.4. <https://CRAN.R-project.org/package=survminer>
5. Hadley Wickham and Jennifer Bryan (2019). readxl: Read Excel Files. R package version 1.3.1. <https://CRAN.R-project.org/package=readxl>
6. Hadley Wickham and Lionel Henry (2019). tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions. R package version 0.8.3. <https://CRAN.R-project.org/package=tidyr>
7. Hadley Wickham (2018). scales: Scale Functions for Visualization. R package version 1.0.0. <https://CRAN.R-project.org/package=scales>
8. Sundar Dorai-Raj (2014). binom: Binomial Confidence Intervals For Several Parameterizations. R package version 1.1-1. <https://CRAN.R-project.org/package=binom>
9. H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
10. Garrett Golemund, Hadley Wickham (2011). Dates and Times Made Easy with lubridate. Journal of Statistical Software, 40(3), 1-25. URL <http://www.jstatsoft.org/v40/i03/>
11. A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18--22.
12. Stef van Buuren, Karin Groothuis-Oudshoorn (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1-67. URL <https://www.jstatsoft.org/v45/i03/>
13. Frank E Harrell Jr, with contributions from Charles Dupont and many others. (2019). Hmisc: Harrell Miscellaneous. R package version 4.2-0. <https://CRAN.R-project.org/package=Hmisc>
14. Stefan Milton Bache and Hadley Wickham (2014). magrittr: A Forward-Pipe Operator for R. R package version 1.5. <https://CRAN.R-project.org/package=magrittr>