A Bayesian Interpretation of Subgroup Analyses: RECOVERY Trial on Tocilizumab and COVID-19*

Arthur M. Albuquerque *Universidade Federal do Rio de Janeiro, Brazil*
Lucas Tramujas, MD *Research Institute, HCor-Hospital do Coração, Brazil*
Lorenzo R. Sewanan, MD, PhD *Columbia University, U.S.A.*
James M. Brophy, MD, PhD *McGill University Health Center, Canada*

Introduction

The Randomised Evaluation of COVID-19 Therapy (RECOVERY) is a platform trial testing several treatments for COVID-19 since the beginning of the pandemics in 2020. Preliminary results on the anti-inflammatory drug tocilizumab were published on medRxiv on February 11, 2021 (Horby 2021). The RECOVERY was a large, open-label randomized controlled trial (RCT) that assessed 28-day mortality as the primary outcome in hospitalized patients with COVID-19 using a frequentist statistical approach. Secondary outcomes were discharge from hospital within 28 days in all patients and risk of invasive mechanical ventilation (IMV) or death among those not receiving IMV at baseline. Overall, tocilizumab was effective for these outcomes.

The authors also reported several pre-specified subgroups analyses. Two clusters of subgroups are of substantial clinical interest: respiratory support and days since symptom onset. For COVID-19 infection, duration of disease and use of oxygen are crucial to stratify the risk of patients. Respiratory support should be interpreted as a direct proxy for the severity of the disease. Patients on invasive mechanical ventilation are naturally at greater risk of death than patients on only simple oxygen. Previous research has shown that dyspnea most commonly develops after a median of 5-7 days since the onset of symptoms (Wang et al. 2020).

It has been postulated that the severity of COVID-19 disease is closely related to the degree of inflammation and disease length (Wiersinga et al. 2020). For example, the results of another anti-inflammatory drug – dexamethasone – showed strikingly different results between subgroups (Group 2021). Thus, one could also expect different results between the subgroups in the tocilizumab trial. However, this trial's subgroup results were consistent with each other.

Subgroup analyses are fundamental to clinical practice. These analyses inform clinicians about the effect of treatments in specific subsets of patients and thus yield to the opportunity of precision medicine. They also allow one to explore or confirm hypotheses. In the context of tocilizumab and COVID-19, it is clinically and economically important to determine if there is sufficient discrepancy to ascertain different treatments for specific subgroups.

Notably, frequentist subgroup analyses pose several limitations. Their low statistical power increases both type 2 error, leading to false-negatives results (Schulz and Grimes 2005), and type 1 error risk, inflating treatment effect (Ioannidis 2005). Estimations suggest that clinical trials designed to have 80% power for the primary outcome are anticipated to have 30% power for the interaction term (Kent, Steyerberg and van Klaveren 2018). Thus, one could quickly draw misleading conclusions about the effect of tocilizumab in specific subgroups if these analyses are not appropriately interpreted.

Also, it is of utmost importance to define if results from a study are clinically meaningful. The frequentist statistical inference relies on p-values and confidence intervals (CIs) to report re-

*April 12, 2021

sults. Unfortunately, p-values only provide information about the data's compatibility to the null hypothesis (usually zero effect). It is impossible to infer whether the actual effect size is large or what is the probability of the alternative hypothesis (result different from zero) of being true based on p-values (Rafi and Greenland 2020). Also, the frequentist approach to subgroup analysis can lead to multiplicity issues (Schulz and Grimes 2005). CIs can improve the interpretation of results beyond p-values. However, they are also easily misinterpreted. CIs contain a range of parameter values that are more compatible with the data than are values outside the interval (Rafi and Greenland 2020). CIs do not indicate what the probability of a range of effects is (e.g., RR < 1, 0.9, 0.8 etc.) to be true. Thus, the frequentist statistical framework poorly aids clinicians to analyze data from subgroups in RCTs thoroughly.

Bayesian re-analysis

We believe that Bayesian analyses can clarify the interpretation of the RECOVERY trial (Wijeysundera et al. 2009). In brief, the Bayesian statistical framework relies on the use of *priors* that represent our belief in the actual effect. Then, these are combined with data (*likelihood*), generating *posterior probabilities* (PP). PP can be described with credible intervals, which, in contrast to confidence intervals, actually inform the probability of an effect within a specified interval. We can also perform multiple looks at the PP without worrying about multiplicity issues. Because of this, Bayesian analyses allows one to assess the probability of numerous effect sizes at once (Spiegelhalter, Abrams and Myles 2004).

Here, we aim to describe the priors we will use in our re-analysis of the RECOVERY trial on tocilizumab. We will re-analyze the primary outcome – 28-day mortality – and a secondary outcome, risk of being discharged alive from the hospital within 28 days. We will analyze these outcomes in all patients and each subgroup regarding respiratory support (no ventilator support, non-invasive mechanical ventilation, and invasive mechanical ventilation) and days since symptom onset ($\leq$ 7 days and $>$ 7 days). We will primarily report the risk difference between the tocilizumab and control groups.

Methods

We will extract data from RCTs on tocilizumab and COVID-19 included in a Cochrane's systematic living review (Ghosn et al. 2021). Then, we will use these results to create evidence-based priors (EBP) for each subgroup if specific data is available. We will also pool data from every subgroup to create an EBP for all patients. In case data is not available for a particular subgroup, we will use the pooled data as the EBP for these subgroups. To dampen this adaptation's influence on the final results, we will also analyze these specific subgroups with both 10% and 50% weight of this adapted prior (Brophy 1995; Goligher et al. 2018). We described the entire data extraction protocol elsewhere (https://osf.io/upd4q/).

Thus, every prior will be normally distributed and described by a *mean* and a *variance*. To test our results' robustness, we will also perform sensitivity analyses using different priors (Sung et al. 2005; Zampieri et al. 2021). We will create four extra priors: skeptical, optimistic, pessimistic, and non-informative. While the latter will have the same parameters (*mean* and *variance*) for every subgroup, we will adapt each subgroup's original EBP to create priors accordingly (White, Pocock and Wang 2005). The skeptical prior uses the EBP's *variance*, but the *mean* equals 0. The optimistic prior uses the EBP's *variance*, but the *mean* equals -0.05. Lastly, the pessimistic prior uses the EBP's *variance*, but the *mean* equals 0.05. We decided on the absolute *mean* value of 0.05 because the RECOVERY group estimated risk difference would be 0.05 for their sample size calculation

(Group 2021; Zampieri et al. 2021). We note that these are values for the analyses regarding the mortality outcome because negative results mean tocilizumab is better. Thus, we will use opposite mean values for the secondary outcome because, in this case, positive results mean tocilizumab is better.

A simulation

To aid in the interpretation of our methods, we simulated data to create multiple priors. In this example, we analyzed the mortality outcome of a fictitious subgroup. The evidence-based prior for this subgroup is described by a *mean* of -0.02 and *standard deviation* of 0.04. We then created the priors following the instructions mentioned above. These priors can be visualized in *Figure 1* and their parameters in *Table 1*.

Conclusion

Here, we discussed the rationale and the methods of a pre-planned Bayesian re-analysis of the RECOVERY trial on tocilizumab and COVID-19.

Table 1: Parameters of priors

			Probability of Treatment Effect		
Prior Belief	Mean RD	SD of RD	Pr(< 0)	Pr(< -0.05)	Pr(< -0.10)
Evidence-based	-0.02	0.04	0.69	0.23	0.02
Skeptical	0.00	0.04	0.50	0.11	0.01
Optimistic	-0.05	0.04	0.89	0.50	0.11
Pessimistic	0.05	0.04	0.11	0.01	0.00
Non-informative	0.00	10.00	0.50	0.50	0.50

Note:
SD = Standard Deviation; RD = Risk Difference

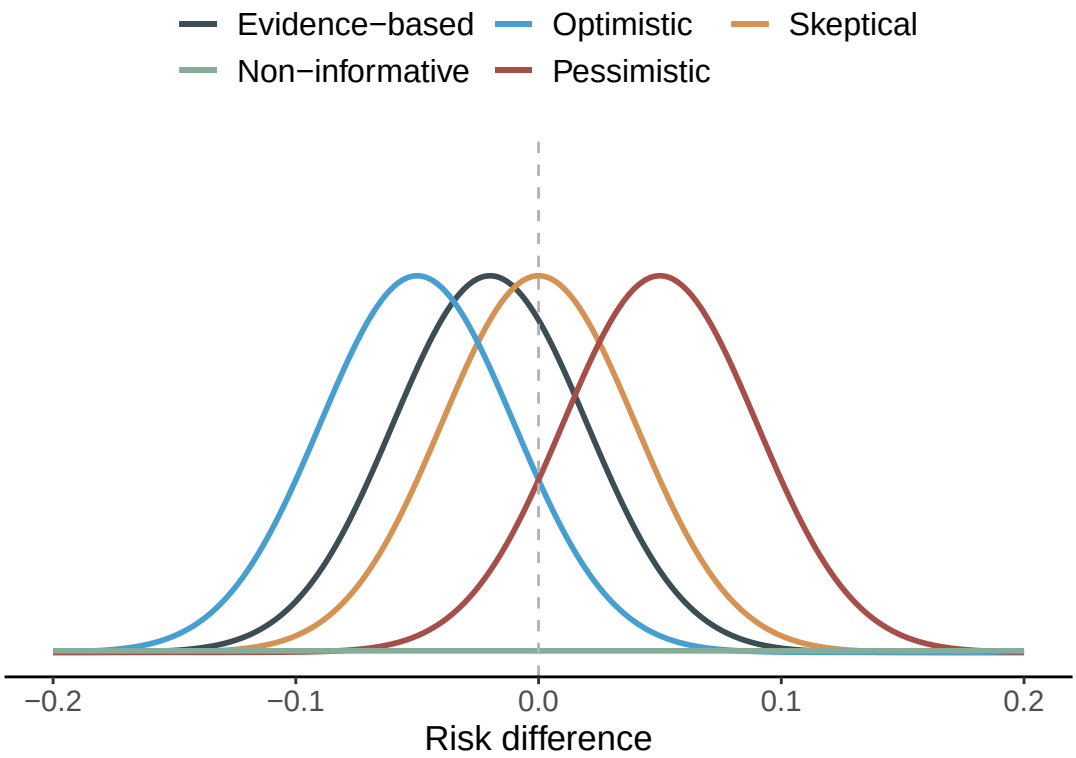

Figure 1: Simulation of priors derived from a fictitious evidence-based prior. Negative values mean tocilizumab is better.

References

Brophy, James M. 1995. "Placing Trials in Context Using Bayesian Analysis: GUSTO Revisited by Reverend Bayes." *JAMA* 273(11):871.

Ghosn, Lina, Anna Chaimani, Theodoros Evrenoglou, Mauricia Davidson, Carolina Graña, Christine Schmucker, Claudia Bollig, Nicholas Henschke, Yanina Sguassero, Camilla Hansen Nejstgaard, Sonia Menon, Thu Van Nguyen, Gabriel Ferrand, Philipp Kapp, Carolina Riveros, Camila Ávila, Declan Devane, Joerg J. Meerpohl, Gabriel Rada, Asbjørn Hróbjartsson, Giacomo Grasselli, David Tovey, Philippe Ravaud and Isabelle Boutron. 2021. "Interleukin-6 Blocking Agents for Treating COVID-19: A Living Systematic Review." *Cochrane Database of Systematic Reviews* (3).

Goligher, Ewan C., George Tomlinson, David Hajage, Duminda N. Wijeysundera, Eddy Fan, Peter Jüni, Daniel Brodie, Arthur S. Slutsky and Alain Combes. 2018. "Extracorporeal Membrane Oxygenation for Severe Acute Respiratory Distress Syndrome and Posterior Probability of Mortality Benefit in a Post Hoc Bayesian Analysis of a Randomized Clinical Trial." *JAMA* 320(21):2251–2259.

Group, RECOVERY Collaborative. 2021. "Dexamethasone in Hospitalized Patients with Covid-19." *New England Journal of Medicine* 384(8):693–704.

Horby, Peter W. 2021. "Tocilizumab in Patients Admitted to Hospital with COVID-19 (RECOVERY): Preliminary Results of a Randomised, Controlled, Open-Label, Platform Trial." *medRxiv* p. 2021.02.11.21249258.

Ioannidis, John P. A. 2005. "Why Most Published Research Findings Are False." *PLOS Medicine* 2(8):e124.

Kent, David M., Ewout Steyerberg and David van Klaveren. 2018. "Personalized Evidence Based Medicine: Predictive Approaches to Heterogeneous Treatment Effects." *BMJ* 363:k4245.

Rafi, Zad and Sander Greenland. 2020. "Semantic and Cognitive Tools to Aid Statistical Science: Replace Confidence and Significance by Compatibility and Surprise." *BMC Medical Research Methodology* 20(1):244.

Schulz, Kenneth F. and David A. Grimes. 2005. "Multiplicity in Randomised Trials II: Subgroup and Interim Analyses." *The Lancet* 365(9471):1657–1661.

Spiegelhalter, D. J., K. R. Abrams and Jonathan P. Myles. 2004. *Bayesian Approaches to Clinical Trials and Health Care Evaluation*. Statistics in Practice Chichester ; Hoboken, NJ: Wiley.

Sung, Lillian, Jill Hayden, Mark L. Greenberg, Gideon Koren, Brian M. Feldman and George A. Tomlinson. 2005. "Seven Items Were Identified for Inclusion When Reporting a Bayesian Analysis of a Clinical Study." *Journal of Clinical Epidemiology* 58(3):261–268.

Wang, Dawei, Bo Hu, Chang Hu, Fangfang Zhu, Xing Liu, Jing Zhang, Binbin Wang, Hui Xiang, Zhenshun Cheng, Yong Xiong, Yan Zhao, Yirong Li, Xinghuan Wang and Zhiyong Peng. 2020. "Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China." *JAMA* 323(11):1061–1069.

White, Ian R., Stuart J. Pocock and Duolao Wang. 2005. "Eliciting and Using Expert Opinions about Influence of Patient Characteristics on Treatment Effects: A Bayesian Analysis of the CHARM Trials." *Statistics in Medicine* 24(24):3805–3821.

Wiersinga, W. Joost, Andrew Rhodes, Allen C. Cheng, Sharon J. Peacock and Hallie C. Prescott. 2020. "Pathophysiology, Transmission, Diagnosis, and Treatment of Coronavirus Disease 2019 (COVID-19): A Review." *JAMA* 324(8):782–793.

Wijeysundera, Duminda N., Peter C. Austin, Janet E. Hux, W. Scott Beattie and Andreas Laupacis. 2009. "Bayesian Statistical Inference Enhances the Interpretation of Contemporary Randomized Controlled Trials." *Journal of Clinical Epidemiology* 62(1):13–21.e5.

Zampieri, Fernando G., Jonathan D. Casey, Manu Shankar-Hari, Frank E. Harrell and Michael O. Harhay. 2021. "Using Bayesian Methods to Augment the Interpretation of Critical Care Trials. An Overview of Theory and Example Reanalysis of the Alveolar Recruitment for Acute Respiratory Distress Syndrome Trial." *American Journal of Respiratory and Critical Care Medicine* 203(5):543–552.