BMC Medical Research Methodology

RESEARCH

Open Access

Developing a Bayesian hierarchical model for a prospective individual patient data meta-analysis with continuous monitoring

Danni Wu[1*], Keith S. Goldfeld[1] and Eva Petkova[1,2,3]

Abstract

Background Numerous clinical trials have been initiated to find effective treatments for COVID-19. These trials have often been initiated in regions where the pandemic has already peaked. Consequently, achieving full enrollment in a single trial might require additional COVID-19 surges in the same location over several years. This has inspired us to pool individual patient data (IPD) from ongoing, paused, prematurely-terminated, or completed randomized controlled trials (RCTs) in real-time, to find an effective treatment as quickly as possible in light of the pandemic crisis. However, pooling across trials introduces enormous uncertainties in study design (e.g., the number of RCTs and sample sizes might be unknown in advance). We sought to develop a versatile treatment efficacy assessment model that accounts for these uncertainties while allowing for continuous monitoring throughout the study using Bayesian monitoring techniques.

Methods We provide a detailed look at the challenges and solutions for model development, describing the process that used extensive simulations to enable us to finalize the analysis plan. This includes establishing prior distribution assumptions, assessing and improving model convergence under different study composition scenarios, and assessing whether we can extend the model to accommodate multi-site RCTs and evaluate heterogeneous treatment effects. In addition, we recognized that we would need to assess our model for goodness-of-fit, so we explored an approach that used posterior predictive checking. Lastly, given the urgency of the research in the context of evolving pandemic, we were committed to frequent monitoring of the data to assess efficacy, and we set Bayesian monitoring rules calibrated for type 1 error rate and power.

Results The primary outcome is an 11-point ordinal scale. We present the operating characteristics of the proposed cumulative proportional odds model for estimating treatment effectiveness. The model can estimate the treatment's effect under enormous uncertainties in study design. We investigate to what degree the proportional odds assumption has to be violated to render the model inaccurate. We demonstrate the flexibility of a Bayesian monitoring approach by performing frequent interim analyses without increasing the probability of erroneous conclusions.

Conclusion This paper describes a translatable framework using simulation to support the design of prospective IPD meta-analyses.

Keywords Bayesian hierarchical models, Bayesian adaptive trial design, Bayesian simulation, International consortium for data sharing, Prospective individual patient data meta-analysis, COVID-19

*Correspondence:
Danni Wu
Danni.Wu@nyulangone.org
Full list of author information is available at the end of the article

Background

With its rapid spread, mutation, and unpredictable local outbreaks, COVID-19 continues to pose a global threat [1, 2]. Since early 2020, COVID-19 has led to over six hundred million cases and over six million deaths [3]. To date, there are still only a few reliable treatments, especially for the new viral variants, for which some existing medications are less or not effective at all [4]. To alleviate this emergency, researchers have initiated numerous clinical trials to find treatments. To date, over 8,000 trials have been launched worldwide for COVID-19, including over 4,000 interventional studies and randomized controlled trials (RCTs) [5]. As the pandemic surged and waned in various parts of the world, these trials were often launched in local settings after the pandemic had peaked in those regions [6]. As a consequence of a decline in new COVID-19 cases, some RCTs were terminated early [7–10], and with a lack of sufficient numbers of patients, many reported inconclusive findings. Other trials opted to pause and wait for the second and third COVID-19 surge in their regions, thus delaying conclusive findings [11]. Without a steady enrollment of trial patients, determining the efficacy and safety of treatments could require additional COVID-19 surges at the same location over several years [6].

Pooling data from ongoing and terminated RCTs holds promise for identifying effective treatments quickly [6, 12]. While conventional meta-analyses pool data from completed RCTs [13, 14], another productive approach involves synthesizing evidence from a collection of both ongoing and completed RCTs [15]. Researchers at New York University (NYU) initiated the COntinuous Monitoring of Pooled International Trials of ConvaLEscent Plasma for COVID-19 Hospitalized Patients (COMPILE) project [16]. The COMPILE study aimed to pool individual patient data (IPD) from ongoing, paused, prematurely-terminated, or completed RCTs assessing the efficacy of COVID-19 convalescent plasma (CCP) in hospitalized patients not on mechanical ventilation at the time of randomization. The goal was to engage RCTs from around the world and to monitor the continuously accumulating data for compelling evidence of CCP's effect in order to obtain answers as soon as possible [17]. Unlike more traditional meta-analyses [18], COMPILE was designed in the presence of considerable uncertainties. The number of RCTs, number of sites within each RCT, number of RCTs within different control conditions, number of patients, the COMPILE study duration, frequency, and timing of interim looks were all unknown. These uncertainties made it difficult to use off-the-shelf statistical methods for analysis and monitoring.

Goldfeld et al. [19] proposed a statistical analysis and interim monitoring plan specifically for the COMPILE study, which included the analytic models and the rules for stopping the study. The actual data analysis of the COMPILE study was published in [17]. In addition, we used patient data from the COMPILE study to study the heterogeneous treatment effects of CCP in [20]. The current paper describes the development process of the analytic models and stopping rules presented in [19] and applied in the COMPILE study. In doing so, we provide a general, translatable framework for developing analytic models and monitoring plans for prospective IPD meta-analyses under uncertainty, which, we believe might become a common practice at least in the face of new pandemics.

Our framework is based on a Bayesian clinical trial paradigm [21, 22]. We use Bayesian hierarchical modeling [23] that explicitly accounts for patient heterogeneity and allows for "borrowing of information" across the trials [24]. This approach enables the implementation of complex statistical models with a variety of hierarchical assumptions (e.g., the number of RCTs, sites, and control conditions) and the reduction of the uncertainty of parameter estimates. Additionally, properly designed Bayesian monitoring allows continuous monitoring without the penalties for multiple interim looks and alpha-spending associated with the frequentist monitoring approach, making it an attractive and efficient strategy [24–27]. The proposed framework is not limited to the COMPILE study but is a translatable tool to guide the design of prospective IPD meta-analyses.

We organize the paper as follows. In the Methods section, we discuss methods of building a Bayesian model for quantifying a treatment's efficacy. We describe how we extended the Bayesian model to accommodate multi-site RCTs and evaluate heterogeneous treatment effects. We outline criteria for selecting prior distribution assumptions and solutions to improve model stability. We describe the proposed goodness-of-fit methodology using posterior predictive checking. We introduce the Bayesian monitoring rules calibrated for type 1 error rate and power. In the Results section, we present our process based on extensive simulations that enabled us to finalize our analysis plan, with an eye towards understanding the impact of the prior distribution assumptions on both posterior estimations and model stability. In addition, we show the simulation results of the proposed goodness-of-fit methodology and provide details about how we set the stopping rules using type 1 error rate and power. In the Discussion and conclusion section, we conclude with a brief discussion and offer a glimpse into the future applications of our work.

Methods

The first step in developing an analytic plan is to determine the functional form of the model, which is largely determined by the study's primary outcome (e.g., a binary outcome suggests a logistic regression model), but also depends on the nature of the intervention as well as the hierarchical structure suggested by the study design. We started with the simplest assumptions that there would be K RCTs in the prospective meta-analysis and each RCT would be conducted in a single site. We also assumed that each of the K RCTs would have n_k subjects, $k = 1, \ldots, K$.

General model for estimating across-study treatment effect

Conceptualizing this project, we started with the most general model that would allow us to estimate RCT-specific treatment effects along with RCT-specific random effects or intercepts while adjusting for patient-level covariates in case there were substantial differences across RCTs. In this model, Y_{ki} denotes the outcome for the i^{th} patient from the k^{th} RCT. Z_{ki} indicates the treatment assignment for the i^{th} subject in the k^{th} RCT; $Z_{ki} = 1$ if the patient is randomized to the experimental treatment arm, $Z_{ki} = 0$ if the patient is randomized to the control treatment arm. $\boldsymbol{X}_{ki}$ denotes a vector of covariates of length p. The expected value of the outcome is related to the linear combination of covariates and treatment assignment via a known link function $g(.)$:

$$\begin{aligned} g(E(Y_{ki})) =& \tau_k + \boldsymbol{\beta}\boldsymbol{X}_{ki} + \theta_k Z_{ki} \\ \tau_k \sim& \text{Normal}\ (\mu = 0, \sigma = \sigma_\tau) \\ \boldsymbol{\beta} \sim& \text{Normal}\left(\boldsymbol{\mu} = \mathbf{0}, \Sigma = \sigma_\beta^2 I_{p\times p}\right) \\ \theta_k \sim& \text{Normal}\ (\mu = \Theta, \sigma = \sigma_\theta) \\ \Theta \sim& \text{Normal}\ (\mu = 0, \sigma = \sigma_\Theta), \end{aligned} \tag{1}$$

where τ_k represents the RCT-specific intercept, $\boldsymbol{\beta}$ is a vector of coefficients for the p covariates, and θ_k is the main effect of experimental treatment for the k^{th} RCT. To avoid potential estimation problems related to small final and/or interim sample sizes in individual participating RCTs, the covariate effects $\boldsymbol{\beta}$ were not modelled as RCT-specific, unlike the treatment effect θ_k and the study-specific intercept τ_k. We were agnostic at this point as to whether the prior distributions would be *Normal*, *Cauchy*, or a $t_{student}$ distribution with 3 degrees of freedom, which is a compromise between the *Normal* and *Cauchy* distributions. The variance assumptions for the prior distributions were also an open-ended target. The prior distributions for all parameters and hyperparameters were determined in later models.

A key feature of the model is that the prior distribution assumes each RCT-specific "experimental treatment effect" θ_k is centered around an overall effect of the experimental treatment Θ. Θ represents the treatment-control effect contrast: $E(g(E(Y_{ki}|\boldsymbol{X}_{ki}, Z_{ki} = 1)) - g(E(Y_{ki}|\boldsymbol{X}_{ki}, Z_{ki} = 0)))$ across all RCTs. This treatment-control effect contrast is the parameter of interest in clinical trials. In the case that the outcome is binary, for example, and $g(.)$ is a *logit* link function, $_\Theta$ would correspond to the experimental arm's treatment effect on the outcome as measured by log odds ratio (log OR).

General model for comparing single treatment against multiple control types

It quickly became apparent that our general approach would be inadequate, as we learned that the individual trials were likely to have varying control conditions. To accommodate this, we adjusted the model so that the experimental treatment represented the reference category. So, $A_{ki} = 0$ if the patient is randomized to the experimental arm, $A_{ki} = 1$ otherwise and we assumed there would be C control types in total ($C > 1$).

The updated model was specified as follows:

$$\begin{aligned} g(E(Y_{ki})) =& \tau_k + \boldsymbol{\beta}\boldsymbol{X}_{ki} + \delta_{k_c} A_{ki} \\ \tau_k \sim& \text{Normal}\ (\mu = 0, \sigma = \sigma_\tau) \\ \boldsymbol{\beta} \sim& \text{Normal}\left(\boldsymbol{\mu} = \mathbf{0}, \Sigma = \sigma_\beta^2 I_{p\times p}\right) \\ \delta_{k_c} \sim& \text{Normal}\ (\mu = \delta_c, \sigma = \eta),\ \ c \in (1, \ldots, C) \\ \eta \sim& \text{Cauchy}\left(\mu = 0, \sigma = \sigma_\eta\right),\ \ \eta \geq 0 \\ \delta_c \sim& \text{Normal}(\mu = -\Delta, \sigma = \eta_0) \\ \eta_0 \sim& \text{Cauchy}\left(\mu = 0, \sigma = \sigma_{\eta_0}\right),\ \ \eta_0 \geq 0 \\ -\Delta \sim& \text{Normal}\ (\mu = 0, \sigma = \sigma_\Delta). \end{aligned} \tag{2}$$

The parameters τ_k and $\boldsymbol{\beta}$ mirror the same parameters in model (1). δ_{k_c} is the main effect of *control* treatment for the k^{th} RCT (i.e., RCT-specific "control effect"). Since the RCTs have the same experimental treatment but have different control types, c denotes the control type for the k^{th} RCT, where $c \in (1, \ldots, C)$. The prior distribution assumes each RCT-specific "control effect" δ_{k_c} is centered closely around its pooled "control effect" δ_c. We have introduced a hyperparameter η that represents the study variation around the control-type mean; this hyperparameter has its own prior distribution, and we need to provide an assumed standard deviation σ_η. In turn, the δ_c's are assumed to center around an overall effect of control treatment $-\Delta$. We use $-\Delta$ to represent the control-treatment effect contrast: $E(g(E(Y_{ki}|\boldsymbol{X}_{ki}, A_{ki} = 1)) - g(E(Y_{ki}|\boldsymbol{X}_{ki}, A_{ki} = 0)))$ across

all RCTs, and Δ represents the treatment-control effect contrast: $E(g(E(Y_{ki}|\boldsymbol{X}_{ki}, A_{ki}=0)) - g(E(Y_{ki}|\boldsymbol{X}_{ki}, A_{ki}=1)))$ across all RCTs. Δ is the parameter of interest in clinical trials. The variation around $-\Delta$ introduces a second hyperparameter η_0; we would need to provide an assumed standard deviation σ_{η_0}. At this point, we set prior distributions for η and η_0 as they are of paramount importance, while σ_τ, σ_β, σ_η and σ_{η_0} have point mass prior distribution as they are nuisance parameters. In evaluating evidence for efficacy, a point-mass prior of σ_Δ is often used [28, 29].

Basic model for ordinal outcome

The COMPILE model was ultimately developed to analyze an ordinal outcome measure (COVID-19 clinical status) proposed by the World Health Organization (WHO) and widely adopted by RCTs around the world to evaluate therapies for COVID-19. This clinical status scale helps RCTs measure COVID-19 severity, with values ranging from 0 = uninfected to 10 = dead (see Additional file 1); larger values on this scale indicate more severe disease [30]. We determined that a cumulative proportional odds (*co*) model would be most appropriate for the ordinal outcome measure [31].

Our goal was to develop a *co* model that reflected the structure of the general model (2) that we were interested in. As before, we assumed there would be K RCTs, each of which would be using one of C possible control conditions. All RCTs have the same experimental treatment: CCP. The ordinal outcome for the i^{th} patient from the k^{th} RCT is denoted by $Y_{ki} = y$, $y = 0, \ldots, 10$, and the cumulative probability is $p_{kiy} = P(Y_{ki} \geq y)$. As before, $A_{ki} = 0$ if the patient is randomized to CCP arm, $A_{ki} = 1$ otherwise.

The basic version of our *co* model was specified as follows:

$$\begin{aligned}
\text{logit}\big(P(Y_{ki} \geq y)\big) =& \tau_{yk} + \boldsymbol{\beta}\boldsymbol{X}_{ki} + \delta_{k_c} A_{ki} \\
\tau_{yk} \sim& \text{Normal}\ (\mu = 0, \sigma = \sigma_\tau) \\
\boldsymbol{\beta} \sim& \text{Normal}\left(\boldsymbol{\mu} = \mathbf{0}, \Sigma = \sigma_\beta^2 I_{p\times p}\right) \\
\delta_{k_c} \sim& \text{Normal}\ (\mu = \delta_c, \sigma = \eta),\ \ c \in (1, \ldots, C) \\
\eta \sim& \text{Cauchy}\ (\mu = 0, \sigma = \sigma_\eta) \\
\delta_c \sim& \text{Normal}(\mu = -\Delta_{co}, \sigma = \eta_0) \\
\eta_0 \sim& \text{Cauchy}\ (\mu = 0, \sigma = \sigma_{\eta_0}) \\
-\Delta_{co} \sim& \text{Normal}\ (\mu = 0, \sigma = \sigma_{\Delta_{co}}).
\end{aligned} \tag{3}$$

The *co* model differs from the general model (2) in several respects. The τ_{yk}'s represent the RCT-specific intercepts, for $y = 1, \ldots, 10$. For any specific RCT k, the τ_{yk}'s satisfy the monotonicity requirements for the intercepts of the *co* model. Since CCP treatment is the reference, the log-odds defined from the cumulative probabilities of the CCP arm are estimated by the τ_{yk}'s. $\boldsymbol{\beta}$ is a vector of coefficients for the p covariates. δ_{k_c} is the k^{th} RCT-specific "control effect", measured as a log OR. Because the RCTs have the same experimental treatment arm of CCP but have different control treatment arms, c denotes the control treatment type: standard of care (SOC) alone without any transfusion, $c = 1$; SOC plus non-convalescent plasma, $c = 2$; SOC plus saline solution, $c = 3$. The prior distribution assumes each RCT-specific "control effect" δ_{k_c} is centered closely around a pooled "control effect" δ_c, the corresponding type c control effect against CCP. The variation of each RCT effect around the groups' mean δ_c is η, estimated from the data. In turn, the δ_c's are assumed to center around $-\Delta_{co}$, the negative of the overall study-wise effect size. Δ_{co}, the key parameter of interest, represents the pooled cumulative log OR across all RCTs. We use $-\Delta_{co}$ so that Δ_{co} will correspond to the conventional difference of log-odds for CCP minus log-odds for control, rather than control minus CCP.

Extended models for ordinal outcome

We explored two major extensions of the basic model (3). First, we anticipated that some RCTs would be conducted at multiple sites, and we were interested in the model that included this added level of variation. Second, we expected that heterogeneity of treatment effect might be essential, so we explored another extension that incorporated an interaction term between the treatment and a pre-specified covariate.

Extended model for multi-site RCTs

We assumed that there would be K RCTs again, but M total sites, where $M > K$. The outcome for the i^{th} patient from the k^{th} RCT and the m^{th} site is denoted by $Y_{kmi} = y$, $y = 0, \ldots, 10$.

$$\text{logit}\big(P(Y_{kmi} \geq y)\big) = \tau_{ykm} + \boldsymbol{\beta}\boldsymbol{X}_{kmi} + \delta_{m_k} A_{kmi}. \tag{4}$$

The notation largely follows model (3). The extended model (4) incorporates new parameters: τ_{ykm}, and δ_{m_k}. The τ_{ykm} indicates the site-specific intercept and δ_{m_k} is the m^{th} site-specific "control effect". Each δ_{m_k} is normally distributed around a RCT-specific "control effect" δ_{k_c}, with a standard deviation η_1. More details of this extended model are in Additional file 2.

Extended model for assessing heterogeneity of treatment effect

To explore the impact of a pre-treatment covariate on the CCP effect, we developed another extension to model (3) for investigating the interaction between treatment and

a categorical pre-treatment covariate, denoted by S. Δ_s denotes the pooled effect of CCP (measured by log OR) for patients with covariate $S = s$. More details of this extended model are in Additional file 3.

Criteria for selecting prior distribution assumptions

In planning for the COMPILE analysis, it was key to establish the values of the standard deviation parameters (e.g., σ_τ, σ_β, σ_η, σ_{η_0}, $\sigma_{\Delta_{co}}$ in the *co* model (3)) as well as the prior distribution families to optimize the models with respect to bias and model stability. Below is illustrated how we identified the most appropriate model parameterization, prior distribution assumptions, and coding implementation strategies. At each stage of development of the statistical analysis plan, we conducted a series of simulations to assess the models under different conditions by varying effect sizes, the numbers of RCTs within each control condition, and the number of patients.

We considered the following specific criteria for choosing prior distributions:

Prior predictive checking: Prior predictive checking is a standard method to determine whether an assumed prior distribution is appropriate [23]. In particular, we used prior predictive checking described in [32] to ensure that all plausible values of the outcome (e.g., WHO 11-point scale) occurred with some probability. Because the WHO clinical status scale is ordinal and not continuous, this criterion was consistently satisfied across all assumed prior distributions.

Bias of estimated posterior distributions: If the sample size of the simulated study is large enough, an appropriate prior distribution should produce an estimated posterior distribution consistent with the data generation process. To assess this, we generated data sets under different scenarios for the effect size, and for each scenario, we generated 2500 studies each with a total of 900 patients. While the Bayesian analysis can provide the *full* posterior distribution of the parameter of interest based on each simulated study (in this case, Δ_{co}), it is challenging to compare models based on thousands of posterior distributions. Rather, we opted to use a single summary statistic, the posterior median, as the basis for comparison. For each effect size scenario, we constructed the distribution of posterior medians based on the 2500 simulated studies. An appropriate prior should result in a distribution of posterior medians that is centered around the true value used for the data generation.

Model stability: The Bayesian models were implemented in Stan software, which provides Bayesian inference over the model conditioned on data using Hamiltonian Monte Carlo (HMC) sampling. By default, the inference engine used is the No-U-Turn sampler (NUTS), an adaptive form of HMC sampling [33]. Divergent transitions that occur in the context of HMC sampling can lead to unreliable estimation of the posterior distributions. This results when the step size in the HMC sampling is too large to capture the highly varying posterior curvature [34]. Both model implementation and poorly conceived prior distribution assumptions can lead to undesirable levels of model divergence. The *proportion* of divergent transitions during HMC sampling is a widely used measure of stability and convergence. A model with a lower divergence rate is considered more reliable [35, 36].

Stable model estimation in Stan depends on two key tuning parameters: *adapt_delta* and *max_treedepth*. *adapt_delta* is the target average proposal acceptance rate applied during the model adaptation period; increasing this value results in a smaller step size for this gradient-based simulation of the Hamiltonian algorithm, allowing better exploration of the sample space [34]. The downsides are two-fold: (i) sampling tends to be slower because a smaller step size means that more steps are required to explore the posterior distribution thoroughly, and (ii) when the step size is too small, the sampler becomes inefficient, and the NUTS may stop before making a U-turn. But, we were able to mitigate these issues by increasing the second tuning parameter, *max_treedepth* [34].

Goodness-of-fit using posterior predictive checking

Any consumer of a statistical model likely will ask if the applied model is a good representation of the observed data. This is particularly important when the model in question, like a *co* model, makes a strong assumption. In this case, the model makes an assumption of proportional cumulative odds. In anticipation of potential deviations from the assumptions, researchers can simulate data under a range of possible violations of the model's assumptions and use posterior predictive checking to examine each model's *goodness-of-fit*.

Posterior predictive checking is a powerful method to assess a model's *goodness-of-fit* [37, 38]. The idea behind this technique is simple: if a model is a good fit, we should be able to use the model to generate replicated data (D^{rep}) that resemble our observed data (D^{original}) [39]. The lack of fit can be measured by the Bayesian p-value, which is the probability that the test statistic (e.g., $P(Y \leq y), y = 0, \ldots, 9$) for D^{rep} is equal to or greater than the test statistic for D^{original} [23]. A Bayesian p-value very close to zero or one is a cause for concern that the model is not fitting the data well, while a Bayesian p-value close to 0.5 means the model captures the data well [23, 33]. The procedure for checking whether the *co* model fits the observed data well and for calculating the Bayesian p-value can be found in Additional file 4.

Interim monitoring for efficacy

COMPILE pre-specified co-primary endpoints, both based on the WHO 11-point scale: the WHO 11-point ordinal scale, and a binary indicator of WHO ≥ 7. These two outcomes accommodate two essential functions: efficiency and interpretability. This section introduces the stopping rule based on the two outcomes.

Basic model for binary outcome

The second primary outcome selected with the goal of ease of clinical interpretation, is derived from the WHO 11-point clinical status scale and indicates that the patient is on mechanical ventilation or worse, i.e., WHO ≥ 7. We determined that a logistic (l) model would be most appropriate for the binary outcome. The l model was included for ease of communication and acceptability by the clinical community.

In model (5), W is an indicator variable for a WHO score ≥ 7, $W = 1$ if the patient has a WHO score ≥ 7, and $W = 0$ otherwise.

$$\text{logit}(P(W_{ki} = 1)) = \tau_k + \boldsymbol{\beta X}_{ki} + \delta_{k_c} A_{ki} \tag{5}$$

The parameters of model (5) mirror the parameters in model (3). The primary parameter of interest is Δ_l, the overall effect of CCP compared to any control.

Interim monitoring

In discussions with experts in the fields of RCTs, Bayesian analysis and monitoring of RCTs, conditions for stopping the COMPILE study were identified. The stopping rules were based on the following posterior probabilities for the ORs ($\text{OR}_{co} = e^{\Delta_{co}}$ and $\text{OR}_l = e^{\Delta_l}$):

$$\begin{aligned} P(\text{OR}_{co} < 1) \geq 0.95 \quad & \& \quad P(\text{OR}_{co} < 0.8) \geq 0.50 \\ & \text{and} \\ P(\text{OR}_l < 1) \geq 0.95 \quad & \& \quad P(\text{OR}_l < 0.8) \geq 0.50 \end{aligned} \tag{6}$$

When $\text{OR}_{co} < 1$ and $\text{OR}_l < 1$, CCP is at least minimally more effective than control; we required a high level of certainty that this be the case. When $\text{OR}_{co} < 0.8$ and $\text{OR}_l < 0.8$, it is considered that the beneficial effect of CCP is more than trivial; we required a moderate level of certainty that this be the case.

The stopping rules pertained to the monitoring of COMPILE study and the execution of the COMPILE meta-analysis itself, and had no direct bearing on the conduct of the individual studies. During the pandemic, the rapid dissemination of high-quality information was viewed as paramount, so once the criteria were met for stopping the prospective meta-analysis study, COMPILE data collection ceased, the final analyses were conducted, and results were published; the individual studies could have chosen to continue to enroll patients or suspend enrollment and continue only to follow up patients already enrolled.

The goal of the COMPILE was to provide answers regarding the efficacy of CCP treatment as soon as possible. Since neither the number of interim looks nor the number of RCTs and number of patients at each interim look could be predicted when we were planning the study, extensive simulations were required to calibrate the Bayesian criteria against the frequentist standards for type 1 error rates and statistical power.

Results

In order to finalize our analysis plan, we used extensive simulations to evaluate and choose prior distribution assumptions for both the basic and the extended models. We also used simulation to validate our proposed method of assessing goodness-of-fit using posterior predictive checking as well as assess the operating characteristics of the proposed Bayesian stopping rules for efficacy.

Evaluating and choosing prior distribution assumptions for the basic model

We started with an initial set of prior distribution assumptions (labeled as *Version 1*) for the parameters in the basic model (3). Using simulations, we evaluated the suitability of this version of assumptions with respect to the criteria described in the section titled Criteria for selecting prior distribution assumptions. Based on the findings from these simulations, we updated and reevaluated a new set of prior distribution assumptions. We iterated through this process a number of times until we were satisfied that the criteria were reasonably met. The sequential versions are shown in Table 1.

Simulation setup - basic model

The evaluation was conducted using the R package *simstudy* [40] to generate simulated data sets with the following parameters:

- We assumed different effect sizes for the three different control types. The overall effect Δ_{co} was set at the simple negative average of the three δ_c's :
  - $\delta_1 = 0.3$
  - $\delta_2 = 0.4$
  - $\delta_3 = 0.5$
  - $\Delta_{co} = -0.4$

Table 1 Prior distributions for different versions of cumulative proportional odds model

Versions	1	2	3	final
α	0	0	0	Normal ($\mu = 0$, $\sigma = 0.1$)
τ_{yk}	Normal ($\mu = 0$, $\sigma = 100$)	Normal ($\mu = 0$, $\sigma = 100$)	Normal ($\mu = 0$, $\sigma = 100$)	$t_{\text{student}}(\text{df} = 3,\ \mu = 0,\ \sigma = 8)$
$\boldsymbol{\beta}$	Normal ($\boldsymbol{\mu} = \mathbf{0}, \boldsymbol{\Sigma} = 100^2 I_{p\times p}$)	Normal ($\boldsymbol{\mu} = \mathbf{0}, \boldsymbol{\Sigma} = 100^2 I_{p\times p}$)	Normal ($\boldsymbol{\mu} = \mathbf{0}, \boldsymbol{\Sigma} = 100^2 I_{p\times p}$)	Normal ($\boldsymbol{\mu} = \mathbf{0}, \boldsymbol{\Sigma} = 2.5^2 I_{p\times p}$)
δ_{k_c}	Normal ($\mu = \delta_c$, $\sigma = \eta$)	Normal ($\mu = \delta_c$, $\sigma = \eta$)	Normal ($\mu = \delta_c$, $\sigma = \eta$)	Normal ($\mu = \delta_c$, $\sigma = \eta$)
η	Cauchy ($\mu = 0$, $\sigma = 100$)	$t_{\text{student}}(\mu = 0,\ \sigma = 100)$	$t_{\text{student}}(\mu = 0,\ \sigma = 100)$	$t_{\text{student}}(\text{df} = 3, \mu = 0,\ \sigma = 0.25)$
δ_c	Normal ($\mu = -\Delta_{co}$, $\sigma = \eta_0$)	Normal ($\mu = -\Delta_{co}$, $\sigma = \eta_0$)	Normal ($\mu = -\Delta_{co}$, $\sigma = \eta_0$)	Normal ($\mu = -\Delta_{co}$, $\sigma = \eta_0$)
η_0	Cauchy ($\mu = 0$, $\sigma = 100$)	$t_{\text{student}}(\mu = 0,\ \sigma = 100)$	$t_{\text{student}}(\mu = 0,\ \sigma = 100)$	0.1
$-\Delta_{co}$	Normal ($\mu = 0$, $\sigma = 100$)	Normal ($\mu = 0$, $\sigma = 100$)	Normal ($\mu = 0$, $\sigma = 0.354$)	Normal ($\mu = 0$, $\sigma = 0.354$)

- The between study and within control type variation was set at $\sigma = 0.1$
- We assumed three RCTs within each control type, with the size of the RCTs being
  - 1 large RCT with $n = 150$
  - 2 small RCTs, each with $n = 75$
- We started with a total sample size 900 as this was our initial aspiration for the COMPILE study

For each simulated individual, we generated a set of covariates: age (a categorical variable with 1 indicating < 50 years old, 2 indicating [50,65), and 3 indicating ≥ 65 years old), gender (a binary variable defined as female and male), WHO score at baseline (an ordinal variable with possible values of 4, 5, and 6) and duration of symptoms before randomization (a categorical variable with 1 for 0-3 days, 2 for 4-6 days, 3 for 7-10 days, 4 for 11-14 days, and 5 for 14+ days). We generated the distributions of these baseline covariates based on data available from the first RCT that joined the COMPILE consortium. In the simulation, we assumed the following distribution of baseline covariates: **age** < 50 years old : [50, 65) : ≥ 65 years old = 1 : 1 : 2; **sex** = female : male: = 1 : 1; **baseline WHO score** = 4 : 5 : 6 = 1 : 1 : 1; **duration of symptoms before randomization** = 0-3 days : 4-6 days : 7-10 days : 11-14 days : 14+ days = 1 : 1 : 1 : 1 : 1. The ordinal WHO outcome was generated as a function of the RCT-specific intercept, the individual-level covariates, and a random treatment assignment. We selected the coefficients from both the available COMPILE RCT data and from the literature describing outcomes of COVID-19 — male, older, and patients with higher severity of symptoms at baseline and patients with longer duration of symptoms were at higher risk for worse outcome. We present the coefficients of the covariates as "true values" in Additional file 5.

We simulated 2500 trials (each trial included 9 RCTs) for model fitting. For each simulated trial, we used 2000 HMC iterations for warm-up and retained 10000 iterations for inference (all simulations in this paper used the same number of HMC iterations. See code for simulations in Additional file 6.).

Figure 1 shows the bias of the posterior estimations as well as the divergences resulting from each set of

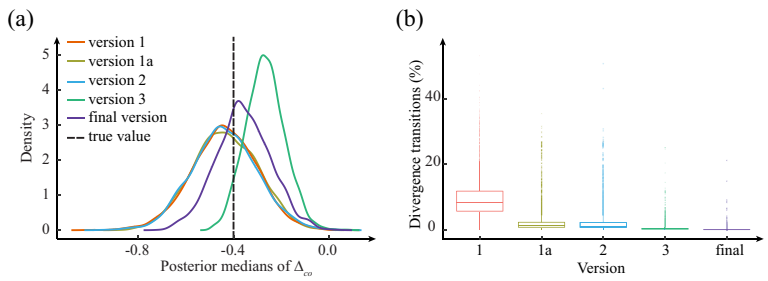

Fig. 1 The performance of the five versions of the basic *co* model: **a** the distribution of posterior medians of the pooled CCP treatment effect Δ_{co} and the true value used to generate the data (true value); **b** boxplots show the median, lower quartile, upper quartile, minimum, and maximum of the number of divergent transitions (%). The proportion of divergent transitions was calculated by (the number of divergent transitions/10,000)× 100% in each simulated trial using the five versions of model

modelling assumptions based on our simulating assumptions. The bias is based on a comparison of the center of the distribution of posterior medians with the "true value" of $\Delta_{co} = -0.4$; the greater the difference, the greater the bias.

Models Version 1 and 1(a)

Version 1 was based largely on non-informative prior distribution assumptions [41]. The proportion of divergent transitions from the simulations was unacceptably high, indicating that the posterior estimation from this model was unlikely to be reliable (see Fig. 1(b)).

Version 1(a) was unchanged from Version 1, except that we used non-centered parameterization to implement the model in Stan. Non-centered parameterization is an additional tool to improve model fitting, and the *Normal* distribution is the best candidate for reparameterization [33]. Based on this, we applied the non-centered parameterization for all *Normal* prior distributions in the model. While the proportion of divergent transitions decreased relative to Version 1, the reduction was not sufficient to ensure model stability (see Fig. 1(b)).

Model Version 2

In the second version of the model, we replaced the *Cauchy* distributions with $t_{student}$ distributions, which are more suitable as prior distributions of the scale parameters [42]. The $t_{student}$ distribution with 3 degrees of freedom ($\sigma = 100$) has tails that present a compromise between the *Normal* distribution ($\sigma = 100$) and the *Cauchy* distribution ($\sigma = 100$). This change resulted in fewer divergent transitions compared to Version 1(a), but again the reduction was insufficient (see Fig. 1(b)).

Model Version 3

In the third iteration, we imposed a skeptical prior on $-\Delta_{co}$ to reflect a very conservative belief in the efficacy of the treatment. A skeptical prior distribution assumes that the probability of large benefit or harm from the experimental treatment is low and that the probability of equivalence between the treatments is high. The standard deviation of the prior ($\sigma_{\Delta_{co}}$) was set at 0.354, which corresponds to a prior for an OR with 95% density between 0.5 and 2. With this prior distribution, the effect of CCP is assumed to be close to zero, and only strong evidence from the data can alter this prior belief. When assessing evidence of efficacy, such a skeptical prior for the treatment effect is considered appropriate [24, 28, 29]. However, the distribution of posterior medians from posterior estimates of the overall effect did not adequately reflect the "true" underlying data generation process (see Fig. 1(a)).

Final Version

We settled on this final version of the *co* model after the sequence of simulation experiments:

$$\begin{aligned}
\text{logit}\big(P(Y_{ki} \geq y)\big) &= \alpha + \tau_{yk} + \boldsymbol{\beta}\boldsymbol{X}_{ki} + \delta_{k_c} A_{ki} \\
\alpha &\sim \text{Normal}\ (\mu = 0, \sigma = 0.1) \\
\tau_{yk} &\sim t_{\text{student}}(\text{df} = 3, \mu = 0, \sigma = 8) \\
\boldsymbol{\beta} &\sim \text{Normal}\ \big(\boldsymbol{\mu} = \mathbf{0}, \Sigma = 2.5^2 I_{p\times p}\big) \\
\delta_{k_c} &\sim \text{Normal}\ \big(\mu = \delta_c, \sigma = \eta\big)\ \ c \in (1, 2, 3) \\
\eta &\sim t_{\text{student}}(\text{df} = 3, \mu = 0, \sigma = 0.25) \\
\delta_c &\sim \text{Normal}\ \big(\mu = -\Delta_{co}, \sigma = 0.1\big) \\
-\Delta_{co} &\sim \text{Normal}\ (\mu = 0, \sigma = 0.354)
\end{aligned} \tag{7}$$

The following updates were made to Version 3.

Global intercept α: α is a nuisance parameter that we expected to be very close to or at 0. Since we modelled each RCT-specific intercept directly, α was effectively fixed to 0 in Version 3. However, model fitting improved when α was freely estimated. In the final version of the model, we used a highly informative prior where most of the probability mass was set close to zero.

RCT-specific intercepts τ_{yk}: We initially assumed that the prior distribution for each τ_{yk} was $Normal(\mu = 0, \sigma = 100)$. The domain experts participating in the study suggested that we use a weakly-informative prior distribution with scale parameter = 8 to act as somewhat of a constraint. However, we used a $t_{student}$ distribution with 3 degrees of freedom ($\sigma = 8$). This $t_{student}$ distribution has heavier tails than the *Normal* distribution ($\sigma = 8$), so we ensured that the HMC simulation would have enough flexibility to explore the sample space.

Covariate coefficients $\boldsymbol{\beta}$: We had little prior information for $\boldsymbol{\beta}$ and expected the observed data to determine the shape of the posterior distribution, so we assumed a diffuse prior distribution.

Standard deviation of RCT-specific "control effect" η: The prior distribution of δ_{k_c} is normally distributed around its own "control effect": δ_c , with a standard deviation η. In the final version, η had an informative prior distribution $t_{student}(df = 3, \mu = 0, \sigma = 0.25)$. This $t_{student}$ distribution has heavier tails than the *Normal* distribution with equivalent scale parameters($\sigma = 0.25$).

Standard deviation of control-type effect η_0: We consulted the domain experts involved in the study, who suggested that the three types of control conditions should not differ greatly; in particular, they believed 95% of the possible values (log OR) should be within 0.2 of the mean, implying a standard deviation of 0.1. Thus, the solution was to use a more informative prior with narrow tails.

The simulation results shown in Fig. 1 indicate that the position with the highest probability is very close to the "true" value for data generation. Because of the

postulated skeptical prior, the posterior estimation was pulled very slightly (rightwards) towards 0, despite the relatively large sample size. The number of divergent transitions was close to or at zero for nearly all simulated trials, indicating that model fitting converged almost every time.

We also assessed the models under different scenarios for the effect sizes. The results were consistent at different effect sizes ($\delta_1, \delta_2, \delta_3$). See Additional file 7 for the bias in posterior estimates and divergent transitions resulting from each version of the *co* model under a different scenario for effect sizes $(\delta_1, \delta_2, \delta_3) = (0.05, 0.1, 0.15)$.

With the final model in hand, we were able to look at extended models, explore goodness-of-fit methods, evaluate operating characteristics of stopping rules, and examine the influence of the sample size we had assumed (see Additional file 8 for model fitting results of the *co* model using various sample sizes).

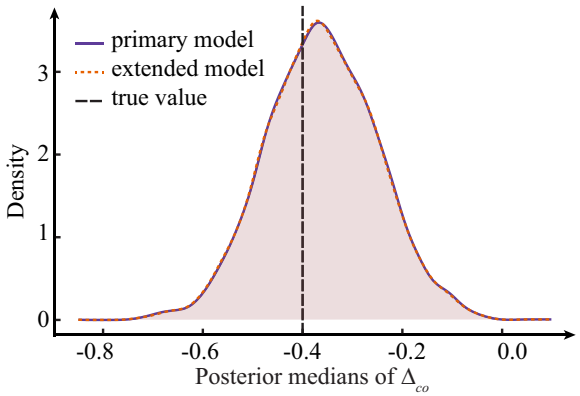

Fig. 2 The distribution of posterior medians of pooled CCP treatment effect Δ_{co} using the final version of basic *co* model and the extended *co* model for multi-site RCTs. The black dashed line represents the true value of parameter used to generate the data

Evaluating and choosing priors for the extended model

After finalizing the basic *co* model (3) as in model (7), we turned our attention to the extended models.

Extended model for multi-site RCTs

In the simulation setup, we assumed that there would be K RCTs again, but M total sites, where $M > K$. The outcome for the i^{th} patient from the k^{th} RCT and the m^{th} site is denoted by $Y_{kmi} = y, y = 0, \ldots, 10$.

- 3 control types with effect sizes: $\delta_1 = 0.3$, $\delta_2 = 0.4$, $\delta_3 = 0.5$
- Between study (within control type) variation $\sigma = 0.1$
- 3 RCTs within each control type
  - 1 large RCT with $n = 150$: 1 large site with $n = 110$ and 2 small sites with $n = 20$
  - 2 small RCTs, each with $n = 75$: 1 large site with $n = 55$ and 2 small sites with $n = 10$
- Between sites (within RCT) variation $\sigma = 0.1$

For each simulated individual, we also generated a set of covariates: age, gender, WHO score at baseline, and duration of symptoms before randomization. The randomization to CCP and control is within sites. The prior distributions for the extended model for multi-site RCTs are in Additional file 9. We conducted 3000 simulations to compare the performance of model (7) and the model for multi-site RCTs. and found that both performed well in recovering the "true value" (see Fig. 2; the estimations of all parameters can be found in Additional file 10). The posterior distributions from both models were virtually identical. Given the similarities of the model estimates, we opted for the simpler model (7) that has fewer hierarchical assumptions and less complexity.

Extended model for assessing heterogeneity of treatment effect

Next, we focused on selecting the prior distributions for the extended model for studying the heterogeneity of the treatment effect (model (A2) in Additional file 3), which includes a term for the interaction between treatment indicators and a categorical pre-treatment variable S.

In the simulation setup, the data were simulated as described in the Simulation setup - basic model section with adjustment for the same set of covariates and a categorical covariate S with three levels (30% patients with $S = 1$, 30% patients with $S = 2$, and 40% patients with $S = 3$), as well as the interaction between covariate S and treatment. The covariate S was not associated with the other covariates. We conducted a series of simulations to assess the models under different conditions by varying treatment's effect sizes for the overall and for the subgroups. The prior distributions of the model with the interaction between treatment and a pre-specified covariate S are specified in model (A4) in Additional file 11.

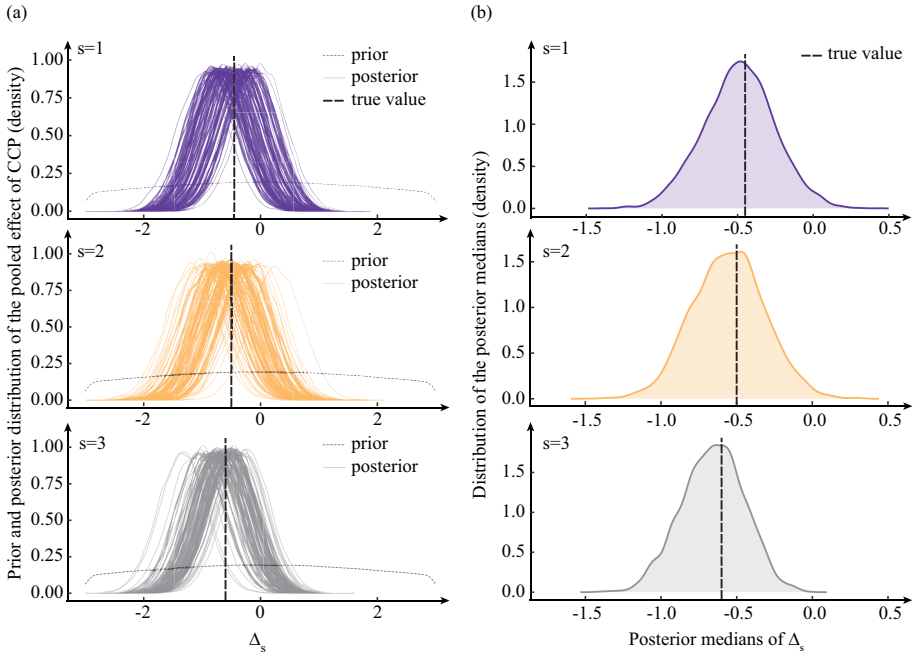

Fig. 3 The performance of the extended model for assessing heterogeneity of treatment effect: **a** prior and 100 posterior distributions of the pooled CCP effect in each level of a pre-specified covariate (Δ_s, $s = 1,2,$ or 3), **b** Distribution of posterior medians of Δ_s (Based on 4500 simulated trials). The black dashed lines represent the true values of parameters used to generate the data: $\Delta_{s=1} = -0.45$, $\Delta_{s=2} = -0.5$, $\Delta_{s=3} = -0.6$

Figure 3 shows the performance of the extended model for assessing the heterogeneity of treatment effect. Figure 3(a) shows the prior distribution and 100 posterior distributions of Δ_s. Figure 3(b) shows the distributions of the posterior medians of Δ_s from 4500 simulated trials. The Bayesian model estimations had a very high probability of recovering "true values".

Simulation results for goodness-of-fit

In the case of the *co* model, there is an assumption of proportional cumulative odds. We investigated to what degree the proportional odds assumption has to be violated to render the model inaccurate. We considered two data-generating mechanisms: the observed data generated under (i) a proportional cumulative odds assumption or (ii) a non-proportional cumulative odds assumption. In addition to the case satisfying the proportionality assumptions, in Fig. 4 and Table 2 we report two cases where the proportionality assumptions were violated: one with small and one with large deviation from proportionality of the odds.

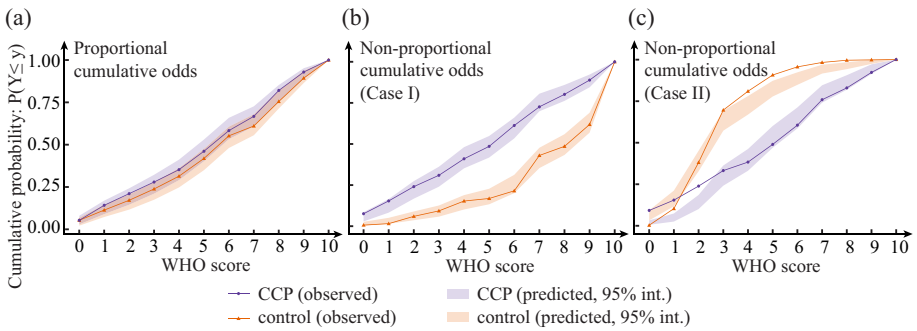

Fig. 4 The observed cumulative probabilities for the CCP arm and the control treatment arm (CCP arm in observed data: solid purple line with marker ◯; control arm in observed data: solid orange line with marker △) as well as the 95% credible interval (the colored bands) for the predicted cumulative probabilities using the posterior predictive checking: **a** observed data was generated under proportional cumulative odds assumption, **b** observed data was generated when the proportional cumulative odds assumption was violated only slightly (Case I), **c** observed data was generated when the proportional cumulative odds assumption was violated more extremely (Case II)

Table 2 Summary of posterior predictive checking based on the ten test statistics

Assumption						
Treatment	**CCP**			**Control**		
Test quantity: % subjects	$T(D^{original})$	**95% int. for** $T(D^{rep})$	**Bayesian *P* value**	$T(D^{original})$	**95% int. for** $T(D^{rep})$	**Bayesian *P* value**
	a Proportional cumulative odds					
WHO ≤ 0	5.32	[2.91, 7.99]	0.48	5.35	[2.18, 6.80]	0.14
WHO ≤ 1	14.19	[8.96, 16.95]	0.24	11.36	[7.28, 14.56]	0.34
WHO ≤ 2	21.06	[14.53, 24.21]	0.23	17.15	[11.65, 20.63]	0.32
WHO ≤ 3	27.94	[21.07, 32.20]	0.30	23.83	[17.48, 27.91]	0.29
WHO ≤ 4	35.25	[29.06, 41.16]	0.45	31.40	[24.76, 36.17]	0.34
WHO ≤ 5	46.12	[40.19, 53.03]	0.53	41.87	[34.95, 47.57]	0.42
WHO ≤ 6	58.31	[53.27, 65.86]	0.67	55.23	[48.06, 60.68]	0.40
WHO ≤ 7	66.74	[60.77, 72.64]	0.52	61.02	[55.58, 67.96]	0.61
WHO ≤ 8	82.04	[75.30, 85.23]	0.28	75.50	[71.36, 82.04]	0.68
WHO ≤ 9	92.90	[88.86, 95.16]	0.35	89.31	[86.41, 93.69]	0.74
	b Non-proportional cumulative odds (Case I)					
WHO ≤ 0	7.57	[2.91, 8.50]	0.07	0.67	[0.24, 2.66]	0.88
WHO ≤ 1	15.37	[8.25, 16.50]	0.07	1.55	[1.45, 5.08]	0.95
WHO ≤ 2	24.05	[16.75, 27.43]	0.21	5.99	[3.63, 8.96]	0.54
WHO ≤ 3	30.96	[24.03, 36.17]	0.37	9.31	[5.81, 12.59]	0.42
WHO ≤ 4	40.98	[35.19, 48.06]	0.55	15.30	[10.17, 18.89]	0.30
WHO ≤ 5	48.55	[41.75, 54.85]	0.44	16.85	[13.32, 22.76]	0.65
WHO ≤ 6	61.25	[53.40, 66.02]	0.32	21.51	[20.10, 31.23]	0.93
WHO ≤ 7	72.61	[69.90, 80.58]	0.84	43.02	[35.35, 47.70]	0.32
WHO ≤ 8	80.18	[76.21, 85.92]	0.65	48.56	[43.34, 55.93]	0.63
WHO ≤ 9	88.86	[84.47, 92.23]	0.42	61.86	[56.90, 69.01]	0.64
	c Non-proportional cumulative odds (Case II)					
WHO ≤ 0	9.11	[0.49, 3.16]	0.00	0.22	[3.15, 9.20]	1.00
WHO ≤ 1	15.33	[2.43, 7.28]	0.00	10.22	[11.62, 21.07]	1.00
WHO ≤ 2	23.78	[9.95, 18.45]	0.00	38.00	[32.69, 45.04]	0.61
WHO ≤ 3	33.11	[24.03, 36.17]	0.16	69.56	[57.14, 69.25]	0.02
WHO ≤ 4	38.22	[33.01, 46.12]	0.65	80.89	[66.83, 77.97]	0.00
WHO ≤ 5	48.89	[46.84, 59.95]	0.90	90.67	[77.72, 87.17]	0.00
WHO ≤ 6	60.44	[58.50, 71.12]	0.92	95.56	[84.75, 92.25]	0.00
WHO ≤ 7	75.78	[74.03, 84.47]	0.92	98.22	[91.53, 96.85]	0.00
WHO ≤ 8	82.89	[81.80, 90.53]	0.94	99.56	[94.19, 98.31]	0.00
WHO ≤ 9	92.22	[91.26, 97.09]	0.92	99.78	[97.34, 99.76]	0.01

Note: Case I: the proportional cumulative odds assumption was violated only slightly. Case II: the proportional cumulative odds assumption was violated more extremely

In Fig. 4, the bands represent the 95% credible interval for test statistics based on 10000 replicated datasets D^{rep}. The solid lines with markers represent the cumulative probabilities in D^{original}. Table 2 shows posterior predictive checking of model (7) using ten test statistics. We confirmed that the model fits the data well if the data generation process satisfied the proportional odds assumption. When the proportional cumulative odds assumption was violated only slightly (Case I), only one Bayesian *p*-value [23, 33] was close to one (i.e., 0.95), which would still give us confidence that our model was a good fit. However, when the proportional cumulative odds assumption was violated more extremely (Case II) in the data generation process, most Bayesian *p*-values were extreme (i.e., close to zero or one), indicating our model might be a poor fit.

Bayesian stopping rules for efficacy

We investigated the probability of stopping early under the proposed Bayesian approach (6) using a range of

effect sizes and sample sizes (Scenario (1) was simulated as in Simulation setup - basic model section: $n = 900$, Scenario (2) doubled the sample size and Scenario (3) tripled the sample size).

Parameters:

- The Bayesian paradigm includes nine data looks at 20%, 33%, 40%, 50%, 60%, 67%, 80%, 90% and 100% of the data. Three sets of control-specific treatment effects as measured by log OR $(\delta_1, \delta_2, \delta_3)$ are considered:
  - (0,0,0), pooled control effect is 0
  - (0.1, 0.2, 0.3), pooled control effect is 0.2
  - (0.4, 0.5, 0.6), pooled control effect is 0.5
- No covariate adjustment in both data generation and analysis

When the simulated effect is $(\delta_1, \delta_2, \delta_3) = (0, 0, 0)$, the sum of the probabilities of meeting the stopping trigger at all interim looks under the Bayesian monitoring approach can be interpreted as the type 1 error rate (see Fig. 5(a)). When data are simulated under the assumption of efficacy, the sum of the probabilities of meeting the stopping trigger (over all interim looks) can be interpreted as statistical power (see Fig. 5(b) and (c)).

The prior distributions in the model for the binary co-primary outcome WHO≥ 7 (model (5)) were selected through a process similar to the process for selecting the prior distributions for model (3) for the ordinal outcome. The prior distributions for the binary co-primary outcome are shown in Additional file 12:

Figure 5 shows the results for all three sets of effect sizes. The type 1 error rates were considerably below the 5% threshold. As expected, the type 1 error rate declined as the sample size increased. In the cases where CCP was assumed to be effective, larger effect sizes and larger sample sizes both increased power. We can see that the proposed Bayesian stopping rule would achieve acceptable type 1 error rates and power.

Continuous monitoring is critical in a pandemic to detect early signals of efficacy and make timely decisions. One concern with the increasing number of interim looks would be inflated type 1 error rate. In our simulation, we expected nine data looks so the stopping rules resulted in acceptable type 1 error rates and power. If researchers expect more interim looks, there are three ways to control for the inflated type 1 error rate: (i) adopt a more skeptical prior distribution for the treatment effect (OR_{co} and OR_l); (ii) increase the sample size; or (iii) set a more restrictive threshold for the stopping criteria. For example, the current threshold for the second criterion of clinically meaningful effect (i.e., $P(OR < 0.8) \geq 0.5$) in the stopping rules is 0.5. We could increase 0.5 to 0.6 to reduce the type I error rate while keeping the prior distributions and sample size constant.

Discussion and conclusions

The presented work describes a translatable framework for developing a rigorous plan for monitoring and analysis of a study that prospectively pools IPD from ongoing, paused, prematurely-terminate, or completed RCTs with the goal of reaching a conclusion regarding the efficacy of a treatment as quickly as possible. Such studies were in particularly high demand during the initial stages of the COVID-19 pandemic, and it is expected that they would be needed not only in future pandemics but also for contributing to more efficient non-pandemic medical research. While the idea of such prospective pooling of data from RCTs at different stages of execution is simple and appealing, the development of the analytic plan for monitoring and analysis is not trivial.

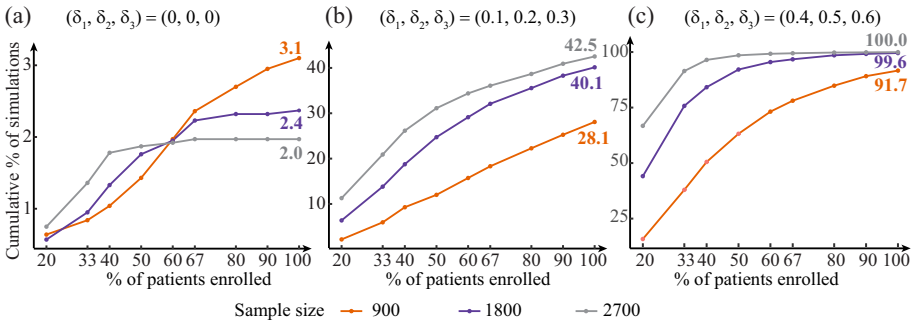

Fig. 5 At different sample sizes, the proportion of times (out of 2000) in which the stopping rules were reached under the Bayesian monitoring approach. The colored numbers in **a** can be interpreted as the type 1 error rates at different sample sizes. The colored numbers in **b** and **c** can be interpreted as the statistical power

In this paper, we report on the extensive simulation investigations that were needed for selecting models and parameters to estimate and for choosing the prior distributions for these parameters. We also show how we can study the operating characteristics of guidelines for continuous monitoring in the absence of information about the total sample size, the rate of patient recruitment, and the number of interim looks at the time of study planning.

Our work should be interpreted in the context of three potential limitations. First, our extended model for assessing heterogeneity of the treatment effect (model (A4)) was designed for the interaction term between a categorical covariate and treatment. It would be useful to extend the model to incorporate the interaction term between a continuous covariate and treatment. While this is important methodologically, it may be less so clinically, since patient characteristics that are best measured using a continuous scale are routinely considered in categorical terms; viewed this way, providing interaction models only for categorical characteristics may not be such a serious limitation. Second, the 95% credible interval for the CCP treatment effect from model for assessing heterogeneity of treatment effect (model (A4)) tends to be wider than model (7) in subgroup analysis because of the diffuse prior in model (A4). Developing more efficient approaches for estimating interactions would be a valuable contribution. Third, our method focuses on sampling from the posterior distribution of the effect size Δ rather than testing the equality of experimental and control treatments, an approach that some believe is more appropriate for this setting [43]. The testing formulation, however, can require high computational overhead compared to the estimation approach we used. Regardless of whether one considers testing or estimation to be more appropriate at the stage when the prospective IPD study is completed and the final data are available, at the stage of developing the analytic plan, there may be less flexibility. Unless advances in computing make the testing approach more practical, when extensive simulations are necessary and must include a range of relatively large sample sizes (here 900, 1800 and 2700), we recommend the estimation approach described in this paper.

To the best of our knowledge, an initiative like COMPILE has not been undertaken previously. In conducting this study, we believe we have developed a translatable framework that can be used to inform such endeavors in the future. This framework can leverage information quickly for other types of therapies under simultaneous investigations around the world. Not only can this framework be a valuable tool for assessing new treatment options for COVID-19, but it can also be useful for the treatment of other diseases.

Abbreviations

COVID-19	Coronavirus disease 2019
COMPILE	COntinuous Monitoring of Pooled International Trials of ConvaLEscent Plasma for COVID-19 Hospitalized Patients
IPD	individual patient data
RCT(s)	randomized controlled trial(s)
SOC	standard of care
CCP	COVID-19 convalescent plasma
WHO	World Health Organization
HMC	Hamiltonian Monte Carlo
NUTS	No-U-Turn sampler
co	cumulative proportional odds
l	logistic
OR(s)	odds ratio(s)
NYU	New York University

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-022-01813-4.

Additional file 1. The WHO 11-point COVID-19 clinical status scale [30].

Additional file 2. Extended model for multi-site RCTs.

Additional file 3. Extended model for assessing heterogeneity of treatment effect.

Additional file 4. Procedure for posterior predictive checks [23].

Additional file 5. The true value of parameters in the data generation process.

Additional file 6. a R code for simulation [33, 44, 45]. **b** Stan code for the final version of the *co* model.

Additional file 7. Assessing models using a different set of effect sizes.

Additional file 8. The effect of sample sizes on the model's performance.

Additional file 9. Final model for multi-site RCTs.

Additional file 10. Posterior estimations of parameters in the extended model for multi-site RCTs.

Additional file 11. Final model for assessing heterogeneity of treatment effect.

Additional file 12. Final model for the binary co-primary outcome.

Acknowledgements
The authors acknowledge Thaddeus Tarpey, Hyung Park, Mengling Liu, and Andrea Troxel for their guidance on developing the models.

Authors' contributions
DW: participated in discussions of the problem, contributed to the creation of the Bayesian hierarchical models, ran all simulations, and wrote the original draft. KSG and EP: supervision, project administration, funding acquisition, participated in discussions of the problem, contributed to the creation of the Bayesian hierarchical models, reviewed and edited the draft. All authors read and approved the final manuscript.

Funding
This work was supported by the National Center for Advancing Translational Sciences of the National Institutes of Health under Award Number UL1TR001445.

Availability of data and materials
The datasets used and analysed during the current study are available from the corresponding author on reasonable request.

Declaration

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Population Health, New York University Grossman School of Medicine, New York, USA. [2]Department of Child and Adolescent Psychiatry, New York University Grossman School of Medicine, New York, USA. [3]Nathan Kline Institute for Psychiatric Research, Orangeburg, USA.

Received: 13 June 2022 Accepted: 5 December 2022
Published online: 25 January 2023

References

1. World Health Organization. COVID-19 coronavirus pandemic. 2020. https://www.euro.who.int/en/health-topics/health-emergencies/coronavirus-covid-19. Accessed 21 Nov 2022.
2. Pinney SP, Giustino G, Halperin JL, et al. Coronavirus historical perspective, disease mechanisms, and clinical outcomes: JACC focus seminar. J Am Coll Cardiol. 2020;76(17):1999–2010. https://doi.org/10.1016/j.jacc.2020.08.058.
3. World Health Organization. WHO Coronavirus (COVID-19) Dashboard. 2021. https://covid19.who.int. Accessed 21 Nov 2022.
4. VanBlargan LA, Errico JM, Halfmann PJ, et al. An infectious SARS-CoV-2 B.1.1.529 Omicron virus escapes neutralization by several therapeutic monoclonal antibodies. bioRxiv. 2021. https://doi.org/10.1101/2021.12.15.472828.
5. US National Library of Medicine. ClinicalTrials.gov. 2021. https://clinicaltrials.gov. Accessed 21 Nov 2022.
6. Petkova E, Antman EM, Troxel AB. Pooling data from individual clinical trials in the COVID-19 era. JAMA J Am Med Assoc. 2020;324(6):543–5. https://doi.org/10.1001/jama.2020.13042.
7. Avendaño-Solà C, Ramos-Martínez A, Muñez-Rubio E, Ruiz-Antorán B, Malo de Molina R, Torres F, et al. Convalescent plasma for COVID-19: a multicenter, randomized clinical trial. MedRxiv. 2020. https://doi.org/10.1101/2020.08.26.20182444.
8. Li L, Zhang W, Hu Y, et al. Effect of convalescent plasma therapy on time to clinical improvement in patients with severe and life-threatening COVID-19: a randomized clinical trial. JAMA J Am Med Assoc. 2020;324(5):460–70. https://doi.org/10.1001/jama.2020.10044.
9. Libster R, Pérez Marc G, Wappner D, Coviello S, et al. Early high-titer plasma therapy to prevent severe COVID-19 in older adults. N Engl J Med. 2021;384(7):610–8. https://doi.org/10.1056/NEJMoa2033700.
10. Gharbharan A, Jordans CCE, Geurtsvankessel C, den Hollander JG, Karim F, Mollema FPN, et al. Convalescent plasma for COVID-19. A randomized clinical trial MedRxiv. 2020. https://doi.org/10.1101/2020.07.01.20139857.
11. Ortigoza MB, Yoon H, Goldfeld KS, Troxel AB, Daily JP, Wu Y, et al. Efficacy and safety of COVID-19 convalescent plasma in hospitalized patients: a randomized clinical trial. JAMA Intern Med. 2022;182(2):115–26. https://doi.org/10.1101/2020.07.01.20139857.
12. Bauchner H, Fontanarosa PB. Randomized clinical trials and COVID-19: managing expectations. JAMA J Am Med Assoc. 2020;323(22):2262–3. https://doi.org/10.1001/jama.2020.8115.
13. Klassen SA, Senefeld JW, Johnson PW, Carter, et al. The effect of convalescent plasma therapy on COVID-19 patient mortality: systematic review and meta-analysis. MedRxiv. 2021. https://doi.org/10.1101/2020.07.29.20162917.
14. Janiaud P, Axfors C, et al. Association of convalescent plasma treatment with clinical outcomes in patients with COVID-19: a systematic review and meta-analysis. JAMA J Am Med Assoc. 2021;325(12):1185–95. https://doi.org/10.1001/jama.2021.2747.
15. Juul S, Nielsen N, Bentzer P, et al. Interventions for treatment of COVID-19: a protocol for a living systematic review with network meta-analysis including individual patient data (the LIVING project). Syst Rev. 2020;9(1):108. https://doi.org/10.1186/s13643-020-01371-0.
16. COMPILE International Study Team. Continuous monitoring of pooled international trials of convalescent plasma for COVID-19 hospitalized patients: a prospective individual patient data meta-analysis. 2021. http://nyulmc.org/compile. Accessed 21 Nov 2022.
17. Troxel AB, Petkova E, Goldfeld KS, Liu M, Tarpey T, Wu Y, et al. Association of convalescent plasma treatment with clinical status in patients hospitalized with COVID-19: a meta-analysis. JAMA Netw Open. 2022;5(1):e2147331. https://doi.org/10.1186/s13643-020-01371-0.
18. Stangl D, Berry DA. Meta-analysis in medicine and health policy. Boca Raton: CRC Press; 2000.
19. Goldfeld KS, Wu D, Tarpey T, Liu M, Wu Y, Troxel AB, et al. Prospective individual patient data meta-analysis: evaluating convalescent plasma for COVID-19. Stat Med. 2021;40(24):5131–51. https://doi.org/10.1002/sim.9115.
20. Park H, Tarpey T, Liu M, Goldfeld KS, Wu Y, Wu D, et al. Development and validation of a treatment benefit index to identify hospitalized patients with COVID-19 who may benefit from convalescent plasma. JAMA Netw Open. 2022;5(1):e2147375. https://doi.org/10.1001/jamanetworkopen.2021.47375.
21. Pedroza C, Tyson JE, Das A, Laptook A, Bell EF, et al. Advantages of Bayesian monitoring methods in deciding whether and when to stop a clinical trial: an example of a neonatal cooling trial. Trials. 2016;17(1):335. https://doi.org/10.1186/s13063-016-1480-4.
22. Lee J, Chu CT. Bayesian clinical trials in action. Stat Med. 2012;31(25):2955–72. https://doi.org/10.1002/sim.5404.
23. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian data analysis. 3rd ed. 2013. https://doi.org/10.1201/b16018.
24. Berry SM, Carlin BP, Lee JJ, Peter M. Bayesian adaptive methods for clinical trials. 1st ed. 2010. https://doi.org/10.1201/EBK1439825488.
25. Lewis RJ, Angus DC. Time for clinicians to embrace their inner bayesian? Reanalysis of results of a clinical trial of extracorporeal membrane oxygenation. JAMA J Am Med Assoc. 2018;320(21):2208–10. https://doi.org/10.1001/jama.2018.16916.
26. Saville BR, Connor JT, Ayers GD, Alvarez J. The utility of Bayesian predictive probabilities for interim monitoring of clinical trials. Clin Trials. 2014;11(4):485–93. https://doi.org/10.1177/1740774514531352.
27. Gelman A, Hill J, Yajima M. Why we (usually) don't have to worry about multiple comparisons. J Res Educ Eff. 2012;5(2):189–211. https://doi.org/10.1080/19345747.2011.618213.
28. Harrell FE, et al. Regression modeling strategies, with applications to linear models, logistic and ordinal regression, and survival analysis. 2nd ed. 2015. https://doi.org/10.1007/978-3-319-19425-7.
29. Casey JD, Johnson NJ, Semler MW, et al. Rationale and design of ORCHID: a randomized placebo-controlled clinical trial of hydroxychloroquine for adults hospitalized with COVID-19. Annals of the American Thoracic Society. 2020;17(9):1144–53. https://doi.org/10.1513/AnnalsATS.202005-478SD.
30. A minimal common outcome measure set for COVID-19 clinical research. Lancet Infect Dis. 2020 8;20(8):e192–e197. https://doi.org/10.1016/S1473-3099(20)30483-7.
31. Agresti A. Categorical data analysis. 2nd ed. 2002. https://doi.org/10.1002/0471249688.
32. Van de Schoot R, Depaoli S, King R, Kramer B, Märtens K, Tadesse MG, et al. Bayesian statistics and modelling. Nat Rev Methods Prim. 2021;1(1):1–26. https://doi.org/10.1038/s43586-020-00001-2.
33. Stan Development Team. Stan modeling language users guide. 2020. https://mc-stan.org/docs. Accessed 21 Nov 2022.
34. Stan Develpment Team. Stan reference manual. 2020. https://mc-stan.org/docs. Accessed 21 Nov 2022.
35. Betancourt M. Diagnosing suboptimal cotangent disintegrations in Hamiltonian Monte Carlo. 2016. arXiv. https://doi.org/10.48550/arXiv.1604.00695.
36. Betancourt M. A conceptual introduction to Hamiltonian Monte Carlo. 2017. arXiv. https://doi.org/10.48550/arXiv.1701.02434.
37. Donald RB. Bayesianly justifiable and relevant frequency calculations for the applied statistician. Ann Stat. 1984;12(4):1151–72. https://doi.org/10.1214/AOS/1176346785.
38. Gelman A, Meng XL, Stern H. Posterior predictive assessment of model fitness via realized discrepancies. Stat Sin. 1996;6(4):733–807.
39. Gabry J, Simpson D, Vehtari A, Betancourt M, Gelman A. Visualization in Bayesian workflow. J R Stat Soc Ser A Stat Soc. 2019;182(2):389–402. https://doi.org/10.1111/rssa.12378.

40. Goldfeld KS, Wujciak-Jens J. Package "simstudy" R topics documented. 2020. https://cran.r-project.org/web/packages/simstudy/simstudy.pdf. Accessed 21 Nov 2022.
41. Kass RE, Wasserman L. The selection of prior distributions by formal rules. J Am Stat Assoc. 1996;91(435):1343–70. https://doi.org/10.1111/rssa.12378.
42. Gelman A. Prior distributions for variance parameters in hierarchical models. Bayesian Anal. 2006;1(3):515–34. https://doi.org/10.1214/06-BA117A.
43. Casella G, Moreno E. Intrinsic meta-analysis of contingency tables. Stat Med. 2005;24(4):583–604. https://doi.org/10.1002/sim.2038.
44. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2020. http://www.r-project.org/index.html.
45. NYU Langone Health. NYU High performance computing core. 2021. https://med.nyu.edu/research/scientific-cores-shared-resources/high-performance-computing-core. Accessed 21 Nov 2022.

Publisher's Note