Protocol A3921165

EFFICACY, SAFETY, TOLERABILITY AND PHARMACOKINETICS OF TOFACITINIB FOR TREATMENT OF SYSTEMIC JUVENILE IDIOPATHIC ARTHRITIS (sJIA) WITH ACTIVE SYSTEMIC FEATURES IN CHILDREN AND ADOLESCENT SUBJECTS

Statistical Analysis Plan
(SAP)

Version: 5

Date: 06 February 2024

TABLE OF CONTENTS

LIST OF TABLES

APPENDICES

1. VERSION HISTORY

Table 1. Summary of Major Changes in SAP Amendments

SAP Version (Date)	Associated Protocol Amendment	Change	Rationale
5 (06 Feb 2024)	Protocol Amendment 8 (01 Sep 2023)	For selected secondary continuous endpoints, the main estimand was updated to treatment policy strategy in Section 2.3.2.2 and Section 6.2.1.2; the supplementary estimand was updated to hypothetical strategy in Section 2.3.4 and Section 6.2.1.3. Appendix 1 was updated accordingly.	In response to FDA feedback on the revised statistical analysis plan
		Clarified in Section 5.1 that the number of flares required for the 2 planned analyses is based on the definition of the primary estimand.	For clarification
		Minor edits and corrections throughout the document	For consistency
4 (12 Jan 2024)	Protocol Amendment 8 (01 Sep 2023)	Updated study design in Section 2.2	Per Protocol Amendment 8
		Clarifications added in Section 2.3: 1) how to handle subjects who are still active but have to discontinue due to study end; 2) subjects who achieve clinical remission will be considered completers in the study	To clarify the definitions of estimands
		Removed the endpoint of clinical remission in OL phase in Sections 3.2.2 and 6.2.2	Clinical remission will only be evaluated in DB phase
		Added details of group sequential design in Section 5.1	Per Protocol Amendment 8

Table 1. Summary of Major Changes in SAP Amendments

SAP Version (Date)	Associated Protocol Amendment	Change	Rationale
		and updated Sections 5.2.1, 7 and 7.2	
		Clarifications added in Section 5.3.1.1 for censoring time, and in Section 5.3.2 for missing data handling in binary endpoint analyses	To clarify the analysis methods
		Revisions in Section 6.4 for subgroup analyses	To clarify which subgroups will be evaluated
		Revisions in Section 6.7 for listings and summaries for COVID-19 impact	To modify the analyses needed to depict the impact
		Added Appendix 2 of efficacy and futility boundaries	To provide examples of the boundaries in different scenarios
		Updated endpoint derivation details in Appendix 3.3 and Appendix 3.6	Per Protocol Amendment 8
		Minor edits and correction throughout the document	Correction and clarification
3 (01 Sep 2022)	Amendment 7 (22 Aug 2022)	Updated study design in Section 2.2	Per Protocol Amendment 7
		Updated throughout the document to clarify and differentiate “discontinuation from investigational product” and “withdrawal from study”	In response to FDA feedback on the Statistical Analysis Plan (Version 2, 24 Feb 2022)
		Updated log-rank test from stratified to unstratified and removed age group from Cox proportional hazards model in	The small expected sample size for some strata based on the current enrollment

Table 1. Summary of Major Changes in SAP Amendments

SAP Version (Date)	Associated Protocol Amendment	Change	Rationale
		the primary analysis (Sections 5.1, 5.2.1 and Appendix 1)	and small expected number of events can cause a significant impact of the efficiency or power of stratified log-rank test.
		Updated analysis method for binary endpoints from CMH to normal approximation (Section 5.2.2 and Appendix 1)	For consistency with updated primary analysis method
		Removed age group from MMRM (Section 5.2.3 and Appendix 1)	For consistency with updated primary analysis method
		Extended the on-study censoring time and risk period to the upper limit of the DB Week 52 visit window post randomization through both studies A3921165 and A3921145 (Section 5.3.1.1 and Appendix 1)	To align with updated study design and to ensure interpretability of results under the treatment policy estimand
		Updated Supplementary Analysis 2 under treatment policy estimand to include data up to 52 weeks post randomization (Section 6.1.2.1 and Appendix 1)	To align with updated study design and to ensure interpretability of results under the treatment policy estimand
		Updated the estimation and sampling method for tipping point analysis under composite estimand from Bayesian to MLE (Section 6.1.2.2 and Appendix 5)	Based on simulation results in response to FDA feedback on the Statistical Analysis Plan (Version 2, 24 Feb 2022)

Table 1. Summary of Major Changes in SAP Amendments

SAP Version (Date)	Associated Protocol Amendment	Change	Rationale
		Added a tipping point analysis for the primary endpoint under treatment policy estimand (Section 6.1.2.2, Appendix 1 and Appendix 5)	In response to FDA feedback in the Statistical Analysis Plan (Version 2, 24 Feb 2022)
		Removed CRP from the definition of JIA ACR Inactive Disease (Appendix 3.7)	Per Protocol Amendment 7
		Clarifications added for analysis windows (Appendix 3.13)	For clarification
		Removed baseline covariates from Weibull regression model in tipping point analyses (Appendix 5)	For consistency with updated primary analysis method
		Typos corrected, minor clarifications added, and references updated throughout the document.	Correction and clarification
2 (24 Feb 2022)	Amendment 2 (16 May 2019)	Remove tofacitinib 10 mg BID dose group	Per Protocol Amendment 2
	Amendment 3 (04 May 2020)	Add in COVID-19 impact analysis in Section 5.3 and Section 6.7	Per Protocol Amendment 3 to assess COVID-19 pandemic impact
	Amendment 5 (21 Jul 2021)	Add in Estimand Framework in Section 2.3	To align with ICH-E9 (R1)
		Add in Tipping Point Analysis for the primary endpoint in Section 6.1.2.2 and Appendix 5	Added as an additional sensitivity analysis to assess impact of missing data on the primary

Table 1. Summary of Major Changes in SAP Amendments

SAP Version (Date)	Associated Protocol Amendment	Change	Rationale
			endpoint of time to disease flare
		Condense Section 6.2 by similar type of endpoints/analyses	To reduce redundancy
1 (30 Oct 2017)	Original (15 Mar 2017)	Not Applicable	Not Applicable

2. INTRODUCTION

This SAP provides the detailed methodology for summary and statistical analyses of the data collected in study A3921165. This document may modify the plans outlined in the protocol; however, any major modifications of the primary endpoint definition or its analysis will also be reflected in a protocol amendment.

The safety and effectiveness of tofacitinib for the treatment of rheumatoid arthritis (RA) has been demonstrated in adult subjects. The Sponsor is conducting a pediatric investigational program to determine the safety and efficacy of tofacitinib in subjects 2 to <18 years of age for the treatment of Juvenile Idiopathic Arthritis (JIA). FDA approved tofacitinib for the treatment of children and adolescents 2 years and older with active pcJIA in September 2020. Two formulations were approved, a tablet and an oral solution, and are dosed based upon patients' weight.

As part of this pediatric investigational program, study A3921165 will evaluate efficacy, safety, tolerability and pharmacokinetics of tofacitinib as treatment for systemic (s)JIA. In this study, after 12 to 40 weeks of treatment with open-label tofacitinib, sJIA patients who are able to taper corticosteroids (CS) while maintaining an Adapted JIA ACR30 response will be identified as "responders". These responders will proceed to a double-blind withdrawal phase in which they will be randomized to either continue with tofacitinib treatment or start placebo treatment. Sustained efficacy of tofacitinib to prevent disease flare will be evaluated in the double-blind randomized withdrawal phase.

2.1. Study Objectives

Primary:

- To assess the sustained efficacy of tofacitinib versus placebo in sJIA patients, as measured by time to sJIA flare in the double-blind randomized withdrawal phase.

Secondary:

- To assess efficacy of tofacitinib versus placebo in sJIA patients at various time points in the double-blind randomized withdrawal phase, as measured by the following.
  a. Percentage of subjects with sJIA disease flares
  b. Percentage of subjects with Adapted JIA ACR 30/50/70/90/100 responses
  c. Changes from baseline in Juvenile Arthritis Disease Activity Score (JADAS-27)
  d. Percentage of subjects achieving inactive disease and clinical remission (JIA ACR)
  e. Percentage of subjects with inactive disease and minimal disease activity (JADAS-27)
  f. Other evaluations specified under "Efficacy endpoints" for the double-blind phase.
- To assess the efficacy of tofacitinib in sJIA subjects in the open-label treatment phase, as measured by:
  a. Percentage of subjects with successful corticosteroids tapering per protocol at the end of the open-label phase in subjects with sJIA receiving corticosteroids at start of open-label phase
  b. Percentage of subjects with Adapted JIA ACR 30/50/70/90/100 responses at every visit from Day 7 onward
  c. Other evaluations specified under "Efficacy endpoints" for the open-label phase.
- To assess the safety and tolerability of tofacitinib in sJIA patients.
- To assess the pharmacokinetics of tofacitinib in sJIA subjects in the open-label phase.

Exploratory:

- To evaluate exploratory biomarker and genomic samples to characterize the effect of tofacitinib.

2.2. Study Design

This is a 2-phase randomized withdrawal study to evaluate efficacy, safety and tolerability, and pharmacokinetics of tofacitinib as a treatment for sJIA. The study will enroll approximately 100 subjects from 2 to <18 years of age with sJIA with active systemic features.

The study will target enrollment of at least 12 subjects in the following age groups: from 12 to <18 years, from 6 <12 years, and from 2 to <6 years.

The study will conclude when the requisite number of flares are achieved during the double-blind phase of the study as described in Section 5.1 Hypotheses and Decision Rules.

Subjects will be enrolled into an open-label phase during which they will receive tofacitinib 5 mg twice daily (BID) oral tablets, or an equivalent weight-based lower dose of tofacitinib oral solution (1 mg/mL) BID for subjects <40 kg. The 5 mg BID dose level of tofacitinib will be evaluated in a staggered fashion (cohorts of 7 subjects followed by safety review).

The open-label phase will be divided into two parts. In Part 1, all subjects must achieve and maintain a protocol-defined minimum level of clinical response for at least 4 weeks. In Part 2 of the open-label phase, subjects treated with background CSs >0.2 mg/kg/day oral prednisone (or equivalent) will attempt to have their CS dose tapered to a predetermined target range, or lower; subjects who successfully taper their CS dose while maintaining the defined clinical response, will be eligible for the double-blind withdrawal phase of the study. The minimum total duration of treatment with a stable dose of tofacitinib for subjects completing Parts 1 and 2 must be 12 weeks to qualify to enter the randomized withdrawal phase of the study.

In the double-blind withdrawal phase, "responders" from Part 1 and Part 2 of the open-label phase will be randomized in a 1:1 ratio, stratified by age group (from 12 to <18 years, from 6 <12 years, and from 2 to <6 years), to either continue tofacitinib or withdraw from tofacitinib and start placebo. The primary endpoint of the study is time to sJIA flare in the double-blind phase. The double-blind withdrawal phase will continue until a sufficient number of flares are reported as described in Section 5.1.

The Sponsor will use external Centralized Coordinating Centers to review and confirm applicable real-time assessments of efficacy, including the Adapted JIA ACR 30/50 response, sJIA flare, and inactive disease status (JIA-ACR). The Coordinating Centers will also monitor and offer recommendations on corticosteroid tapering of study participants. The Centralized Coordinating Centers will be blinded to study drug assignment in the double-blind withdrawal phase of the study.

All subjects participating in this study, including those who discontinued in the open-label or double-blind phase, will have the option, if eligible (based on inclusion and exclusion criteria), of enrolling in the tofacitinib JIA long-term extension study (A3921145) after completion of this study.

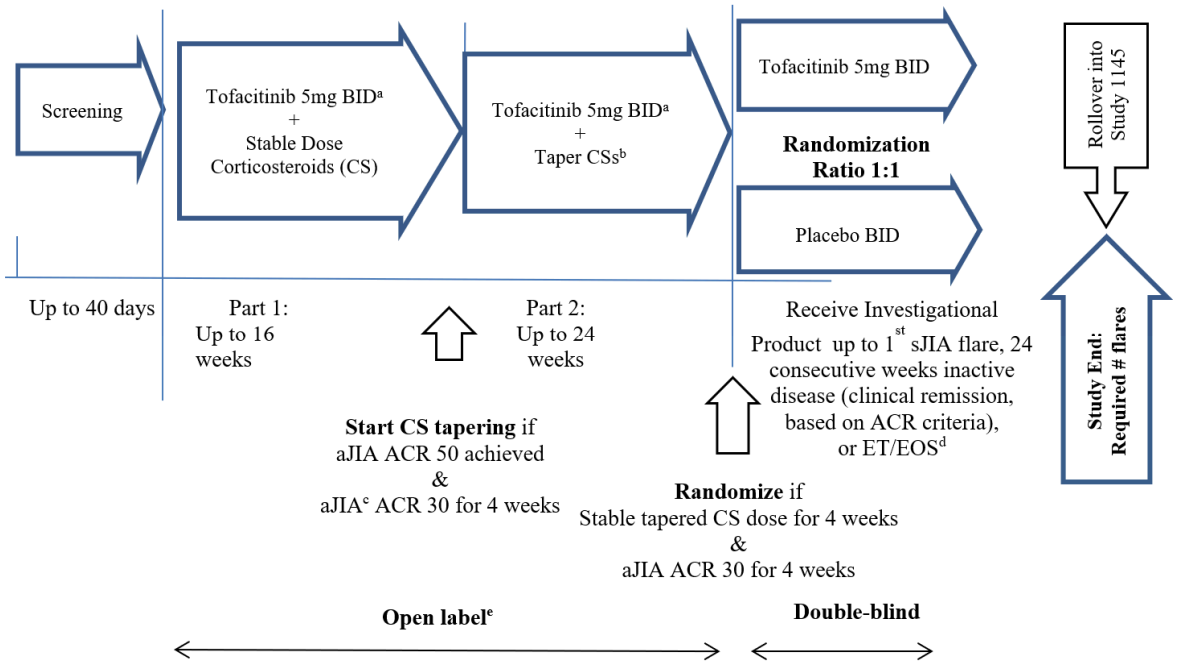
Subjects who flare in the double-blind withdrawal phase will be given the opportunity to participate in A3921145 and be treated with open-label tofacitinib. Subjects who achieve ACR clinical remission (ie, maintain JIA ACR inactive disease during 24 consecutive weeks in the double-blind withdrawal phase) will complete their participation in A3921165 and be given the opportunity to enter A3921145.

Subjects who discontinue investigational product in the double-blind phase and who do not enter A3921145 will continue in A3921165 for follow up of efficacy and safety endpoints.

Subjects will be required to perform all scheduled visits until Week 52 after randomization or until the study concludes, whichever comes first. Subjects should receive standard-of-care treatment in accordance with local treatment guidelines.

Subjects who discontinue the study in the Open-Label Phase and do not enter A3921145, will be required to perform a follow-up visit 28 days after the last dose of investigational product.

A schematic of the study design is shown below:



a Subjects <40 kg will receive an equivalent weight-based lower dose of tofacitinib 5 mg BID.
b CS Tapering is only required for subjects treated with CS >0.2 mg/kg/day oral prednisone (or equivalent). During the active CS tapering period in Part 2 subjects must maintain an Adapted JIA ACR50 response.
c aJIA: Adapted JIA.
d Subjects who discontinue investigational product in the randomized withdrawal phase continue in study until Week 52 after randomization, or until the study concludes, whichever comes first.
e Subjects who discontinue the study in the open-label phase and do not enter A3921145 within 4 weeks, will be required to perform a follow-up visit 28 days after the last dose of investigational product.

2.3. Estimands

2.3.1. Primary Estimand

The primary estimand (Estimand 1) of this study is a composite estimand (accounting for both treatment adherence and response), defined according to the primary objective and in alignment with the primary endpoint. It includes the following 5 attributes:

- Population: sJIA patients, as defined by the inclusion criteria for the double-blind phase, who are randomized.
- Treatment: tofacitinib 5 mg BID, placebo

- Variable: time to sJIA disease flare in the double-blind phase.

  Patients who complete study participation after maintaining JIA ACR inactive disease for at least 24 weeks (ie, clinical remission) in the double-blind phase, and patients who are still active in the double-blind withdrawal phase of the study but have to discontinue due to study end, will be considered as non-flare and censored at the last available flare assessment while on investigational product. (More details of on-treatment censoring time and risk period are provided in Section 5.3.1.1.)

  Patients who discontinue from investigational product for any other reasons are considered as having a disease flare on the investigational product discontinuation day.

- Intercurrent event: investigational product discontinuation is the intercurrent event. The intercurrent event is captured through the variable definition. Data collected after discontinuation of the investigational product (ie, off-treatment data) will not be included to derive the endpoint, ie, only on-treatment data will be used. (See Appendix 3.13 for definition of on-treatment data.)

- Population-level summary: hazard ratio of disease flare for those assigned to tofacitinib at randomization versus those assigned to placebo at randomization in the double-blind phase.

2.3.2. Secondary Estimand(s)

As this is a time-to-event trial, patients will have varying follow-up times. For efficacy endpoints analyzed by visit, only the data through 52 weeks in the double-blind phase will be analyzed statistically. The data after Week 52 will not be analyzed statistically but summarized descriptively. The reason for choosing 52 weeks is that majority of the patients are expected to complete 52 weeks of the study in the double-blind phase and 52 weeks are deemed adequate to evaluate secondary efficacy by visit.

The secondary estimands are only defined for the double-blind phase secondary endpoints.

2.3.2.1. Binary Secondary Endpoints in Double-Blind Phase

For binary endpoints such as adapted JIA ACR 30/50/70/90/100 response at each double-blind phase visit, the estimand also uses composite strategy (Estimand 2) and includes the following 5 attributes:

- Population: sJIA patients, as defined by the inclusion criteria for the double-blind phase, who are randomized.

- Treatment: tofacitinib 5 mg BID, placebo

- Variable: binary endpoints such as aJIA ACR 30/50/70/90/100 response at each double-blind phase visit through Week 52.

  Patients who complete study participation after maintaining JIA ACR inactive disease for at least 24 weeks (clinical remission) in the double-blind phase, and patients who are still active in the double-blind withdrawal phase of the study but have to discontinue before Week 52 due to study end, will use the last on-treatment windowed visit (Appendix 3.13) ACR value (ie, last observation carried forward (LOCF)), from that visit onward until Week 52 visit.

  Patients who discontinue from the investigational product for any other reasons before Week 52 are considered as treatment failure, ie, ACR non-responders, from that visit onward until Week 52 visit. If there is an on-treatment ACR assessment that falls within the same windowed visit with the investigational product discontinuation, the ACR assessment will prevail.

- Intercurrent event: investigational product discontinuation will be the intercurrent event. The intercurrent event is captured through the variable definition. Data collected after discontinuation of the investigational product (ie, off-treatment data) will not be included to derive the endpoint, ie, only on-treatment data (Appendix 3.13) collected prior to or at Week 52 visit will be used.

- Population-level summary: difference in proportion of subjects achieving binary endpoint response such as aJIA ACR 30/50/70/90/100 response at each double-blind phase visit through Week 52 between those assigned to tofacitinib at randomization versus those assigned to placebo at randomization.

Estimands for all other binary secondary endpoints except occurrence of disease flares in double-blind phase will follow the same composite strategy as the above secondary estimand. See endpoint definitions in Section 3.2.1 and Appendix 1 and missing data handling in Section 5.3 for specific differences.

For occurrence of disease flare at each double-blind phase visit through Week 52, the estimands follow the estimands for the primary endpoint of time to sJIA disease flare in the double-blind phase as in Section 2.3.1 and Section 2.3.3.

2.3.2.2. Continuous Secondary Endpoints in Double-Blind Phase

For change from open-label baseline in each sJIA ACR core variable, except for disability index from the CHAQ for which change from double-blind baseline will be considered (i.e., change from OL baseline in number of joints with active arthritis, number of joints with limited range of motion, physician global evaluation of disease activity, parent/legal guardian/subject evaluation of overall well-being [from the CHAQ] and ESR; and change from DB baseline in disability index from the CHAQ) at each scheduled visit through Week

52 in the double-blind phase, the estimand uses treatment policy strategy (Estimand 3) and includes the 5 attributes below:

- Population: sJIA patients, as defined by the inclusion criteria for the double-blind phase, who are randomized.
- Treatment: tofacitinib 5 mg BID, placebo
- Variable: change from open-label baseline in each sJIA ACR core variable, except for disability index from the CHAQ for which change from double-blind baseline will be considered, at each scheduled visit through Week 52 in the double-blind phase.
- Intercurrent event: the intercurrent event of investigational product discontinuation will not be considered. All the data collected while On- and Off-investigational product until Week 52 will be used to derive the endpoints.
- Population-level summary: the mean difference in change from open-label baseline in each sJIA ACR core variable, except for disability index from the CHAQ for which change from double-blind baseline will be considered, at each scheduled visit through Week 52 in the double-blind phase between all those assigned to tofacitinib at randomization and all those assigned to placebo at randomization.

Estimands for all the other secondary continuous endpoints in double-blind phase (Section 3.2.1) will follow the hypothetical estimand defined in Section 2.3.4.

2.3.3. Additional Estimands for Primary Endpoint

A supplementary (Estimand 4) estimand for the primary endpoint will use the hypothetical strategy, which estimates the treatment effect if all patients adhere to the protocol and the intercurrent event of investigational product discontinuation has not occurred. It differs from the primary estimand in the following 2 attributes:

- Variable: time to sJIA disease flare in the double-blind phase.

  Patients who discontinue from investigational product for reasons other than flare in the double-blind phase and who complete study participation after achieving clinical remission, will be considered as non-flare and censored at the last available flare assessment while on investigational product. (More details of on-treatment censoring time and risk period are provided in Section 5.3.1.1.)

- Intercurrent event: all data after an intercurrent event (investigational product discontinuation), if collected, will be excluded.

A second supplementary estimand (Estimand 5) for the primary endpoint will use the treatment policy strategy and estimate the treatment difference regardless of whether an intercurrent event occurs. It differs from the primary estimand in the following 2 attributes:

- Variable: time to sJIA disease flare in the double-blind phase.

- Intercurrent event: investigational product discontinuation is not considered for analysis. All the flare data collected both on-treatment and off-treatment in the double-blind phase of Study A3921165 will be used. For subjects who withdraw from study early, before reaching 52 weeks of double-blind treatment from Study A3921165 and enrolled in the LTE (Study A3921145), the flare data collected through Week 52 post randomization from both studies A3921165 and A3921145 will be included (see Section 6.1.2.1). (More details of on-study censoring time and risk period are provided in Section 5.3.1.1.)

2.3.4. Additional Estimands for Selected Secondary Endpoints

For aJIA ACR 30/50/70/90/100 response at each double-blind phase visit, a supplementary estimand (Estimand 6) will use the treatment policy strategy. It estimates the effect regardless of treatment adherence. It differs from the composite estimand (Section 2.3.2.1) in the following 2 attributes:

- Variable: aJIA ACR 30/50/70/90/100 response at each double-blind phase visit through Week 52.

- Intercurrent event: the intercurrent event of investigational product discontinuation will not be considered. More specifically, unlike Estimand 2, LOCF is not applicable and subjects who discontinue from investigational product will not be considered as ACR non-responders. All the aJIA ACR data collected while On- and Off-investigational product until Week 52 will be used to derive the endpoints.

For all secondary continuous endpoints in double-blind phase (Section 3.2.1), the estimand will use the hypothetical strategy (Estimand 7). It differs from the treatment policy estimand (Section 2.3.2.2) in the following 2 attributes:

- Variable: secondary continuous endpoints such as change from double-blind baseline in JADAS-27 CRP and JADAS-27 ESR at each scheduled visit through Week 52 in the double-blind phase.

- Intercurrent event: investigational product discontinuation will be the intercurrent event. All data after the intercurrent event, if collected, will be excluded; ie, only the on-treatment data (Appendix 3.13) will be used.

3. ENDPOINTS AND BASELINE VARIABLES: DEFINITIONS AND CONVENTIONS

3.1. Primary Endpoint

- Time to sJIA disease flare in the double-blind randomized withdrawal phase.

Time to the disease flare event is defined as the number of days from randomization to flare in the double-blind phase and calculated as date of disease flare – date of randomization +1.

Details for the derivation of disease flare are described in Appendix 3.3. Additional disease flares related to intercurrent events occurrence are differently defined for different estimands

in Sections 2.3.1 and 2.3.3. Different imputation methods of disease flare dates for the censored data are described separately in Sections 6.1.1, 6.1.2.1 and 6.1.2.2.

3.2. Secondary Efficacy Endpoints

3.2.1. Double-blind Randomized Withdrawal Phase

- Occurrence of disease flares in the double-blind phase at each visit.
- Adapted JIA ACR 30/50/70/90/100 response at every visit in the double-blind phase, with individual component improvements calculated relative to the open-label baseline.
- "Absence of fever", defined as absence of fever (fever defined as oral temperature >38°C/100.4°F) due to sJIA in the week preceding the assessment at every visit in the double-blind phase.
- Change from DB baseline in Juvenile Arthritis Disease Activity Score (JADAS-27, calculated based on the ESR and CRP inflammation biomarkers separately) at every visit in the double-blind phase.
- Change from baseline (both open-label and double-blind baseline) in each sJIA ACR core variable (i.e., number of joints with active arthritis, number of joints with limited range of motion, physician global evaluation of disease activity, parent/legal guardian/subject evaluation of overall well-being (from the CHAQ), disability index from the CHAQ, and ESR) at every visit in the double-blind phase.
- Occurrence of inactive disease status and minimal disease activity (JADAS-27, calculated based on the ESR and CRP inflammation biomarkers separately) at every visit in the double-blind phase. Refer to Appendix 3.6 for detailed definitions.
- Occurrence of JIA ACR inactive disease status at every visit in the double-blind phase. Refer to Appendix 3.7 for detailed definitions.
- Clinical remission (24 weeks consecutive JIA ACR inactive disease, starting from double-blind phase day 1) in the double-blind phase.
- Change from DB baseline in Child Health Questionnaire (CHQ) responses every 6 months thereafter in the double-blind phase.
- Change from DB baseline in Child Health Assessment Questionnaire (CHAQ) - Discomfort Index at every visit in the double-blind phase.

3.2.2. Open-Label Treatment Phase

- Achievement of corticosteroid tapering per protocol at the end of the open-label active treatment period in applicable subjects receiving corticosteroids on study Day 1 of the open-label phase.

  A successfully tapered subject is one that has completed Part 2 of the OL by reaching their target corticosteroid dose and maintaining an adapted JIA ACR30 response for four weeks on this dose.

- Achievement of a corticosteroid dose of ≤0.2 mg/kg/day or 10 mg/day (whichever is lower) at the end of the open-label treatment period in subjects receiving corticosteroids on Day 1 of the open-label phase.

- Occurrence of CRP ≤ 10 mg/L at every visit of the open-label phase.

- Occurrence of fever (Temp >38 Degrees Celsius) attributed to sJIA at Day 3, Day 7 and Day 14 of the open-label phase.

- Time from first tofacitinib dose to first adapted JIA ACR30 response in Part 1 of the open-label phase.

- Adapted JIA ACR 30/50/70/90/100 response at every visit from Day 7 onward in the open-label phase.

- "Absence of fever", defined as absence of fever due to sJIA in the week preceding the assessment at every visit from Day 7 onward in the open-label phase.

- Change from OL baseline in Juvenile Arthritis Disease Activity Score (JADAS-27, calculated based the ESR and CRP inflammation biomarkers separately) at every visit from Day 7 onward in the open-label phase.

- Change from OL baseline in each JIA ACR core variable at every visit from Day 7 onward in the open-label phase.

- Occurrence of inactive disease status and minimal disease activity (JADAS-27, calculated based the ESR and CRP inflammation biomarkers separately) at every visit from Day 7 onward in the open-label phase.

- Occurrence of JIA ACR inactive disease status at every visit from Day 7 onward in the open-label phase.

- Change from OL baseline in Child Health Questionnaire (CHQ) responses at the end of Part 1 and Part 2 of the open-label phase.

- Change from OL baseline in Child Health Assessment Questionnaire (CHAQ) - Discomfort Index at every visit from Day 7 onward in the open-label phase.

For details of the derivation of the secondary endpoints, see Appendix 3.1 to Appendix 3.9.

3.3. Other Endpoints

3.3.1. PK Endpoints

- Tofacitinib concentrations during the open-label phase.

3.3.2. Exploratory Endpoints

- Change from OL baseline in various genomic and serum biomarkers following treatment with tofacitinib.

3.4. Baseline Variables

This study will have a separate type of baseline for each of its two phases. The open-label baseline will be defined as the last value collected prior to day 1 of tofacitinib administration in this phase. For the analyses in the double-blind phase, baseline will refer to the values collected at the randomization visit – before the administration of the assigned treatment in this phase for all endpoints, unless noted otherwise. For Adapted JIA ACR responses the open-label baseline will be utilized as the reference point.

Selected analyses in the double-blind phase, such as changes in laboratory parameters and vital signs, will also refer to the open-label baseline as a separate analysis, in addition to using the double-blind baseline.

In the double-blind phase change from baseline in each sJIA ACR core variable will be calculated with respect to both the open-label and double-blind baseline.

Age groups for stratification will be based on the age at enrollment in the open-label phase of the study. The age groups are: 2 to <6 years, 6 to <12 years, and 12 to <18 years.

3.5. Safety Endpoints

The safety endpoints of this study are:

- All adverse events (AEs), including Serious Adverse Events (SAEs).
- Adjudicated macrophage activation syndrome (MAS) events.
- Serious infections, including tuberculosis, varicella and herpes zoster and adjudicated opportunistic infections.
- Clinically significant abnormal laboratory parameters, including abnormal hematology parameters, lipid parameter changes, liver enzymes, serum creatinine elevation.
- Adjudicated malignancies, including lymphoma and non-melanoma skin cancer.
- Adjudicated gastrointestinal perforations.

- Adjudicated cardiovascular diseases.
- Assessments of growth and puberty development.

4. ANALYSIS SETS

4.1. Double-Blind Analysis Sets

4.1.1. Double-Blind Full Analysis Set

The Double-Blind Full Analysis Set (DBFAS) will consist of all randomized subjects who received at least one dose of investigational product in the double-blind phase. Subjects will be reported under the treatment they were randomized to. In this study, the primary efficacy analysis will be performed on the DBFAS. Secondary endpoints for the double-blind randomized withdrawal phase will also be analyzed using the DBFAS.

4.1.2. Double-Blind Safety Analysis Set

The double-blind safety analysis set (DBSAS) will consist of all randomized subjects who have received at least one dose of investigational product in the double-blind phase. Subjects will be reported under the treatment that they received. This analysis set will be very similar, if not the same, with the DBFAS; however, the subjects will be analyzed according to the treatment received rather than randomized.

4.2. Open-Label Analysis Sets

4.2.1. Open-Label Part 1 Analysis Set

Open-label part 1 (OLPT1) analysis set will consist of all subjects who were enrolled into the open-label part 1 phase of the study and received at least one dose of investigational product in part 1.

4.2.2. Open-Label Part 2 Analysis Set

Open-label part 2 (OLPT2) analysis set will consist of all subjects who were enrolled into the open-label part 2 phase of the study and received at least one dose of investigational product in part 2.

5. GENERAL METHODOLOGY AND CONVENTIONS

5.1. Hypotheses and Decision Rules

The primary null hypothesis in this study is that there is no difference between the distributions of time to disease flare in the double-blind randomized withdrawal phase between tofacitinib and placebo. The alternative hypothesis is that, among the subjects that meet the randomization criteria, time to disease flare is increased in the subjects remaining on tofacitinib in the double-blind withdrawal phase versus placebo subjects.

The difference between the distributions of time to disease flare will be tested using an unstratified log-rank test (Schoenfeld, 1987; Shih, 1999; Yang, 2012)[1,2,3] .

There will be potentially 2 planned analyses. The First Analysis will be performed after 28 subjects have reported flare in the double-blind phase. Number of flares required for the 2 planned analyses will be counted according to the definition of having a flare in the primary estimand (Section 2.3.1). The purpose of the First Analysis is to allow early stopping of the study for efficacy or futility, and to assess safety of tofacitinib. The investigators, subjects, and sponsor study team will remain blinded to treatment assignment in the double-blind phase through the entire duration of the trial until final database release.

If the study does not end for efficacy or futility at the First Analysis, a total of 37 subjects with flares will be required for the Final Analysis to yield an overall power of 80% using a log-rank test to detect the treatment difference, assuming a 1-sided type-I error rate of 2.5% and a 64% improvement in median time to flare (i.e., a hazard ratio of 0.36 and a median time to flare of 236 days for placebo; assumptions based on a similar sJIA randomized withdrawal trial). The number of flares needed at the Final Analysis will be based upon the actual number of flares at the First Analysis (additional details are provided below and in Appendix 2).

To protect the integrity of the study and to preserve the type I error rate at 0.025 (1-sided test) and overall study power at 80% (type II error rate β=0.2), a fraction of α for efficacy and a fraction of β for futility will be spent at the First Analysis and accounted for in the overall type I error rate and type II error rate, respectively. A formal efficacy boundary for rejecting the null hypothesis is constructed by using the spending function methodology of the gamma family design with γ= 4. Similarly, a formal futility boundary for not rejecting the null hypothesis is constructed by using the spending function methodology of the gamma family design with γ= -4.

- If the value of the test-statistic at the First Analysis crosses the efficacy boundary ($z \leq -1.973$, 1-sided $p \leq 0.0242$), the trial may be stopped for efficacy.
- If the value of the test-statistic at the First Analysis crosses the futility boundary ($z \geq -1.240$, 1-sided $p \geq 0.1074$), the trial may be stopped for futility.
- Otherwise, the trial will continue as planned. The Final Analysis will be performed after flares have been reported in 37 subjects.

Table 2. Efficacy and Futility Stopping Boundaries with 28 Flares at First Analysis Expressed as Hazard Ratio, Z Scales and P-values

Efficacy Stopping Boundary				
	Number of Flares	Hazard Ratio	Z-score	P-value (1-sided)
First Analysis	28	0.474	-1.973	0.0242
Final Analysis	37	0.427	-2.586	0.0049
Futility Stopping Boundary				
	Number of Flares	Hazard Ratio	Z-score	P-value (1-sided)
First Analysis	28	0.626	-1.240	0.1074

Table 2. Efficacy and Futility Stopping Boundaries with 28 Flares at First Analysis Expressed as Hazard Ratio, Z Scales and P-values

Final Analysis	37	0.427	-2.586	0.0049

Assuming a hazard ratio of 0.36 (tofacitinib vs placebo) and a median time to flare of 236 days for placebo
Randomization ratio = 1:1, type I error rate (1-sided) = 2.5%, power = 80.4%
Efficacy boundary: gamma family spending function with $\gamma = 4$
Futility boundary (non-binding): gamma family spending function with $\gamma = -4$

The efficacy and futility boundaries will depend on the number of flares. For logistical and administrative reasons, the actual number of flares at the First Analysis and the Final Analysis might differ slightly from those that have been pre-specified here. In that case appropriate adjustments will be made to the efficacy and futility boundaries. The actual boundaries used for the First or the Final Analysis will be re-calculated from the specified spending function based on the actual number of flares achieved at the time of First or Final analysis. These boundaries will be calculated prior to the evaluation of comparative data. More specifically, for the First Analysis, the study team will complete evaluation of subject data to determine the total number of flares in the primary analysis. The stopping boundaries and the number of flares needed at the Final Analysis will be calculated from the specified spending functions to preserve the type I error rate at 0.025 (1-sided test) and overall study power at 80% (type II error rate β=0.2). Examples of efficacy and futility boundaries in various scenarios are provided in Appendix 2.

The First Analysis will be completed by the independent Statistical Data Analysis Center (SDAC), and the results will be reviewed by the Data Safety Monitoring Board (DSMB). The efficacy and futility boundaries will be calculated as described above by the study statistician, and provided to DSMB prior to the First Analysis. Before the First Analysis is performed, the details of the objectives, decision criteria, dissemination plan and method of maintaining the study blind as per Pfizer's standard operating procedures will be documented and approved in a DSMB charter.

If the study continues and the total number of flares observed at the Final Analysis exceeds the number calculated at the First Analysis (e.g., 37 planned and 38 observed), the boundaries will be adjusted based on the actual information fraction (e.g, 38/37=1.03) to preserve the type I error rate. The adjusted boundaries will be determined prior to the evaluation of comparative data.

5.2. General Methods

In addition to the statistical methods described below, all efficacy endpoints listed in Section 3 will be summarized descriptively. In general, number and percent will be presented for binary variables; number, mean, median, standard deviation, quartiles, minimum, and maximum will be presented for continuous variables. The Kaplan-Meier method will be used to obtain the estimates of median (and other percentiles) event free time for time-to-event endpoints and estimates of event free rates at specific time points.

The analyses in the double-blind phase will be performed by treatment group (tofacitinib and placebo), per Section 5.2.5. For this time-to-event trial, subjects will have variable follow-up

times. Efficacy analyses by-visit in the double-blind randomized withdrawal phase will be performed at Week 4 and every 4 weeks through the last visit of the whole study. However, only the data through 52 weeks in the double-blind phase will be analyzed statistically (Section 2.3.2).

The analyses in the open-label phase will be performed separately for Part 1 and Part 2. The summaries will be presented for all tofacitinib.

All efficacy endpoints will be derived programmatically by Pfizer with the exception of double-blind flare, which will be based on real-time calculations by the CCC.

5.2.1. Analyses for Time-to-Event Data

Time-to-event endpoint, i.e., time-to-flare in the double-blind randomized withdrawal phase, will be summarized using the Kaplan-Meier method and estimated survival curves will be displayed graphically. The median and quartiles for the time to event and survival (event-free) probabilities at particular time points along with 95% confidence intervals (CIs) will be estimated. The 95% CIs for both the estimated probability of event and the probability differences between the treatment groups at the target day of each windowed visit (Appendix 3.13) will be generated using Greenwood's formula.

Details on the censoring of time-to-flare for the primary analysis can be found in Section 6.1.1 and for supportive analyses can be found in Section 6.1.2.

The unstratified log-rank test will be used to compare the time to event between the treatment groups.

Hazard ratios (tofacitinib/placebo) and 95% CIs will be obtained from a Cox Proportional hazards model with treatment group as covariate. Repeated confidence interval will also be provided to account for alpha spending, when appropriate. This will be done as a supportive analysis for the log-rank test comparison between the treatment groups.

5.2.2. Analyses for Binary Data

Open-label binary endpoints will be analyzed using the normal approximation to binomial proportion.

The normal approximation 95% CI is calculated as:

$$\hat{p} \pm z_{0.975}\sqrt{\frac{\hat{p}(1-\hat{p})}{N}}$$

where $\hat{p}$ is the estimated response rate, $z_{0.975}$ is the 97.5th percentile of the standard normal distribution and N is the number of subjects evaluable for the endpoint at the time point. If the lower bound is calculated to be negative, it will be set to 0%; if the upper bound is calculated to be larger than 100%, it will be set to 100%. In case when response rate is 0 or 100%, standard error will be reported as "NA" and the 95% CI bounds will be the same as the response rate.

For the double-blind binary endpoint at a single time point (eg, adapted JIA ACR 30/50/70/90/100 response at each DB phase visit), the normal approximation for the difference in binomial proportions can be used to test the superiority of tofacitinib 5 mg BID to placebo at DB phase timepoints. The normal approximation to the test statistic for the difference in binomial random variables is calculated as

$$Z = \frac{\hat{p}_t - \hat{p}_c}{\sqrt{\frac{\hat{p}_t(1-\hat{p}_t)}{N_t} + \frac{\hat{p}_c(1-\hat{p}_c)}{N_c}}},$$

where $\hat{p}$ refers to the relative frequency, N to the number of observations, the subscript c refers to the comparator group (e.g., placebo group) and the subscript t refers to the test group (e.g., tofacitinib 5 mg BID group) so that test statistics and p-values can be calculated for the contrast comparing the two treatment groups.

Two-sided 95% confidence intervals will be formed by

$$\hat{p}_t - \hat{p}_c \pm z_{0.975}\sqrt{\frac{\hat{p}_t(1-\hat{p}_t)}{N_t} + \frac{\hat{p}_c(1-\hat{p}_c)}{N_c}}$$

Two-sided p-value for the test of the 0 difference between tofacitinib 5 mg BID and placebo groups will be calculated as:

$$p = 2(1 - \Phi(|Z|)),$$

where Φ(.) is the Gaussian cumulative density function.

If there is no (ie, 0) response or 100% response in any one or both of the two treatment groups for the comparison, eg, tofacitinib 5 mg BID vs. placebo, when calculating the proportions above, 0.5 will be added to the number of responses (ie, numerator) and 1 will be added to the denominator in each treatment corresponding to the pair of comparison for calculating the treatment difference, standard error (ie, $\sqrt{var(d)}$), 95% CI and 2-sided p-value (Agresti, 2002)[4]. (Note that this adjusted method also applies to the case where the response rate is 0% in one treatment group and 100% in the other for comparison.)

When response rate of 0% is observed in both treatments or 100% in both treatments in comparison, no formal comparison will be performed. Estimated response rate of 0% or 100% will be reported as observed. Standard error will be reported as 0.

The final results will be expressed in percentages, i.e., (proportions x 100)%.

5.2.3. Analyses for Continuous Data

For the continuous secondary endpoints in the DB phase, a mixed model for repeated measures (MMRM) will be applied. The repeated measures mixed model will contain fixed effects for treatment (tofacitinib and placebo), visit, treatment by visit interaction, baseline value, baseline value by visit interaction. The baseline value in the model will be either DB

baseline value or OL baseline value, depending on the definition of the endpoint (ie, change from DB baseline or change from OL baseline).

An unstructured covariance matrix will be assumed with a Kenward and Roger method for degrees of freedom approximation provided the model converges, otherwise an alternative covariance matrix, eg, heterogeneous compound symmetry (CSH) will be attempted. If it still does not converge, compound symmetry will be attempted. If none of the covariance structures is computationally feasible, other approaches may be considered.

5.2.4. Treatments in the Open-Label Phase

Though tofacitinib 10 mg BID dose is no longer evaluated in this study, a dose increase from 5 mg BID to 10 mg BID was allowed based on previous protocol for individuals who are unable to control fever within the first 2 weeks of treatment with tofacitinib 5 mg BID in Part 1, at the investigator's discretion. These dose switchers will be grouped together with the other pure 5 mg BID treated subjects for the OL phase analyses, with the number of dose switchers footnoted in related TFLs.

5.2.5. Treatments in Double-Blind Withdrawal Phase

At the beginning of the double-blind randomized withdrawal phase, subjects from the open-label run-in phase will be randomized into 2 sequences: tofacitinib 5 mg BID (continuation of tofacitinib treatment), or placebo (withdrawal of tofacitinib). In the primary and secondary double-blind analyses, the two treatment groups (tofacitinib or placebo) will be compared to assess the treatment effect. Some descriptive analyses will be performed.

5.3. Methods to Manage Missing Data

In general, missing values will not be imputed for descriptive statistics. In addition, missing values for safety endpoints will not be imputed and no data in the open-label phase analyses will be imputed, except for derivations of specific variables outlined below. For the endpoints not mentioned in the following sections, missing data will not be imputed. As limited number of subjects were enrolled with short follow up time in the double-blind phase before the COVID-19 pandemic anchor date, the impact of COVID-19 pandemic is limited. Missing due to COVID-19 pandemic will be treated the same way as the other missing data.

5.3.1. Disease Flare

Disease flare in the double-blind phase will be assessed in real-time by the Centralized Coordinating Center (CCC) according to the CCC's SOP and used in the analyses.

Every effort will be made to collect the components required for the real-time flare calculation. Per CCC's SOP, in the event that individual JIA core set components are missing at a given double-blind visit (while the subject is still ongoing in the DB) and that the available data are inconclusive for the determination of flare, the CCC will utilize the missing component(s) value from the previous visit. This type of mixed components LOCF method is based on calculating the composite value using a mixture of values at a visit and values carried forward from previous visits, which may include the double-blind baseline if no post-baseline values are available.

Different imputation methods of disease flare dates for the censored data are described separately in Sections 6.1.1, 6.1.2.1 and 6.1.2.2.

5.3.1.1. Censoring Times and Risk Periods for Disease Flare

The first event will be counted within the defined risk periods below. If a subject does not have an event or has an event but outside the risk period, the subject will be censored at the last available flare assessment within the risk period. Subjects with no post-randomization flare assessments will be censored at randomization date.

Define last contact date as the maximum of (last study visit date, subject withdrawal from study date, telephone contact date). If a subject dies, last contact date is the death date.

On-Treatment Censoring Time and Risk Period

The risk period corresponding to the on-treatment censoring time is defined from the date of the randomization to the minimum of (last contact date, last investigational product dose date + 2 days). The addition of 2 days to the last investigational product dose date permits the data collected 2 days after the last investigational product dose be considered as on-treatment. Explicitly, the risk period is calculated as the minimum of (last contact date, last investigational product dose date + 2 days) – date of randomization + 1.

This risk period will be used in the primary analysis (Estimand 1) and a supplementary analysis (Estimand 4) for time to disease flare.

On-Study Censoring Time and Risk Period

The risk period corresponding to the on-study censoring time is defined from the date of the randomization to the last contact date. Explicitly, the risk period is calculated as the last contact date – date of randomization + 1. For subjects who withdraw from study early, before reaching 52 weeks of double-blind treatment from Study A3921165 and enrolled in the LTE study (Study A3921145), the last contact date will be extended to the upper limit of the DB Week 52 visit window post randomization through both studies A3921165 and A3921145 (ie, Randomization Day + 378) (See Appendix 3.13).

This risk period will be used in a supplementary analysis (Estimand 5) for time to disease flare.

5.3.2. Binary Endpoints

For binary secondary endpoints in double-blind phase, intercurrent event will be handled as described in Section 2.3.2.1 (Estimand 2) or Section 2.3.4 (Estimand 6). After the intercurrent event is handled as appropriate, any remaining missing responses will be handled by setting the response value to nonresponsive for visits up to DB Week 52. This method of handling missing responses is known as missing response as non-response (MR=NR).

5.3.2.1. Adapted JIA ACR30-50-70-90-100 Response

The Adapted JIA ACR30-50-70-90-100 responses will be derived programmatically and used in analyses. If the value in any of the components at a timepoint is missing, the component variables that are not missing will be used to determine the response status. As a general principle, if there are sufficient non-missing components to determine whether the ACR endpoint is a response or non-response, then ACR endpoint is not missing, else if the available non-missing components are not sufficient to determine the response status of ACR endpoint then it is considered missing.

Missing responses will not be imputed for descriptive summaries. For Estimand 2 and Estimand 6 analyses, missing responses will be handled as described in Section 5.3.2.

5.3.2.2. JADAS-27 CRP, JADAS-27 ESR, and Occurrence of JADAS Minimum Disease Activity and Inactive Disease

For continuous measures (change from double-blind baseline in JADAS-27 scores), if any component is missing, then JADAS-27 CRP/ESR will be set to missing. These missing data will not be imputed otherwise and will be handled using the MMRM as described in Section 5.2.3.

For binary assessments (occurrence of JADAS minimum disease activity and inactive disease), detailed definitions based on JADAS-27 CRP/ESR are in Appendix 3.6. If JADAS-27 CRP/ESR value is missing, the corresponding binary assessments will be considered missing. Missing responses will not be imputed for descriptive summaries. For Estimand 2 analysis, missing responses will be handled as described in Section 5.3.2.

5.3.2.3. Presence of JIA ACR Inactive Disease and Clinical Remission

Calculations of JIA ACR inactive disease using ESR are as in Appendix 3.7. If any of the components is missing, the components that are not missing will be used to determine the Clinical Inactive Disease Status as follows:

- Inactive disease: all components must be present (non-missing) and meet all conditions.
- Active disease: at least one component is present (non-missing) and does not meet the condition.
- Not determined: any other scenarios.

Clinical remission will be derived from inactive disease except when LOCF is applied in Estimand 2 as described in Section 2.3.2.1. Clinical remission corresponds to at least 24 weeks of inactive disease continuously while on investigational product (for evaluation of this endpoint in the double-blind phase, the counting of 24 weeks starts from the double-blind phase Day 1).

5.3.3. JIA ACR Core Set Variables

All JIA ACR Core Set Variables are continuous endpoints, missing data will be handled using the MMRM as described in Section 5.2.3 and will not be imputed otherwise.

5.3.4. CHQ and CHAQ Responses

CHQ and CHAQ responses are all continuous endpoints, missing scores will be handled using the MMRM as described in Section 5.2.3 and will not be imputed otherwise.

6. ANALYSES AND SUMMARIES

6.1. Primary Endpoint: Time to Disease Flare in Double-Blind Withdrawal Phase

6.1.1. Primary Analysis

- Estimand strategy: Composite (Section 2.3.1).
- Analysis set: DBFAS (Section 4.1.1)
- Analysis methodology: Methods described in Section 5.2.1 and decision rule in Section 5.1 will be followed.
- Censoring time and risk period: On-Treatment Censoring Time and Risk Period (Section 5.3.1.1).
- Intercurrent events and missing data: See definitions described in Section 2.3.1 for intercurrent events; and Section 5.3.1 for missing values.

6.1.2. Supplementary/Sensitivity Analyses

6.1.2.1. Supplementary Analyses

Supplementary Analysis 1

- Estimand strategy: hypothetical (Section 2.3.3).
- Analysis set: DBFAS (Section 4.1.1)
- Analysis methodology: methods described in Section 5.2.1.
- Censoring time and risk period: On-Treatment Censoring Time and Risk Period (Section 5.3.1.1).
- Intercurrent events and missing data: See definitions described in Section 2.3.3 for intercurrent events; and Section 5.3.1 for missing values.

Supplementary Analysis 2

- Estimand strategy: treatment policy (Section 2.3.3).
- Analysis set: DBFAS (Section 4.1.1)

- Analysis methodology: methods described in Section 5.2.1.
- Censoring time and risk period: On-Study Censoring Time and Risk Period (Section 5.3.1.1).
- Intercurrent events and missing data: See definitions described in Section 2.3.3 for intercurrent events; and Section 5.3.1 for missing values.

Supplementary Analysis 2 will include observations that occur after the intercurrent event of investigational product discontinuation. Subjects discontinuing investigational product in the double-blind phase, for reasons other than flare, before reaching 52 weeks of double-blind treatment will continue to be followed up for flares either via scheduled study visits in this study or in the LTE study A3921145 (for subjects who withdraw from this study and roll over) for up to 52 weeks post randomization. For data coming from A3921145 study, disease flare will be programmatically evaluated using the same approach by CCC as in this study. Flares occurring after investigational product discontinuation in A3921165 study come from either the CCC flare assessment at regularly scheduled study visits or the follow-up visit post investigational product discontinuation.

6.1.2.2. Sensitivity Analyses: Tipping Point Analysis

Sensitivity Analysis 1

This analysis will be performed only when superiority is achieved based on the primary analysis for the primary endpoint (Section 6.1.1).

The impact of missing data (ie, censoring) on the primary analysis (composite estimand) will be evaluated through model-based multiple imputation. For subjects who are censored under the composite estimand, the imputation will be done to the LSLV of Study A3921165. Imputation model will be fit for patients assigned to tofacitinib and placebo.

Tipping point analysis will be conducted to explore the plausibility of missing data assumptions under which the conclusions change, i.e., under which there is no longer evidence of a treatment effect. The tipping point analysis will be two-dimensional, i.e., will allow assumptions about the censored times on the two arms to vary independently.

In the tipping point analysis approach, the post-censoring hazard of flares for subjects with censored time to flare will be varied over a grid of sensitivity parameters that increase or decrease the hazard function relative to the CAR (censoring at random) analysis in both the tofacitinib and placebo groups.

The primary endpoint is time to disease flare in the double-blind phase. Patients who are censored without disease flare contribute partial information about the endpoint, which can be analyzed using standard statistical methods for survival data. Weibull regression model will be fit for both observed and censored data as the imputation model. Statistical simulations were conducted suggesting estimation bias for Weibull parameters was minimal when the Weibull shape parameter was ≥ 1. The assumed Weibull shape parameter of 1 was supported by data of a prior completed pJIA study (A3921104). Conditional on the observed

censored time, a time-to-flare will be imputed, subject to the end-of-study constraint as follows: use the imputed time-to-flare if the imputed time-to-flare is prior to the end of study (ie, LSLV of Study A3921165); but censored at the end of the study, if the imputed time-to-flare exceeds the LSLV date of Study A3921165.

The completed imputed data sets will be analyzed using the same approach as the primary analysis (ie, unstratified log-rank test) in Section 5.2.1. This will be repeated for R times per each scenario. The multiple imputation analysis will be applied to each set of the two sensitivity parameters in the two-dimensional tipping point analysis.

More detailed descriptions are provided in Appendix 5.

Sensitivity Analysis 2

This analysis will be performed only when superiority is achieved based on the primary analysis for the primary endpoint (Section 6.1.1) and the result of Supplementary Analysis 2 (Section 6.1.2.1) is consistent with the primary analysis.

The impact of missing data (ie, censoring) on Supplementary Analysis 2 (treatment policy estimand, Section 6.1.2.1) will be evaluated through model-based multiple imputation. For subjects who are censored under the treatment policy estimand and not enrolled in the LTE study (Study A3921145), the imputation will be done to the LSLV of Study A3921165. For subjects who are censored under the treatment policy estimand and enrolled in the LTE study (Study A3921145), the imputation will be done to the LSLV of Study A3921165 or to Randomization Day + 378 (the upper limit of the DB Week 52 visit window), whichever comes later. Imputation model will be fit for patients assigned to tofacitinib and placebo.

Weibull regression model will be fit for both observed and censored data as the imputation model. Conditional on the observed censored time, a time-to-flare will be imputed, subject to the constraint as follows: use the imputed time-to-flare if the imputed time-to-flare is within the risk period for imputation defined above; but censored at the end of the risk period for imputation, if the imputed time-to-flare exceeds the risk period for imputation.

A tipping point analysis similar to Sensitivity Analysis 1 will be conducted to explore the plausibility of missing data assumptions under which the conclusions change, i.e., under which there is no longer evidence of a treatment effect.

More detailed descriptions are provided in Appendix 5.

6.2. Secondary Endpoints

The secondary endpoints will be analyzed in the open-label phase and by treatment group in the double-blind phase. For further information on the treatment groups please refer to Section 5.2.4 and Section 5.2.5. For missing data handling of the secondary endpoints, please refer to Section 5.3.

6.2.1. Secondary Endpoints Analysis – Double-Blind Withdrawal Phase

The double-blind secondary endpoint analyses will be carried out using the DBFAS population, by treatment group (tofacitinib and placebo).

6.2.1.1. Binary Secondary Endpoints in Double-Blind Phase

Except the endpoint of occurrence of disease flare by visit, which is estimated in the analyses of the primary endpoint (Section 6.1.1 and Section 6.1.2.1), all other DB binary secondary endpoints listed in Section 3.2.1 (aJIA ACR 30/50/70/90/100 response, absence of fever, occurrence of inactive disease and minimal disease activity, occurrence of JIA ACR inactive disease and clinical remission), are analyzed as below:

- Estimand strategy: Composite (Section 2.3.2.1).
- Analysis set: DBFAS (Section 4.1.1).
- Analysis methodology: Descriptive summary in Section 5.2 and analysis methods in Section 5.2.2 will be followed.
- Intercurrent events and missing data: See definitions described in Section 2.3.2.1 for intercurrent events and Section 5.3.2 for missing data.
- The basic summary statistics will be calculated by visit. The normal-approximated 95% CI and 2-sided p-value for the treatment effect will be presented at each scheduled visit in the double-blind phase through Week 52.

6.2.1.2. Continuous Secondary Endpoints in Double-Blind Phase

Change from open-label baseline in each sJIA ACR core variable, except for disability index from the CHAQ for which change from double-blind baseline will be considered, will be analyzed as below:

- Endpoints:
  - Change from OL baseline in number of joints with active arthritis
  - Change from OL baseline in number of joints with limited range of motion
  - Change from OL baseline in physician global evaluation of disease activity
  - Change from OL baseline in parent/legal guardian/subject evaluation of overall well-being [from the CHAQ]
  - Change from DB baseline in disability index from the CHAQ
  - Change from OL baseline in ESR
- Estimand strategy: Treatment policy (Section 2.3.2.2).

- Analysis set: DBFAS (Section 4.1.1).
- Analysis methodology: Descriptive summary in Section 5.2 and analysis methods in Section 5.2.3 will be followed.
- Intercurrent events and missing data: See definitions described in Section 2.3.2.2 for intercurrent events and MMRM described in Section 5.2.3 for missing data.
- The basic summary statistics will be calculated by visit. The least-squares (LS) means, the 95% confidence interval for the LS means, the between treatment difference, the corresponding 95% confidence interval and 2-sided p-values for the main treatment effect will be presented for change from baseline value at each scheduled visit in the double-blind phase through Week 52.

Analysis details of other continuous secondary endpoints are provided in Section 6.2.1.3.

6.2.1.3. Supplementary Analyses

For aJIA ACR 30/50/70/90/100 response at each scheduled visit in the double-blind phase through Week 52, an additional analysis using the treatment policy estimand strategy (Section 2.3.4) will be performed. It will use the same methodology and summary as in Section 6.2.1.1 but will also include data collected after the intercurrent events of investigational product discontinuation, ie, all data collected while On- and Off-investigational product until Week 52 will be used to derive the endpoints.

For all DB continuous secondary endpoints listed in Section 3.2.1 (change from DB baseline in JADAS-27 ESR/CRP, change from both OL and DB baseline in each sJIA ACR core variable, change from DB baseline in CHQ responses and CHAQ discomfort index), an analysis using the hypothetical estimand strategy (Section 2.3.4) will be performed. It will use the same methodology and summary as in Section 6.2.1.2 but only the on-treatment data (Appendix 3.13) will be used.

6.2.2. Secondary Endpoints Analysis – Open-Label Phase

In general open-label analyses will be broken down by Part 1 and Part 2, and open-label overall, unless otherwise specified. There will be no treatment comparisons for the open-label analyses and missing data will not be imputed, unless otherwise specified for the calculation of specific endpoints. Summaries by visit will be done separately for Part 1 and Part 2, using the corresponding Part 1 or Part 2 open-label analysis sets (OLPT1 and OLPT2).

For below binary endpoints, descriptive summary statistics (number and percentage) mentioned in Section 5.2 will be presented. Percentages will be calculated out of the number of subjects with non-missing values at each visit.

- Successfully tapering at the end of Part 2 of the open-label phase and Achievement of corticosteroid dose of ≤0.2 mg/kg/day or 10 mg/day at the end of OL Part 2 phase

  Analyses will be conducted using the OLPT2 analysis set.

- Adapted JIA ACR30/50/70/90/100 responses at each visit from Day 7 onward
- JADAS-27 Inactive Disease and Minimal Disease Activity at each visit
- Absence of Fever (lack of fever due to sJIA in the week preceding the assessment) at each visit
- JIA ACR Inactive Disease at each visit
- Occurrence of fever attributed to sJIA on Day 3, 7 and 14

  Summary will be for all the tofaticinib OLPT1 subjects.

- CRP ≤ 10 mg/L at every visit of the open-label phase

Descriptive statistics as described in Section 5.2 will be tabulated for:

- JADAS-27 ESR and CRP scores and change from OL baseline values at each visit
- CHAQ responses and change from OL baseline values at each visit

  Details on the calculation of CHAQ responses can be found in Appendix 3.9.

- CHQ responses and change from OL baseline values at the end/last visit of Part 1 and Part 2 of the open-label phase
- JIA ACR Core Set Variables and change from OL baseline values at each visit

Besides the descriptive summary mentioned in Section 5.2, Kaplan-Meier methods described in Section 5.2.1 will be used to analyze below time-to-event endpoint:

- Time to Adapted JIA ACR30 Response in Part 1 of the open-label phase

  Analysis will be done using the OLPT1. Subjects that do not achieve an adapted JIA ACR30 response in Part 1 (withdraw from the study) will be censored at their last available response assessment in Part 1. Time to response will be counted since the first dose of tofacitinib in the study for all the subjects.

6.3. Other Endpoint(s)

6.3.1. Analysis of PK Endpoints

Plasma tofacitinib concentration data will be listed by individual subject and time point. Oral clearance (CL/F) and other pharmacokinetic (PK) parameters calculated from plasma tofacitinib concentrations are out of scope of this SAP and will be reported separately.

6.3.2. Analysis of Exploratory Endpoints

Exploratory or pooled analyses, if conducted utilizing the biobanked genomic and/or biomarker samples from this study, will be documented in a separate analysis plan.

6.4. Subset Analyses

No subgroup analyses will be done for OL endpoints. The primary endpoint (time to flare in the DB phase) and some secondary endpoints (occurrence of disease flare, achieving ACR 30/50/70 and JIA ACR inactive disease at each scheduled visit in double-blind phase) will be summarized by age group at enrollment, baseline body weight, geographical region, formulation, and treatment history (OL Day 1 oral corticosteroid use, OL Day 1 MTX use, and prior bDMARD experience) by treatment and visit. Details are as follows.

Age Group Analyses

Subgroup analyses will be presented for age group at enrollment: from 2 to <6 years, from 6 <12 years, and from 12 to <18 years.

The primary endpoint of time to flare in the DB phase will be analyzed by these subgroups. The same methodology and summary as the primary analysis in Section 6.1.1 will be done (Estimand 1). I.e, K-M estimates of median time to flare and quartiles, as well as probabilities of not having flare up to all the DB monthly visits will be generated for the tofacitinib and placebo treatment groups by age subgroup. Hazard ratios (tofacitinib/placebo) and corresponding 95% confidence intervals may be provided for the time-to-flare analysis. The secondary endpoints of ACR 30/50/70 response and JIA ACR inactive disease by visit in the DB phase will be analyzed similarly as in Section 6.2.1.1 by age subgroup (Estimand 2).

Subgroup analyses similar to the above age group analyses (Estimand 1 and Estimand 2) will also be presented for:

- Baseline body weight: <40 kg and ≥40 kg
- Geographic region: North America [US and Canada], South and Central America [Brazil, Argentina, Mexico, Costa Rica], Europe [Poland, Belgium, Great Britain, Spain, Hungary, Germany, and Italy], Asia [China, India], and Rest of World [Ukraine, Turkey, Russia, Australia, Israel, and South Africa]
- Formulation: tablet and solution

  Note: subjects who take both formulations will be excluded
- OL Baseline (Day 1) oral corticosteroid use: Yes and No
- DB Baseline (Randomization Day) oral corticosteroid use: Yes and No
- OL Baseline (Day 1) MTX use: Yes and No
- Prior bDMARD experience (any time on or before OL Day 1): Yes and No

6.5. Baseline and Other Summaries and Analyses

Open-label baseline demographic characteristics will be summarized by age group at enrollment.

Subject Disposition

Subject disposition will be summarized for both Part 1 and Part 2 of the open-label phase (based on the OLPT1 and OLPT2 analysis sets), including how many subjects entered and completed each part, discontinued investigational product and withdrew from study in each part. The reasons for both investigational product discontinuation and study withdraw from Part 1 and Part 2, and from the double-blind phase will be summarized. In addition, the number of subjects being randomized directly after completion of Part 1 and those randomized after completing Part 2 will be provided. The number of subjects entering, not entering and reasons for not entering A3921145 study from open-label or double-blind phase will also be presented.

The Part 1 and 2 open-label disposition tables will be for all tofacitinib treated subjects. Number of subjects who increased from 5 mg BID to 10 mg BID on Day 14 per previous protocol (original protocol and Protocol Amendment 1) will be footnoted.

For the double-blind phase the number of subjects randomized to each treatment group: tofacitinib 5 mg BID, placebo will be summarized and the reasons for discontinuation from investigational product in DB phase will be displayed based on the DBFAS.

Drug Exposure

For each phase (OL and DB) the subject's weight, prescribed daily dose, actual daily dose, route, formulation (tablet or solution) along with the start and stop dosing dates and reasons for missed doses will be displayed in listings. The start date of Part 1, Part 2 will be included in the listings to distinguish the doses taken in each OL phase. For subjects receiving oral solution prescribed daily dose will be determined based on the subject's assigned nominal dose and the subject's weight (Table 3).

Table 3. Investigational Product Dosing and Administration

Body Weight (kg)	Dosage Regimen (Run-In Phase: Tofacitinib, Double-Blind Phase: Tofacitinib/Placebo)
	5 mg BID nominal dose
5-<7	2 mg (2 mL oral solution) BID
7-<10	2.5 mg (2.5 mL oral solution) BID
10-<15	3 mg (3 mL oral solution) BID
15-<25	3.5 mg (3.5 mL oral solution) BID

Table 3. Investigational Product Dosing and Administration

25-<40	4 mg (4 mL oral solution) BID
≥40	5 mg (one 5 mg tablet or 5 mL oral solution) BID

Descriptive statistics for duration of treatment will be provided for Part 1 of the OL (based on OLPT1), Part 2 of the OL (based on OLPT2), and the DB (based on DBSAS).

Additionally, the number and percentage of subjects falling into categories of treatment duration will be summarized. For Part 1 of the OL, the categories will be ≤4 weeks, >4 weeks-≤8 weeks, >8 weeks-≤12 weeks, >12 weeks-≤16 weeks. For Part 2 of the OL 4-week categories will be similar and going up to a maximum of 24 weeks. For the DB 4-week categories will be used as well, with the maximum interval determined by the duration of the study.

6.6. Safety Summaries and Analyses

Safety analyses will be carried out using the safety analysis sets for the open-label Part 1, Part 2, and double-blind phase (OLPT1, OLPT2, DBSAS).

All the safety data, including the following, will be summarized descriptively through appropriate data tabulations, descriptive statistics, and graphical presentations:

- Safety laboratory tests will be summarized according to Pfizer standards. Changes from baseline in lab and vital sign values of randomized subjects will utilize both the open-label baseline and the double-blind baseline.
- All the following safety endpoints will be summarized using non-standard tables: serious infections (based on AE database with SOC="Infections and Infestations" and marked as serious), herpes zoster (including both AE database and adjudicated data), adjudicated opportunistic infections, cytopenias, adjudicated malignancies, adjudicated gastrointestinal (GI) perforations, adjudicated cardiovascular events, adjudicated MAS, clinically significant abnormal laboratory parameters (absolute neutrophil, absolute lymphocyte, serum creatinine, liver tests, lipid levels, and hemoglobin), and validated assessments of growth (body weight and height, refer to Appendix 3.10 for details) and pubertal development (Tanner Stage of Development).
- Subjects meeting discontinuation and monitoring criteria will be summarized.

6.6.1. Adverse Events

Adverse events will be summarized according to Pfizer standards. AE's will be summarized by the separate phases: Part 1, Part 2 of the open-label, and double-blind.

An adverse event will be considered a treatment-emergent AE (TEAE) if the event starts during the treatment phase of interest. In all these analyses, all AEs (ie, an infinite lag) will be counted within each study phase, with the reception of the incidence rate (IR) calculation

for events of interest, where the risk period will be up to 28 days within the last investigational product dose, as defined below.

Selected AEs in double-blind phase will be analyzed using incidence rate (IR) and 95% CI and hazard ratio (HR) and 95% CI using a Cox model, including any TEAE, SAE, adjudicated MAS, serious infections (SI) including tuberculosis, varicella and herpes zoster (HZ) and adjudicated opportunistic infections (OI), adjudicated malignancies, adjudicated GI perforations, and adjudicated CV events.

The IR (expressed in unique number of subjects with events per 100 subject-years) for double-blind phase will be defined as the number of subjects with at least one event during the risk period divided by the sum of the durations of exposures for the subjects during the risk period.

The risk period (RP) for the double-blind phase for each subject will span from day of randomization to the earliest of last contact date and last investigational product dose date + 28 days. The last contact date will be the maximum of: AE start date, AE stop date, last study visit date, withdrawal from study date and telephone contact date. If a subject dies, last contact date will be the death date. First events will be counted within the risk period. If a subject did not have an event or had an event but outside the risk period, the subject will be censored at the end of risk period.

IR for DB= (#of subjects with first events during the DB risk period)/

(sum of durations of subject exposure during the risk period)

The numerator of the IR will include subjects counted only once as soon as they experience a first event during their individual risk period defined above.

The denominator will be the sum of the times to event for those who experience an event during the risk period and the lengths of risk periods for those who do not experience an event. If a subject experiences multiple instances of the same event, the time to first occurrence of the event will be considered.

The calculated rate is then multiplied by 100 to yield an IR in units of the number of subjects with events per 100 subject-years. Exact 95% Poisson CI will be provided for IRs as follows.

The 95% CI (corresponding to α=0.05) for this IR will be calculated based on the assumption that the actual count of cases (ie, unique subjects with events) arises from an exact Poisson distribution. The approach used in Daly (1992)[5] describing the calculation of exact confidence limits will be used. If x denotes the actual number of subjects with events, the $(1 - \alpha)\times 100\%$ confidence limit, x_L the lower confidence limit and x_U the upper confidence limit, then the following formula yields values for x_L and x_U,

When $x > 0$,

$$x_L = 0.5 * \chi^2(\alpha/2; 2x) \text{ and } x_U = 0.5 * \chi^2(1- \alpha/2; 2x+2),$$

When $x = 0$,

$$x_L = 0 \text{ and } x_U = 0.5 * \chi^2(1- \alpha/2; 2x+2).$$

In the above notation, $\chi^2(p; n)$ represents the p^{th} quantile of the chi-square distribution with n degrees of freedom. Once x_L and x_U are obtained, they are to be adjusted for the at-risk population per 100 subject-years by multiplying by 100.

In addition, the number and percentage of subjects with MAS including both the adjudicated events and per protocol lab criteria will be obtained for Part 1, Part 2 of OL, and DB. The later will be programmatically derived using the following algorithm from protocol Section 7.2.3:

Per protocol, a subject is classified as having MAS if the subject has:

- Fever at a visit plus Ferritin >684 ng/mL

AND at least 2 of the following 4 laboratory variables:

- Platelets ≤181 x 10^9 /L.
- AST >48 U/L.
- Triglycerides >156 mg/dL.
- Fibrinogen ≤360 mg/dL.

A supporting listing with the subjects above mentioned lab values and fever will be provided for the subjects meeting MAS criteria per protocol. Subjects with adjudicated MAS will also be listed.

6.7. Additional Analyses Depicting COVID-19 Pandemic Impact

As this study just randomized a few subjects with short follow up time in the double-blind phase before the COVID-19 pandemic anchor date (09 January 2020 for China sites, and 11 March 2020 for the remaining countries) and majority subjects are enrolled during the pandemic, the impact of COVID-19 pandemic is limited. Thus, no additional efficacy analysis will be done.

In order to report the impact of COVID-19 on clinical trial populations and study data, the following additional listings and summaries will be produced for both open-label and double-blind phases, unless noted otherwise:

- A separate summary table solely for subject discontinuations from investigational product and withdrawal from study related to COVID-19 pandemic, if any, will be provided

- Protocol deviations related to COVID-19 pandemic will be summarized and listed separately. Both important and non-important PDs related to COVID-19 pandemic will be reported

- COVID-19 related AEs, if any, will be reported separately.

7. INTERIM ANALYSES

The planned analysis is as described in Section 5.1. This analysis will be performed after 28 subjects have reported flare in the double-blind phase. The First Analysis will be completed by the independent SDAC, and the results will be reviewed by the DSMB. More details of the decision criteria, dissemination plan and method of maintaining the study blind will be provided in a separate First Analysis SAP and the DSMB charter.

Unplanned interim analyses (blinded) may be performed to support regulatory submissions in other indications.

7.1. Centralized Coordinating Center (CCC)

This study will utilize a CCC to review and confirm, in real time, efficacy assessments, including sJIA fever assessment, JIA core set variables, extra-articular sJIA features, Adapted JIA ACR response, sJIA disease flare and inactive disease, according to validated criteria, at various time points throughout the study.

In addition, the CCC will review corticosteroid (CS) regimens of study participants. The CCC will suggest a customized 4-week CS tapering schedule for each subject entering Part 2. Subjects are required to contact the investigator if they experience any symptoms of a sJIA flare; if required, an extra unscheduled visit will be conducted to assess the adverse event. The tapering schedule will be re-evaluated in consultation with the CCC.

The CCC will also provide recommendations in case participants are not able to maintain a stable CS dose in the double-blind phase. Furthermore, the CCC will provide guidance regarding discontinuation of any subject based on CS regimen.

For data analysis Pfizer will calculate all the efficacy endpoints in-house, except for double-blind flare, which will be obtained from the CCC's real-time assessments mentioned above.

7.2. Data Safety Monitoring Board (DSMB)

An independent Data Safety Monitoring Board (DSMB), a group of experts external to Pfizer, will review accumulating safety data from this study on an ongoing basis within the context of the Phase 3 pediatric program as well as adult program. Based on these reviews, the DSMB will have the capacity to make recommendations to Pfizer that might impact the future conduct of the trial. The recommendations made by the DSMB to alter the conduct of the study will be forwarded to Pfizer for final decision. Pfizer will forward such decisions, which may include summaries of aggregate analyses of endpoint events and of safety data which are not endpoints, to regulatory authorities, as appropriate. The DSMB will have access to unblinded treatment information from concurrently ongoing double-blind studies

during the clinical trial. The management and process of this committee will be in accordance with Pfizer's Standard Operating Procedures and will be documented in the DSMB Charter. The DSMB members will all be individuals who are independent of Pfizer. A DSMB Liaison will be appointed; this is an individual who represents Pfizer to coordinate communications and facilitates access to Pfizer's resources, but is not involved in the study design, study management, site management, data accrual, or study analysis. Records of DSMB meetings, interactions with Pfizer contacts, assessments and recommendations and materials reviewed will be maintained and kept proprietary and confidential by the DSMB.

In addition, the DSMB will also evaluate efficacy data at the First Analysis and make a recommendation regarding study continuation based on observed results of the study. The recommendations made by the DSMB will be forwarded to the Sponsor for final decision. The independent Statistical Data Analysis Center (SDAC) will prepare data for DSMB review. Clinical sites and sponsor study team will be restricted from access to study results until the conclusion of the study.

Further information about the DSMB can be found in the DSMB Charter, which outlines the operating procedures of the committee, including specific description of the scope of their responsibilities, including a plan where communication timelines are defined.

7.3. Steering Committee

The Sponsor will use a Steering Committee to provide guidance regarding evaluation of the 5 mg BID dose level of tofacitinib in the initial cohorts of this study. In the open-label phase the Steering Committee will assess preliminary efficacy and safety of the 5 mg BID dose level based on a review of 2 cohorts, each receiving tofacitinib 5 mg BID. The Steering Committee will provide recommendations regarding further enrollment at the 5 mg BID dose level.

Additional details will be provided in a separate Steering Committee Charter.

7.4. Safety Endpoint Adjudication Committee

To help assess specific safety events in this and other studies in the tofacitinib program, adjudication committees have been established to harmonize and standardize selected safety event assessment. Members of these safety event adjudication committees will be blinded to treatment assignment in order to allow for unbiased assessments. These committees include a Cardiovascular Endpoint Adjudication Committee (CV EAC), Malignancy Adjudication Committee (MAC), Opportunistic Infection Review Committee (OIRC), Hepatic Event Review Committee (HERC) and Gastrointestinal Perforation Review Committee (GIPRC), and Macrophage Activation Syndrome (MAS) Review Committee.

Further information about these committees can be found in the respective charters, including specific descriptions of the scope of their responsibilities and the processes and definitions used to review and assess specific safety events.

In addition to these external committees, an internal committee of medically qualified Pfizer personnel with expertise in the assessment and diagnosis of respiratory disease will review

and categorize potential events of interstitial lung disease (Interstitial Lung Disease Committee, ILDRC).

Additional safety event adjudication or review committees may be established to harmonize and standardize selected safety event assessments. As described above, individual committee charters will provide specific descriptions of the scope of responsibilities and the processes and definitions used to review and assess specific safety events.

In addition to the event adjudication or review committees described above, all biopsies of potentially malignant tumors, suspicious lymphadenopathy, or possible extranodal lymphoproliferative disorder should be submitted to the central laboratory for review by central laboratory pathologists. In some instances, additional expert pathology review of submitted samples may be performed. Description of the scope of review and the processes used to obtain and assess biopsies is described in the Histopathology Review for Potential Malignancy charter.

8. REFERENCES

1. Schoenfeld, D. A., & Tsiatis, A. A. (1987). A modified log rank test for highly stratified data. Biometrika, 74(1), 167-175.

2. Shih, J. H. (1999). A class of permutation tests for stratified survival data. Biometrics, 55(4), 1156-1161.

3. Yang 2012 (2012). The effects of overstratification on the stratified log rank test for survival analysis; http://d-scholarship.pitt.edu/11951/1/SY_ETD_04202012.pdf

4. Agresti A. (2002). Categorical Data Analysis, 2nd Edition, New York: John Wiley & Sons.

5. Daly LB. (1992). Simple SAS macros for the calculation of exact binomial and Poisson confidence limits. Comp. Biol. Med. 22: 351-361.

6. Brunner HI, Lovell DJ, Finck BK, Giannini EH. Preliminary definition of disease flare in juvenile rheumatoid arthritis. J Rheumatol 2002; 29(5):1058-64.

7. Giannini EH, Ruperto N, Ravelli A, Lovell DJ, Felson DT, Martini A. Preliminary definition of improvement in juvenile arthritis. Arthritis Rheum 1997; 40: 1202-1209.

8. Consolaro A, Ruperto N, et al. Development and validation of a composite disease activity score for juvenile idiopathic arthritis. Arthritis Care and Research, 2009; 61(5):658-666.

9. Bazso A, Consolaro A, et al. Development and testing of reduced joint counts in juvenile idiopathic arthritis. J Rheumatol, 2009; 36:183 -90.

10. Nordal EB, Zak M, et al. Validity and predictive ability of the juvenile arthritis disease activity score based on CRP versus ESR in a Nordic population based setting. Ann Rheum Dis, 2012; 71:1122 -7..

11. Consolaro A, Bracciolini G, et al. Remission, minimal disease activity and acceptable symptom state in juvenile idiopathic arthritis. Arthritis Rheum, 2012; 64:2366 -74.

12. Consolaro A, Giancane, G, Schiappapietra, B. et al. Clinical outcome measures in juvenile idiopathic arthritis. Pediatr Rheumatol 14, 23 (2016). https://doi.org/10.1186/s12969-016-0085-5

13. Wallace CA, Giannini EH, Huang B, Itert L, Ruperto N for the Childhood Arthritis and Rheumatology Research Alliance (CARRA), the Pediatric Rheumatology Collaborative Study Group (PRCSG), and the Pediatric Rheumatology InterNational Trials Organisation (PRINTO). American College of Rheumatology provisional criteria for defining clinically inactive disease in select categories of juvenile idiopathic arthritis. Arthritis Care Res 2011;63;929-936.

14. Landgraf JM, Abetz L Ware JE. In:Child-Health Questionnaire(CHQ): a user's manual. Boston, MA: Health Institute, New England Medical Center; 1996: 34-52, 90-95.

15. Singh G, Athreya BH, Fries JF, Goldsmith DP. Measurement of health status in children with juvenile rheumatoid arthritis. Arthritis Rheum 1994; 37:1761–1769.

16. Ruperto N, Ravelli A, Pistorio A, Malattia C, Cavuto S, Gado-West L, Tortorelli A, Landgraf JM, Singh G, Martini A, for the Paediatric Rheumatology International Trials Organisation (PRINTO). Cross-cultural adaptation and psychometric evaluation of the Childhood Health Assessment Questionnaire (CHAQ) and the Child Health Questionnaire (CHQ) in 32 countries. Review of the general methodology. Clin Exp Rheumatol. 2001; 19 (Suppl. 23): S1-9.

17. Brinkhof MW, Spycher BD, Yiannoutsos C, Weigel R, Wood R, Messou E, Boulle A, Egger M, & Sterne JA (2010). Adjusting mortality for loss to follow-up: analysis of five ART programmes in sub-Saharan Africa. PloS one, 5(11), e14149. doi:10.1371/journal.pone.0014149

18. Atkinson A, Kenward MG, Clayton T, Carpenter JR (2019). Reference-based sensitivity analysis for time-to-event data. *Pharmaceutical Statistics,* 18(6): 645-658.

19. Wilson, E.B., Hilferty, M.M. (1931). "The distribution of chi-squared." Proceedings of the National Academy of Sciences, Washington, 17:684-688.

20. Moscovici, J.L., Ratitch, B. (2017). "Combining survival analysis results after multiple imputation of censored event times." PharmaSUG 2017 – Paper SP05.

21. Ratitch, B., Lipkovich, I., O'Kelly, M. (2013). "Combining analysis results from multiply imputed categorical data." PharmaSUG 2013 – Paper SP03.

22. Rubin DB. (1987). Multiple imputation for nonresponse in surveys. New York: Wiley.

9. APPENDICES

Appendix 1. SUMMARY OF EFFICACY ANALYSES

Table 4. Summary of Efficacy Analyses

Endpoint	Estimand and Analysis Type	Population	Data Inclusion and Rules for Handling Intercurrent Events and Missing Data	Analysis Model
Time to disease flare in DB phase	Primary estimand Estimand 1 Composite Strategy Primary analysis	DBFAS	Investigational product discontinuation due to study end and end of study treatment after clinical remission are treated as non-flare and censored. Investigational product discontinuation for any other reason are treated as disease flare. Missing data are not imputed, assuming censoring at random.	Unstratified log-rank test (primary), KM Cox Proportional Hazard Ratio (supportive)
	Primary estimand Estimand 1 Composite Strategy Sensitivity analysis	DBFAS	Investigational product discontinuation due to study end and end of study treatment after clinical remission are treated as non-flare and censored. Investigational product discontinuation for any other reason are treated as disease flare. Missing data will be imputed, assuming censoring not at random.	Tipping point analysis, unstratified log-rank test
	Supplementary estimand Estimand 4 Hypothetical Strategy Supplementary analysis	DBFAS	Investigational product discontinuation for reasons other than disease flare are treated as non-flare and censored. Missing data are not imputed, assuming censoring at random.	Unstratified log-rank test, KM Cox Proportional Hazard Ratio

Table 4. Summary of Efficacy Analyses

Endpoint	Estimand and Analysis Type	Population	Data Inclusion and Rules for Handling Intercurrent Events and Missing Data	Analysis Model
	Supplementary estimand Estimand 5 Treatment Policy Strategy Supplementary analysis	DBFAS	All data collected including those through Week 52 post randomization from study A3921165 and flare data from study A3921145 and will be included regardless of intercurrent events. Missing data are not imputed, assuming censoring at random.	Unstratified log-rank test, KM Cox Proportional Hazard Ratio
	Supplementary estimand Estimand 5 Treatment Policy Strategy Sensitivity analysis	DBFAS	All data collected including those through Week 52 post randomization from study A3921165 and flare data from study A3921145 and will be included regardless of intercurrent events. Missing data will be imputed, assuming censoring not at random.	Tipping point analysis, unstratified log-rank test
Time to disease flare in DB phase by subgroups: age, body weight, geographic region, formulation, OL baseline oral corticosteroids use, DB baseline oral corticosteroid use, OL baseline MTX use, prior bDMARD experience	Primary estimand Estimand 1 Composite Strategy Subset analysis	DBFAS	Investigational product discontinuation due to study end and end of study treatment after clinical remission are treated as non-flare and censored. Investigational product discontinuation for any other reason are treated as disease flare. Missing data are not imputed, assuming censoring at random.	KM, Cox Proportional Hazard Ratio
Occurrence of disease flares by DB visit	Primary estimand Estimand 1 Composite Strategy	DBFAS	Investigational product discontinuation due to study end and end of study treatment after clinical remission are treated as non-flare and censored. Investigational product	KM

Table 4. Summary of Efficacy Analyses

Endpoint	Estimand and Analysis Type	Population	Data Inclusion and Rules for Handling Intercurrent Events and Missing Data	Analysis Model
	Secondary analysis		discontinuation for any other reason are treated as disease flare. Missing data are not imputed, assuming censoring at random.	
	Supplementary estimand Estimand 4 Hypothetical Strategy Supplementary analysis	DBFAS	Investigational product discontinuation for reasons other than disease flare are treated as non-flare and censored. Missing data are not imputed, assuming censoring at random.	KM
	Supplementary estimand Estimand 5 Treatment Policy Strategy Supplementary analysis	DBFAS	All data collected including those through Week 52 post randomization from study A3921165 and flare data from study A3921145 and will be included regardless of intercurrent events. Missing data are not imputed, assuming censoring at random.	KM
Occurrence of disease flares by DB visit by subgroups: age, body weight, geographic region, formulation, OL baseline oral corticosteroids use, DB baseline oral corticosteroid use, OL baseline MTX use, prior bDMARD experience	Primary estimand Estimand 1 Composite Strategy Subset analysis	DBFAS	Investigational product discontinuation due to study end and end of study treatment after clinical remission are treated as non-flare and censored. Investigational product discontinuation for any other reason are treated as disease flare.	KM

Table 4. Summary of Efficacy Analyses

Endpoint	Estimand and Analysis Type	Population	Data Inclusion and Rules for Handling Intercurrent Events and Missing Data	Analysis Model
			Missing data are not imputed, assuming censoring at random.	
Adapted JIA ACR responses by DB visit	Secondary estimand Estimand 2 Composite Strategy Secondary analysis	DBFAS	For investigational product discontinuation due to study end and end of study treatment after clinical remission, use last ACR value before discontinuation. Investigational product discontinuation for any other reason are treated as non-response. Missing responses are treated as non-response.	Normal approximation
	Estimand 6 Treatment Policy Strategy Supplementary analysis	DBFAS	The intercurrent event of investigational product discontinuation will not be considered. All the data collected while On- and Off-investigational product until Week 52 will be used to derive the endpoints. Missing responses are treated as non-response.	Normal approximation

Table 4. Summary of Efficacy Analyses

Endpoint	Estimand and Analysis Type	Population	Data Inclusion and Rules for Handling Intercurrent Events and Missing Data	Analysis Model
JADAS-27 ESR inactive disease by DB visit	Secondary estimand Estimand 2 Composite Strategy Secondary analysis	DBFAS	For investigational product discontinuation due to study end and end of study treatment after clinical remission, use last JADAS-27 ESR value before discontinuation. Investigational product discontinuation for any other reason are treated as active disease. Missing responses are treated as active disease.	Normal approximation
JADAS-27 ESR minimal disease activity by DB visit	Secondary estimand Estimand 2 Composite Strategy Secondary analysis	DBFAS	For investigational product discontinuation due to study end and end of study treatment after clinical remission, use last JADAS-27 ESR value before discontinuation. Investigational product discontinuation for any other reason are treated as non-response for minimal disease activity. Missing responses are treated as non-response for minimal disease activity.	Normal approximation
JADAS-27 CRP inactive disease by DB visit	Secondary estimand Estimand 2 Composite Strategy Secondary analysis	DBFAS	For investigational product discontinuation due to study end and end of study treatment after clinical remission, use last JADAS-27 CRP value before discontinuation. Investigational product discontinuation for any other reason are treated as active disease. Missing responses are treated as active disease.	Normal approximation

Table 4. Summary of Efficacy Analyses

Endpoint	Estimand and Analysis Type	Population	Data Inclusion and Rules for Handling Intercurrent Events and Missing Data	Analysis Model
JADAS-27 CRP minimal disease activity by DB visit	Secondary estimand Estimand 2 Composite Strategy Secondary analysis	DBFAS	For investigational product discontinuation due to study end and end of study treatment after clinical remission, use last JADAS-27 CRP value before discontinuation. Investigational product discontinuation for any other reason are treated as non-response for minimal disease activity. Missing responses are treated as non-response for minimal disease activity.	Normal approximation
Absence of fever by DB visit	Secondary estimand Estimand 2 Composite Strategy Secondary analysis	DBFAS	For investigational product discontinuation due to study end and end of study treatment after clinical remission, use last oral temperature value before discontinuation. Investigational product discontinuation for any other reason are treated as fever. Missing responses are treated as fever.	Normal approximation
JIA ACR inactive disease by DB visit	Secondary estimand Estimand 2 Composite Strategy Secondary analysis	DBFAS	For investigational product discontinuation due to study end and end of study treatment after clinical remission, use last JIA ACR value before discontinuation. Investigational product discontinuation for any other reason are treated as non-response. Missing responses are treated as active disease.	Normal approximation
Clinical remission in DB phase	Secondary estimand	DBFAS	For investigational product discontinuation due to study end and end of study treatment after clinical remission, use last clinical	Normal approximation

Table 4. Summary of Efficacy Analyses

Endpoint	Estimand and Analysis Type	Population	Data Inclusion and Rules for Handling Intercurrent Events and Missing Data	Analysis Model
	Estimand 2 Composite Strategy Secondary analysis		remission status before discontinuation. Investigational product discontinuation for any other reason are treated as non-response for active disease. Missing responses are treated as active disease.	
ACR 30/50/70 response and JIA ACR inactive disease by DB visit and by subgroups: age, body weight, geographic region, formulation, OL baseline oral corticosteroids use, DB baseline oral corticosteroid use, OL baseline MTX use, prior bDMARD experience	Secondary estimand Estimand 2 Composite Strategy Subset analysis	DBFAS	For investigational product discontinuation due to study end and end of study treatment after clinical remission, use last JIA ACR value before discontinuation. Investigational product discontinuation for any other reason are treated as non-response/active disease. Missing responses are treated as non-response/active disease.	Normal approximation
JADAS-27 ESR scores change from DB baseline by DB visit	Estimand 7 Hypothetical Strategy Secondary analysis	DBFAS	Only data collected On-investigational product are included. Missing data are not imputed and handled using MMRM.	MMRM with terms treatment, visit, treatment*visit, DB baseline, DB baseline*visit
JADAS-27 CRP scores change from DB baseline by DB visit	Estimand 7 Hypothetical Strategy Secondary analysis	DBFAS	Only data collected On-investigational product are included. Missing data are not imputed and handled using MMRM.	MMRM with terms treatment, visit, treatment*visit, DB baseline, DB baseline*visit

Table 4. Summary of Efficacy Analyses

Endpoint	Estimand and Analysis Type	Population	Data Inclusion and Rules for Handling Intercurrent Events and Missing Data	Analysis Model
CHAQ - Discomfort Index responses change from DB baseline by DB visit	Estimand 7 Hypothetical Strategy Secondary analysis	DBFAS	Only data collected On-investigational product are included. Missing data are not imputed and handled using MMRM.	MMRM with terms treatment, visit, treatment*visit, DB baseline, DB baseline*visit
CHQ responses change from DB baseline by DB visit	Estimand 7 Hypothetical Strategy Secondary analysis	DBFAS	Only data collected On-investigational product are included. Missing data are On-investigational product not imputed and handled using MMRM.	MMRM with terms treatment, visit, treatment*visit, DB baseline, DB baseline*visit
sJIA ACR core set variables change from OL baseline by DB visit, including • number of joints with active arthritis • number of joints with limited range of motion • physician global evaluation of disease activity • parent/legal guardian/subject evaluation of overall well-being (from the CHAQ) • ESR	Secondary estimand Estimand 3 Treatment Policy Strategy Secondary analysis	DBFAS	All data collected On- and Off-investigational product are included. Missing data are not imputed and handled using MMRM.	MMRM with terms treatment, visit, treatment*visit, OL baseline, OL baseline*visit
	Estimand 7 Hypothetical Strategy Supplementary analysis	DBFAS	Only data collected On-investigational product are included. Missing data are not imputed and handled using MMRM.	MMRM with terms treatment, visit, treatment*visit, OL baseline, OL baseline*visit
CHAQ disability index change from OL baseline by DB visit	Estimand 7 Hypothetical Strategy Secondary analysis	DBFAS	Only data collected On-investigational product are included. Missing data are not imputed and handled using MMRM.	MMRM with terms treatment, visit, treatment*visit, OL baseline, OL baseline*visit

Table 4. Summary of Efficacy Analyses

Endpoint	Estimand and Analysis Type	Population	Data Inclusion and Rules for Handling Intercurrent Events and Missing Data	Analysis Model
sJIA ACR core set variables change from DB baseline by DB visit • number of joints with active arthritis • number of joints with limited range of motion • physician global evaluation of disease activity • parent/legal guardian/subject evaluation of overall well-being (from the CHAQ) • CHAQ disability index • ESR	Estimand 7 Hypothetical Strategy Secondary analysis	DBFAS	Only data collected On-investigational product are included. Missing data are not imputed and handled using MMRM.	MMRM with terms treatment, visit, treatment*visit, DB baseline, DB baseline*visit
CHAQ disability index change from DB baseline by DB visit	Secondary estimand Estimand 3 Treatment Policy Strategy Secondary analysis	DBFAS	All data collected On- and Off-investigational product are included. Missing data are not imputed and handled using MMRM.	MMRM with terms treatment, visit, treatment*visit, DB baseline, DB baseline*visit
Steroid taper success in OL Part 2	Summary	OLPT2	Only use observed data	N/A
Steroid reduction to ≤0.2 mg/kg/day or 10 mg/day in OL Part 2	Summary	OLPT2	Only use observed data	N/A
Adapted JIA ACR responses by OL Part 1 / 2 visit	Summary	OLPT1 / OLPT2	Only use observed data	N/A
JADAS-27 inactive disease CRP by OL Part 1 / 2 visit	Summary	OLPT1 / OLPT2	Only use observed data	N/A

Table 4. Summary of Efficacy Analyses

Endpoint	Estimand and Analysis Type	Population	Data Inclusion and Rules for Handling Intercurrent Events and Missing Data	Analysis Model
JADAS-27 inactive disease ESR by OL Part 1 / 2 visit	Summary	OLPT1 / OLPT2	Only use observed data	N/A
JADAS-27 minimal disease activity CRP by OL Part 1 / 2 visit	Summary	OLPT1 / OLPT2	Only use observed data	N/A
JADAS-27 minimal disease activity ESR by OL Part 1 / 2 visit	Summary	OLPT1 / OLPT2	Only use observed data	N/A
Absence of fever by OL Part 1 / 2 visit	Summary	OLPT1 / OLPT2	Only use observed data	N/A
JIA ACR inactive disease by OL Part 1 / 2 visit	Summary	OLPT1 / OLPT2	Only use observed data	N/A
JADAS-27 scores ESR (actual and change from OL baseline) by OL Part 1 / 2 visit	Summary	OLPT1 / OLPT2	Only use observed data	N/A
JADAS-27 scores CRP (actual and change from OL baseline) by OL Part 1 / 2 visit	Summary	OLPT1 / OLPT2	Only use observed data	N/A
CHAQ responses (actual and change from OL baseline) by OL Part 1 / 2 visit	Summary	OLPT1 / OLPT2	Only use observed data	N/A
CHQ responses (actual and change from OL baseline) at end of OL Part 1 / 2	Summary	OLPT1 / OLPT2	Only use observed data	N/A
JIA ACR core set variables (actual and changes from OL baseline) by OL Part 1 / 2 visit	Summary	OLPT1 / OLPT2	Only use observed data	N/A

Table 4. Summary of Efficacy Analyses

Endpoint	Estimand and Analysis Type	Population	Data Inclusion and Rules for Handling Intercurrent Events and Missing Data	Analysis Model
Fever attributed to sJIA on Day 3, 7 and 14 of OL Part 1	Summary	OLPT1	Only use observed data	N/A
Time to adapted JIA ACR30 response in OL Part 1	Summary	OLPT1	Only use observed data	N/A
CRP ≤ 10 mg/L by OL Part 1 / 2 visit	Summary	OLPT1 / OLPT2	Only use observed data	N/A

Appendix 2. EFFICACY AND FUTILITY STOPPING BOUNDARIES

Efficacy: A 2-look Gamma family spending function with gamma parameter 4 is used to determine the efficacy boundaries.

Futility (non-binding): A 2-look Gamma family spending function with gamma parameter -4 is used to determine the futility boundaries.

Assumptions:

- True hazard ratio of tofacitinib vs placebo is 0.36
- Median time to flare for placebo is of 236 days

The efficacy and futility boundaries will depend on the number of flares. Boundaries when the number of flares at First Analysis is 28 are provided in Table 2. Boundaries when the number of flares at First Analysis is 29-31 are shown below.

Table 5. Stopping Boundaries at First Analysis Expressed as P-values and Number of Flares Needed at Final Analysis in Various Scenarios

# Flares at First Analysis	28	29	30	31
Efficacy Boundary: 1-sided p-value ≤	0.0242	0.0245	0.0246	0.0247
Futility Boundary: 1-sided p-value ≥	0.1074	0.0815	0.0604	0.0506
# Flares Needed at Final Analysis	37	36	35	35
Overall Study Power (%)	80.4	80.5	80.9	81.8

Efficacy and Futility Stopping Boundaries with 29 Flares at First Analysis Expressed as Hazard Ratio, Z Scales and P-values

Efficacy Stopping Boundary				
	Number of Flares	Hazard Ratio	Z-score	P-value (1-sided)
First Analysis	29	0.481	-1.969	0.0245
Final Analysis	36	0.425	-2.570	0.0051
Futility Stopping Boundary				
	Number of Flares	Hazard Ratio	Z-score	P-value (1-sided)
First Analysis	29	0.596	-1.395	0.0815
Final Analysis	36	0.425	-2.570	0.0051

Efficacy and Futility Stopping Boundaries with 29 Flares at First Analysis Expressed as Hazard Ratio, Z Scales and P-values

Assuming a hazard ratio of 0.36 (tofacitinib vs placebo) and a median time to flare of 236 days for placebo
Randomization ratio = 1:1, type I error rate (1-sided) = 2.5%, power = 80.5%
Efficacy boundary: gamma family spending function with $\gamma = 4$
Futility boundary (non-binding): gamma family spending function with $\gamma = -4$

Efficacy and Futility Stopping Boundaries with 30 Flares at First Analysis Expressed as Hazard Ratio, Z Scales and P-values

Efficacy Stopping Boundary				
	Number of Flares	Hazard Ratio	Z-score	P-value (1-sided)
First Analysis	30	0.488	-1.966	0.0246
Final Analysis	35	0.425	-2.534	0.0056
Futility Stopping Boundary				
	Number of Flares	Hazard Ratio	Z-score	P-value (1-sided)
First Analysis	30	0.567	-1.552	0.0604
Final Analysis	35	0.425	-2.534	0.0056

Assuming a hazard ratio of 0.36 (tofacitinib vs placebo) and a median time to flare of 236 days for placebo
Randomization ratio = 1:1, type I error rate (1-sided) = 2.5%, power = 80.9%
Efficacy boundary: gamma family spending function with $\gamma = 4$
Futility boundary (non-binding): gamma family spending function with $\gamma = -4$

Efficacy and Futility Stopping Boundaries with 31 Flares at First Analysis Expressed as Hazard Ratio, Z Scales and P-values

Efficacy Stopping Boundary				
	Number of Flares	Hazard Ratio	Z-score	P-value (1-sided)
First Analysis	31	0.494	-1.965	0.0247
Final Analysis	35	0.429	-2.503	0.0062
Futility Stopping Boundary				
	Number of Flares	Hazard Ratio	Z-score	P-value (1-sided)
First Analysis	31	0.555	-1.639	0.0506
Final Analysis	35	0.429	-2.503	0.0062

Assuming a hazard ratio of 0.36 (tofacitinib vs placebo) and a median time to flare of 236 days for placebo
Randomization ratio = 1:1, type I error rate (1-sided) = 2.5%, power = 81.8%
Efficacy boundary: gamma family spending function with $\gamma = 4$
Futility boundary (non-binding): gamma family spending function with $\gamma = -4$

Appendix 3. DATA DERIVATION DETAILS

Appendix 3.1. The JIA Core Set Variables

The sJIA ACR response determination is a derived measurement utilizing each component of the JIA core set variables, which include:

- Number of joints with active arthritis;
- Number of joints with limited range of motion;
- Physician global evaluation of disease activity;
- Parent/legal guardian/subject evaluation of overall well-being (from the CHAQ);
- Functional ability (Disability Index from the CHAQ);
- ESR. For "real-time" assessment of the Adapted JIA ACR30 response during a study visit the subject's ESR will be used to determine the percent change in inflammation biomarker.

The ACR defines a joint with active arthritis as a joint with swelling or, in the absence of swelling, limitation of motion accompanied by pain on motion, or tenderness.

See details of the calculation of functional ability (disability index from CHAQ) in Appendix 3.9.

Appendix 3.2. The JIA Joint Counts

For swollen joints, painful/tender joints and joints with limitation of motion, respectively, before calculating the joint count, the following pre-processing step for each joint should be performed:

- If a joint receives an intra-articular injection (either at baseline or post-baseline visits), set the joint status to "Present" on or after the date of injection.
- If the site of injection is not recorded and/or cannot be identified, the joint assessment on the CRF (either "Present" or missing) will be used.

After this pre-processing, a joint assessment that is missing is considered "absent", thus a good joint. To calculate the number of joints with active arthritis and the number of joints with limited range of motion, only "present" joints will be taken into account.

Appendix 3.3. Disease Flare

The Disease Flare (Brunner 2002)[6] determination is a derived measurement utilizing each component of the JIA core set variables (Appendix 3.1).

Flare is defined as a worsening of 30% or more in 3 or more of the 6 variables of the JIA core set, with no more than one variable improving by 30% or more. In addition, for sJIA a Disease Flare may also constitute a recurrence of fever.

sJIA Flare is defined as at least one of the following:

- Recurrence of fever (oral temperature >38°C/100.4°F) on 2 or more consecutive days considered to be due to sJIA activity.

- Worsening of 30% or more in three or more of the six variables of the JIA core set with no more than one variable of the JIA core set improving by 30% compared to the day of randomization into the withdrawal phase.*

*Note: For "real-time" assessment of sJIA flares during a study visit the subject's ESR will be used to determine the JIA core component related to percent change in inflammation biomarker.

For evaluation of sJIA flare the following applies: If the Physician or parent/legal guardian global scores are used in the definition of flare, then there must be an increase of at least 2 units on a 0-10 VAS scale. If the number of active joints or number of joints with loss of motion is used in the definition of flare, then there must be an increase of at least 2 joints. If the ESR is used in the definition of flare, then the value at the visit in which flare is being assessed must be out of the normal range (ESR >20 mm/hr).

Investigators will be required to complete relevant CRF pages regarding sJIA Disease Status while the subject is at the clinic, and email or fax the sJIA Disease Assessment worksheets to the Coordinating Center for confirmation of the sJIA flare based on ESR. The confirmation whether a subject has a sJIA Flare per protocol will need to be entered in the eCRF upon receipt of the assessment from the Coordinating Center. At each visit in the double-blind phase, the investigator will evaluate the subject for evidence of disease flare. Subjects who experience a single episode of disease flare in the double-blind phase (based on the investigator's and parent/legal guardian's assessment of components of the JIA core set variables), will be discontinued from the study. Refer to Appendix 3.12 for detailed disease flare information collected on CRF pages.

For the real-time calculation of disease flare, ESR will be used by the CCC. This CCC calculation will be used in final analysis of disease flare.

Appendix 3.4. The Adapted JIA ACR30-50-70-90-100 Response

The disease activity of sJIA will be measured using an adaptation of the JIA American College of Rheumatology (ACR) 30 response (Giannini 1997)[7]. The Adapted JIA ACR30 Response criteria include the components of the pediatric ACR core set (Appendix 3.1) along with the absence of fever, defined as temperature ≤38°C/100.4°F in the preceding 7 days. The Adapted JIA ACR30, 50, 70, 90, 100 response is defined as the absence of fever due to sJIA (oral temperature ≤38°C/100.4°F) in the preceding 7 days along with at least 3 out of 6 JIA core set variables improved ≥30%, 50%, 70%, 90%, 100%, respectively, with no more

than 1 out of 6 JIA core set variables worsened by >30%. Only ESR will be used to calculate Adapted JIA ACR responses.

The Adapted JIA ACR30, 50, 70, 90, 100 response will be determined at every visit from Day 7 onward based on the investigator's and parent/legal guardian's assessment of components of the JIA core set variables.

If the value in any of the components at a timepoint is missing, the component variables that are not missing will be used to determine the response status. As a general principle, if there are sufficient non-missing components to determine whether the JIA ACR endpoint is a response or non-response, then JIA ACR endpoint is not missing, else if the available non-missing components are not sufficient to determine the response status of JIA ACR endpoint then it is considered missing.

In order to avoid numerical difficulty, if the baseline value of any component is equal to 0, the following algorithm will be used in evaluating the percent change from baseline:

1. If change from baseline is also equal to 0, then percent change from baseline is set to be 0%;

2. If change from baseline is > 0, then percent change from baseline is set to be 999999%.

These percentages will be used to derive the JIA ACR endpoints. Change from baseline cannot be < 0 since none of the components should have negative value.

Appendix 3.5. JADAS-27 CRP and JADAS-27 ESR

JADAS-27 CRP (JADAS-27 ESR) score will be determined based on four components (Consolaro 2009, Bazso 2009, Nordal 2012) [8,9,10]:

- Physician global assessment of disease activity;
- Parent/legal guardian/subject global assessment of well-being (from the CHAQ);
- Number of joints with active disease (27 joint assessment);
- CRP or ESR as applicable.

The 27 joints used here are bilateral Elbow, Wrist, MCP 1-3, PIP 1-5, Hip, Knee and Ankle, plus Cervical Spine.

The CRP value will be normalized to a 0 – 10 scale according to the following formula:

$$(\text{CRP (mg/L)} - 2.87) / 10$$

Before calculation, CRP values < 2.87 mg/L will be converted to 2.87 and CRP values >102.87 mg/L will be converted to 102.87. The value 2.87 (mg/L) here is the upper bound of the reference range provided by the lab.

The ESR value will be normalized to 0 – 10 scale as below:

$$(\text{ESR} - 20) / 10$$

Before calculation, ESR values < 20 mm/h will be converted to 20 and ESR values >120 mm/h will be converted to 120.

JADAS-27 CRP/ESR score will be calculated as the simple linear sum of the scores of its 4 components, which yields a global score of 0 – 57 for the JADAS-27 CRP/ESR score. If any component is missing, the JADAS-27 CRP/ESR score will be set to missing.

Appendix 3.6. JADAS-27 High Disease Activity, Moderate Disease Activity, Low Disease Activity, Minimum Disease Activity and Inactive Disease

The cutoff values in the JADAS-27 that correspond to inactive disease and minimal disease activity (Consolaro 2012, Consolaro 2016)[11, 12] are defined as follows:

Polyarthritis (>4 active joints) at baseline:

- Minimal Disease Activity: ≤3.8.
  - Inactive Disease: ≤1.0.
  - Low Disease Activity: >1.0 – ≤3.8.
- Moderate Disease Activity: >3.8 – ≤8.5.
- High Disease Activity: >8.5.

Oligoarthritis (≤4 active joints) at baseline:

- Minimal Disease Activity: ≤2.
  - Inactive Disease: ≤1.0.
  - Low Disease Activity: >1.0 – ≤2.0.
- Moderate Disease Activity: >2.0 – ≤4.2.
- High Disease Activity: >4.2.

For JADAS-27 inactive disease and minimal/low/moderate/high disease activity for systemic JIA subjects with >4 active joints at the OL baseline visit, the cutoff values of polyarthritis will be used. For JADAS-27 inactive disease and minimal/low/moderate/high disease activity for systemic JIA subjects with ≤ 4 active joints at the baseline visit, the cutoff values of oligoarthritis will be used.

Note that both JADAS-27 CRP and JADAS-27 ESR score will be used as cut-off values, so there will be two sets of JADAS minimum disease activity and inactive disease, one uses JADAS-27 CRP and the other uses JADAS-27 ESR.

Appendix 3.7. Presence of JIA ACR Inactive Disease and Clinical Remission

The American College of Rheumatology (ACR) Clinical Inactive Disease and Clinical Remission (Wallace 2011)[13] criteria is defined as follows:

Clinical Inactive Disease:

- No joints with active arthritis;
- No fever, rash, serositis, splenomegaly, hepatomegaly, or generalized lymphadenopathy attributable to sJIA;
- No active uveitis (as defined by the SUN Working Group);
- Normal ESR* (within normal limits of the method used where tested) or, if elevated, not attributable to JIA;
- Physician global assessment of disease activity score of 'best possible' on the scale used;
- Duration of morning stiffness of ≤15 minutes;

* The subject's ESR result will be used for 'real-time' assessment. The normal range of ESR is 0 – 20 mm/hr.

If any of the components is missing, the components that are not missing will be used to determine the Clinical Inactive Disease Status as follows:

- Inactive disease: all components must be present (non-missing) and meet all conditions.
- Active disease: at least one component is present (non-missing) and does not meet the condition.
- Not determined: any other scenarios.

Clinical Remission:

- Clinical inactive disease for 24 weeks continuously while on medications.

Appendix 3.8. CHQ Responses

The CHQ (Landgraf 1996)[14] is a validated general pediatric quality of life instrument. The CHQ assesses for 14 physical and psychosocial domains: general health perceptions, physical functioning, role/social physical functioning, bodily pain, role/social emotional functioning, role/social behavioral functioning, parent impact-time, parent impact-emotional, self-esteem, mental health, behavior, family activities, family cohesion, and change in health.

The response options for the CHQ are ordinal scales that vary by the item. Each item consists of 4–6 response options. Additionally, each scale consists of varying numbers of items.

Appendix 3.9. CHAQ Responses

The CHAQ, derived from the adult Health Assessment Questionnaire, comprises two indices, Disability and Discomfort, and parent global assessment of Overall Well Being (Singh 1994, Ruperto 2001)[15,16].

Disability Index

The parent/legal guardian/subject will be asked to provide responses to questions designed to assess function in 8 areas, including 1. dressing and grooming, 2. arising, 3. eating, 4. walking, 5. hygiene, 6. reach, 7. grip, and 8. activities-distributed, among a total of 30 items. Each question is rated on a four-point scale of difficulty in performance, scored from 0-3. The question with the highest score determines the score for the functional area. If aids or devices are used or assistance is required, the minimum score for that functional area is 2.

Each question is rated on a four-point ordinary scale with 0 for “no difficulty”, 1 for “some difficulties”, 2 for “much difficulties”, and 3 for “unable to do”. The eight areas of the CHAQ are averaged to calculated the CHAQ disability index which ranges from 0 (no or minimal physical dysfunction) to 3 (very severe physical dysfunction). A subject must have a score for at least six of the eight categories, otherwise a CHAQ-DI score cannot be validly computed.

When a subject uses aids or devices (eg, crutches, jar openers, etc.), then they answer the question based on their usual equipment or way of performing the activity. If they have no difficulty with a sub-category item when using aids/devices, then they mark the “no difficulty” column. In scoring, the use of aids/devices results in an adjusted score for that item.

Algorithm:

1. Sum the 8 functional area scores by using the highest score from each functional area.

2. Adjust for use of aids/device and/or help from another person when indicated. If aid/device is used, adjust a 0 or 1 to 2, 2 or 3 remains unchanged. The below identifies the aid/device companion variable for each CHAQ-DI functional area.

3. Divide the summed category scores by the number of categories answered (must be a minimum of 6) to obtain a CHAQ-DI score of 0-3 (3=worst functioning).

Functional area	Corresponding Devices
Dressing and grooming	Devices used for dressing (button hook, zipper pull, long handled shoe horn, etc.)
Arising	Special or built-up chair
Eating	Built up pencil or special utensils
Walking	Cane, walker, crutches, wheelchair
Hygiene	Raised toilet seat, bathtub seat, bathtub bar, long handled appliances in bathroom
Reach	Long handled appliances for reach
Grip	Jar opener (for jars previously opened)

Notes about Aids/Devices:

The assignment of devices to particular disability categories assumes that the devices are used only for their intended purposes, eg, when a subject indicates that they use a cane, it is presumed that the cane is used as an aid in walking. However, it is possible that subject uses the cane as an aid in performing other activities. For example, the subject may check off the cane listed at the bottom of the page (or write "cane" under the "other" option) and then write a little note in the margin stating that the cane is also used on a regular basis to help them rise out of a chair and to rise off of the toilet. In such a case, the variables should be scored a "1" to reflect the subject's use of a cane in these three areas of functioning. If unsure whether the subject is using one of the devices specified above for the purpose for which it is designed, the subject should be called for clarification.

When there are devices entered in the "Other" Section or notes written next to an item, they are considered as being used for any of the stated categories. Permanent adaptations of the person's environment (eg, changing faucets in the bathroom or kitchen, using fabric hook and loop closures on clothing) should also be counted as aids/devices.

In our experience, few subjects have reported "other" items, and when they have, it has usually been either a duplicate of an aid/device already listed, or they have listed something that does not count (eg, a wrist splint). Thus, it is usually acceptable to exclude the "Other" option, if desired.

Discomfort Index

For the assessment of discomfort, the parent/legal guardian/subject will be asked to provide a response to the following question:

"How much pain do you think your child had because of his or her illness IN THE PAST WEEK? Please fill a circle below to indicate how severe your child's pain has been:"

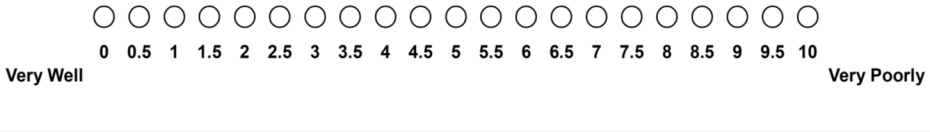
The parent/legal guardian/subject will rate the overall pain by entering a number from 0 to 10 (in 0.5 increments), with '0' as 'No Pain' and '10' as 'Very Severe Pain' on a 21-circle visual analog scale (VAS), as shown below.

Parent/Legal Guardian/Subject Global Assessment of Overall Well-Being

For the assessment of overall well-being, the parent/legal guardian/subject will be asked to provide a response to the following question:

"Considering all the ways in which the illness affects your child AT THIS TIME, please indicate below how your child is doing by filling a circle."

The parent/or legal guardian/subject will rate the overall well-being by entering a number from 0 to 10 (in 0.5 increments), with '0' as 'Very Well' and '10' as 'Very Poorly' on a 21-circle visual analog scale (VAS), as shown below.



Appendix 3.10. Body Weight and Height Standardized Z-Scores

For assessment of growth as measured by body weight and height, a table of descriptive statistics for body weight and height standardized Z-scores (actual values and changes from baseline) at each visit will be constructed by age group (2 to <6 years, 6 to <12 years, and 12 to <18 years) for males and females separately.

Height (weight) Z-score is calculated using the following formula:

$$Z = \frac{(X/M)^L - 1}{LS}, L \neq 0$$

Or

$$Z = \ln(X/M)/S, L = 0$$

where X is the physical measurement (e.g., height, weight) and L, M, and S are the values from standardized growth charts provided by the US Centers for Disease Control website (http://www.cdc.gov/GrowthCharts/) by gender for children from birth to 240 months. The Z-score cannot be calculated if age >240 months.

Appendix 3.11. Others

Methotrexate background therapy: WHO-Drug Preferred Term = "Methotrexate" or "Methotrexate sodium" as contra-indicated therapy will be considered.

Prior biologic failures: If the biologic is discontinued due to one of the following 4 reasons, the subject has a biologic failure: 1. adverse event, 2. intolerability, 3. lack of efficacy, failure to achieve a response, 4. Both AE/LOE, failure to achieve a response. Biologics includes Actemra, Cimzia, Enbrel, Humira, Kineret, Orencia, Remicade, Rituxan, Simponi, Ilaris.

History of uveitis: It will be identified by MedDRA preferred term as medical history (or AE) reported.

Appendix 3.12. Investigational Product Discontinuation due to Disease Flare

To look for subjects who discontinue investigational product due to disease flare, the following CRF forms should be checked:

1. "End of treatment (EOT) double-blind period" form: "Subject Status" question checked as "withdrawn during double-blind treatment period"; "Reason for discontinuation of treatment" question checked as "withdrawn due to insufficient clinical response"

2. "Adverse Event" form: Confirm an adverse event of "sJIA Flare" was recorded on the Adverse Event CRF

3. "Pediatric ACR Response and Disease Flare" form at the Early Termination visit: "Flare Status" checked as "Flare"

Appendix 3.13. Definition and Use of Visit Windows in Reporting

Visit windows will be used for efficacy variables, and for any safety displays that display by scheduled visits.

For an endpoint that consists of multiple components, e.g., Adapted JIA ACR30 and JADAS-27 CRP, the components should be collapsed first in the same visit window to derive the endpoint.

If two or more observations fall into the same visit window, the observation closest to the target day will be used in the analyses. If there is a tie, the later observation will be used.

For efficacy data, the visit windows will be applied to both on-treatment and on-study (both on- and off-treatment) data separately. Note: Data collected in visits on or before the last dosing day +2 days are considered on treatment data. Data collected in visits after investigational product discontinuation day + 2 days are considered off-treatment data. For safety data, only the on-study data visit window will be defined and used, except IR calculations for select adverse events.

Table 6. Visit Windows for Efficacy and Safety Endpoints

Open-Label Part 1		
Visit Label	**Target Day**	**Definition [Day window]**
Baseline	Day 1	Last non-missing assessment on or before Day 1 and prior to first dose of investigational product administration
Part 1 Day 3	Day 3	Days 2-4
Part 1 Day 7	Day 7	Days 5-9
Part 1 Week 2	Day 14	Days 10-21
Part 1 Week 4	Day 28	Days 22-42/ Part 2 Start Day
Part 1 Week 8	Day 56	Days 43-70/ Part 2 Start day
Part 1 Week 12	Day 84	Days 71-98/ Randomization Day or Part 2 Start Day
Part 1 Week 16	Day 112	Days 99-126/ Randomization Day or Part 2 Start Day
Open-Label Part 2		
Visit Label	**Target Day**	**Definition [Day window]**
Part 2 Week 4	Part 2 Start Day + 28	Part 2 Start Day + 1-42
Part 2 Week 8	Part 2 Start Day + 56	Part 2 Start Day + 43-70/ Randomization Day
Part 2 Week 12	Part 2 Start Day + 84	Part 2 Start Day + 71-98/ Randomization Day
Part 2 Week 16	Part 2 Start Day + 112	Part 2 Start Day + 99-126/ Randomization Day

Table 6. Visit Windows for Efficacy and Safety Endpoints

Part 2 Week 20	Part 2 Start Day + 140	Part 2 Start Day + 127-154/ Randomization Day
Part 2 Week 24	Part 2 Start Day + 168	Part 2 Start Day + 155-182/ Randomization Day
Double-Blind Phase		
Visit Label	**Target Day**	**Definition [Day window]**
Randomization	Double-blind Day 1 (Randomization Day)	Randomization Day
DB Week 4	Randomization Day + 28	Randomization Day + 1-42
DB Week 8	Randomization Day + 56	Randomization Day + 43-70
DB Week 12	Randomization Day + 84	Randomization Day + 71-98
DB Week 16	Randomization Day + 112	Randomization Day + 99-126
DB Week 20	Randomization Day + 140	Randomization Day + 127-154
DB Week 24	Randomization Day + 168	Randomization Day + 155-182
DB Week 28	Randomization Day + 196	Randomization Day + 183-210
DB Week 32	Randomization Day + 224	Randomization Day + 211-238
DB Week 36	Randomization Day + 252	Randomization Day + 239-266
DB Week 40	Randomization Day + 280	Randomization Day + 267-294
DB Week 44	Randomization Day + 308	Randomization Day + 295-322
DB Week 48	Randomization Day + 336	Randomization Day + 323-350
DB Week 52	Randomization Day + 364	Randomization Day + 351-378
…		

Day 1: the first day of treatment with investigational product
Part 2 Start Day: the first day of CS tapering
Randomization Day (DB Phase Day 1): the day of randomization into the withdrawal phase

Appendix 4. Dose Evaluation in the Open-Label Phase

In Part 1 of the open-label phase, dosing decisions will be made based on the evaluation of cohorts of 7 subjects treated with tofacitinib. Enrollment will be conducted in a staggered fashion. The upper limit of a 95% confidence interval for an Adapted JIA ACR 30 response will serve as guidance to evaluate the initial cohorts of 7 subjects dosed with tofacitinib 5 mg BID. Using the confidence interval, a given cohort will be deemed to have sufficient efficacy for further study if at least 2 responders out of 7 subjects are observed. The following will serve as guidance for making decisions about the cohorts:

- If ≤ 1 out of 7 responders is observed in 2 consecutive cohorts, the dose level will be discontinued.
- If ≥ 2 out of 7 responders are observed in each of 2 consecutive cohorts at the 5 mg BID dose level, efficacy at this dose level will be considered sufficient to allow further evaluation in additional subjects.
- In the event that intermediate response rates are observed (for example, 1 out of 7 in the first cohort and 2 out of 7 in the second), the decision on whether or not to evaluate the dose level in additional subjects would be based on review of the totality of the data by the Sponsor and Steering Committee and further cohort(s) may be needed to evaluate this dose.

Refer to Section 9.3 in the protocol for additional information on the confidence intervals. A Steering Committee Charter will describe the dose selection process in detail.

Appendix 5. Tipping Point Analysis for the Primary Endpoint Based on Weibull Regression

A tipping point analysis will be conducted to assess the robustness of the conclusion from the time-to-event analysis performed under Censoring At Random (CAR) assumption to different alternative Censoring Not At Random (CNAR) assumptions (ie, non-ignorable or informative censoring). In this tipping point analysis, two treatment groups are considered: tofacitinib 5 mg BID and placebo. For the data, we denote the observed event time as T_i and censoring variable as δ_i for subject i with $i = 1, \ldots, N$ where N is the total number of subjects in the DBFAS. If the event of interest occurred at T_i, $\delta_i = 0$ and if the event of interest did not occur at T_i, $\delta_i = 1$. We also let indicator variables, X_{1i} to represent treatment assignment as follows: for tofacitinib 5 mg BID, $X_{1i} = 1$ and for placebo, $X_{1i} = 0$. This assigns the placebo group as the comparator group. The Weibull survival model has been proposed for sensitivity analysis (Brinkhof et al. 2010; Atkinson et al. 2019)[17,18]. A Weibull regression model will be fitted using the following parameterization:

$$P(T_i = t_i) = \frac{k t_i^{k-1}}{\gamma_i^k} \exp\left(-\left(\frac{t_i}{\gamma_i}\right)^k\right)$$

where k is the Weibull shape parameter (same for all subjects) and γ_i is the Weibull scale parameter for subject i. The treatment indicator variable is modeled as

$$\gamma_i = \exp(\beta_0 + \beta_1 X_{1i}).$$

The survival function and hazard function of subject i are then given by

$$S_i(t) = \exp\left(-\left(\frac{t}{\gamma_i}\right)^k\right)$$

and

$$h_i(t) = \frac{k t^{k-1}}{{\gamma_i}^k}$$

respectively.

The tipping point analysis will be implemented in the following steps.

Step 1: Estimate the parameters using Maximum Likelihood method.
Fit the data (both observed and censored times) to the Weibull regression model specified above using maximum likelihood (ML) method. The ML estimates of the parameters' means and variance-covariance matrix (ML multivariate sampling distribution) will be generated. This corresponds to the assumption of CAR.

Step 2: Sample parameters from the ML sampling distributions.
Using the estimated ML multivariate sampling distribution from Step 1, we simulate (or sample) R sets of values and we denote one set for example as (β_0^*, β_1^*, k^*). For subject i, the Weibull scale parameter will be computed on this set as

$$\gamma_i^* = \exp(\beta_0^* + \beta_1^* X_{1i}).$$

This is done similarly for all R sets of sampled parameters.

Step 3: Define and apply sensitivity parameter for each of two treatment groups.
For subject i with censored time-to-event (ie, $\delta_i = 1$), we define a sensitivity parameter c which is multiplied to the subject's hazard function estimated under CAR assumption, the new hazard function $h_i^c(t)$ for time beyond the observed censored time-to-event ($t > T_i$) represents the hazard under an alternative CNAR assumption:

$$h_i^c(t) = ch_i(t) = c\frac{kt^{k-1}}{{\gamma_i}^k} = \frac{kt^{k-1}}{(\gamma_i/(c^{1/k}))^k}$$

with the sensitivity parameter in the range of $c > 0$. If $0 < c < 1$, the CNAR hazard is lower than the CAR hazard and if $c > 1$, the CNAR hazard is higher than the CAR hazard. If $c = 1$, it is the same as the CAR hazard. Therefore, specifying a range of values in $0 < c < 1$ and $c > 1$ for this sensitivity parameter will represent a series of deviations from the CAR assumption in favorable and unfavorable directions respectively for time-to-event. Different sensitivity parameter c can be specified for different treatment groups.

Step 4: Impute the time-to-event under a CNAR assumption (c) using the sampled Weibull parameters.
For subject i with censored time-to-event (ie, $\delta_i = 1$) and with a specified CNAR sensitivity parameter c, using a single set of sampled Weibull parameters (γ_i^*, k^*) in Step 2, we can reconstruct the conditional predictive distribution as follows:

$$\gamma_i^{c*} = \gamma_i^*/(c^{1/k^*})$$
$$F_i^{c*}(t_i|t_i > T_i;\ c, \gamma_i^*, k^*) = 1 - exp\left[\left(\frac{T_i}{\gamma_i^{c*}}\right)^{k^*} - \left(\frac{t}{\gamma_i^{c*}}\right)^{k^*}\right].$$

In order to impute t_i (imputed value as $t_i{}^*$), we generate one u_i^*~Uniform(0,1) for subject i, then we can solve for $t_i{}^*$,

$$t_i{}^* |\ t_i > T_i = \gamma_i^{c*}\left[\left(\frac{T_i}{\gamma_i^{c*}}\right)^{k^*} - \log(1 - u_i^*)\right]^{1/k^*}.$$

Different CNAR sensitivity parameter (c) values will give different imputed $t_i{}^*$ using the same u_i^* for this subject i.

The risk period and censoring time for imputation are defined for Sensitivity Analysis 1 and Sensitivity Analysis 2 separately:

Sensitivity Analysis 1:
The risk period for imputation is defined from the date of the randomization to the LSLV of Study A3921165, for subjects who are censored under the composite estimand.

Sensitivity Analysis 2:

1) the risk period for imputation is defined from the date of the randomization to the LSLV of Study A3921165, for subjects who are censored under the treatment policy estimand and not enrolled in the LTE study (Study A3921145);
2) the risk period for imputation is defined from the date of the randomization to the LSLV of Study A3921165 or to Randomization Day + 378 (the upper limit of the DB Week 52 visit window), whichever comes later, for subjects who are censored under the treatment policy estimand and enrolled in the LTE study (Study A3921145). "Randomization Day + 378" will include time spent post randomization in both Study A3921165 and Study A3921145. For example, if a subject is randomized on Study Day 80, then "Randomization Day + 378" is Study Day 458 and DB Phase Day 379; so imputation will be done up to max (LSLV of A3921165 – date of randomization + 1, DB Phase Day 379) for this subject enrolled in Study A3921145.

To implement the imputation constraint, if the imputed t_i^* is ≤ the day of the end of risk period for imputation, the imputed time is used as is; if the imputed t_i^* is > the day of the end of risk period for imputation, the subject will be censored at the end of risk period for imputation. This can be performed for all subjects with $\delta = 1$ and across all R sets of sampled Weibull parameters.

Step 5: Perform Rubin's rule for R complete imputed datasets.
For a given combination of sensitivity parameters specified for the treatment groups, Step 4 above will produce R sets of complete imputed datasets. Each of the complete imputed dataset will be analyzed using the unstratified log-rank test providing comparison of tofacitinib 5 mg BID versus placebo.

Each analysis above will give the Chi-squared test statistic (X) and degree of freedom (d) of a treatment comparison based on unstratified log-rank test. d is 1 if there are only two treatments in comparison across strata. The Wilson-Hilferty-transformed Chi-squared test statistic will be combined using Rubin's rules (Wilson and Hilferty, 1931; Moscovici and Ratitch, 2017; Ratitch, Lipkovich, and O'Kelly, 2013; Rubin, 1987) [19,20,21,22]. The Wilson-Hilferty transformation of the log-rank Chi-squared test statistic below, under null hypothesis of no treatment difference, will yield an asymptotic standard normal statistic (Z):

$$Z = \frac{\sqrt[3]{X/d} - (1 - 2/(9d))}{\sqrt{2/(9d)}}$$

with standard error of 1. These R sets of results will be combined via the Rubin's rules to obtain a combined Z that follows Student's t distribution according to the Rubin's rules and the upper-tailed 1-sided p-value corresponding to combined Z can be generated. The upper-tailed 1-sided p-value from the Student's t test of the Rubin's rules (Ratitch, Lipkovich, and O'Kelly, 2013)[21] will then be compared to pre-specified 2-sided efficacy boundary (Section 5.1).

Step 6: Assessment of tipping points.

Step 5 above can be repeated for a range of sensitivity parameter (c) values specified for each treatment group. The goal is to evaluate the impact of different CNAR assumptions for subjects with censored time-to-event on the conclusion. The specific implementation (ie, seeds for random number generation, number of multiple imputations (R), and values of the sensitivity parameter by treatment group) of this tipping point analysis will be specified in the statistical programming plan.

Appendix 6. List of Abbreviations

Abbreviation	Term
ACR	American College of Rheumatology
AE	adverse event
aJIA	adapted juvenile idiopathic arthritis
AST	aspartate aminotransferase
bDMARD	biologic disease modifying antirheumatic drug
BID	twice daily
CAR	censoring at random
CCC	Centralized Coordinating Center
CHAQ	Childhood Health Assessment Questionnaire
CHAQ-DI	Childhood Health Assessment Questionnaire-Disability Index
CHQ	Childhood Health Questionnaire
CI	confidence interval
CL/F	oral clearance
CNAR	censoring not at random
COVID-19	corona virus disease 2019
CRF	case report form
CRP	C-reactive protein
CS	corticosteroids
CSH	heterogeneous compound symmetry
CV	cardiovascular
DB	double-blind
DBFAS	double-blind full analysis set
DBSAS	double-blind safety analysis set
DSMB	Data Safety Monitoring Board
EAC	Endpoint Adjudication Committee
EOT	end of treatment
ESR	erythrocyte sedimentation rate
FDA	Food and Drug Administration (United States)
GI	gastrointestinal
GIPRC	Gastrointestinal Perforation Review Committee
HERC	Hepatic Event Review Committee
HR	hazard ratio
HZ	herpes zoster
ICH	International Council for Harmonisation
ILDRC	Interstitial Lung Disease Committee
IR	incidence rate
JADAS	juvenile arthritis disease activity score
JIA	juvenile idiopathic arthritis
K-M	Kaplan-Meier
LOCF	last observation carried forward
LOE	lack of efficacy
LS	least squares

Abbreviation	Term
LSLV	last subject last visit
LTE	Long term extension
MAC	Malignancy Adjudication Committee
MAS	macrophage activation syndrome
MCP	metacarpophalangeal
MedDRA	Medical Dictionary for Regulatory Activities
ML	maximum likelihood
MMRM	mixed model for repeated measures
MR	missing response
MTX	methotrexate
NR	non response
OI	opportunistic infection
OIRC	Opportunistic Infection Review Committee
OL	open-label
OLPT	open-label part
PIP	proximal interphalangeal
pcJIA	polyarticular course juvenile idiopathic arthritis
PK	pharmacokinetic(s)
RA	rheumatoid arthritis
RP	risk period
SAE	serious adverse event
SAP	statistical analysis plan
SDAC	Statistical Data Analysis Center
SE	standard error
SI	serious infection
sJIA	systemic juvenile idiopathic arthritis
SOC	system organ class
SOP	standard operating procedure
SUN	Standard Uveitis Nomenclature
TEAE	treatment-emergent adverse event
TFL	table figure listing
US	United States
VAS	visual analog scale
WHO	World Health Organization