The Association of Guideline Directed Prophylaxis with the Incidence of Postoperative Nausea and Vomiting in Pediatric Patients[*]

Proposal for A Multicenter Retrospective Observational Cohort Study

Benjamin Y. Andrew, MD, MHS [ID][†]
Duke University School of Medicine; Department of Anesthesiology
Critical Care and Perioperative Population Health Research (CAPER) Group

Lucy Everett, MD
Massachusetts General Hospital, Department of Anesthesia, Critical Care and Pain Medicine

T. Wesley Templeton, MD
Wake Forest University School of Medicine, Department of Anesthesiology

Timothy T. Houle, PhD [ID]
Massachusetts General Hospital, Department of Anesthesia, Critical Care and Pain Medicine

Lisa N. Vitale, MD
University of Michigan, Department of Anesthesiology

Vikas O'Reilly-Shah, MD, PhD [ID]
University of Washington, Department of Anesthesiology and Pain Medicine

Meredith Bailey, MSN, RN
Multicenter Perioperative Outcomes Group

Vijay Krishnamoorthy, MD, PhD [ID]
Duke University School of Medicine; Department of Anesthesiology
Critical Care and Perioperative Population Health Research (CAPER) Group

Ashraf Habib, MBBCh, MHS [ID]
Duke University School of Medicine, Department of Anesthesiology

Brad Taicher, DO, MBA [ID]
Duke University School of Medicine, Department of Anesthesiology

---

[*]This is a preliminary proposal prepared for review by the Perioperative Clinical Research Committee (PCRC).
[†]Corresponding author (benjamin.andrew@duke.edu).

ABSTRACT   Postoperative nausea and vomiting (PONV) is a frequent complication following general anesthesia in pediatric patients. PONV is associated with a number of negative patient outcomes including dehydration, pulmonary aspiration, delayed mobilization, risk for wound dehiscence, delayed discharge, and unanticipated hospital admission. Given its high incidence and associated costly outcomes, a systematic approach to reducing PONV through appropriate prophylaxis represents a key opportunity for optimization of patient care. Consensus guidelines for prophylaxis against PONV in pediatric patients recommend a risk-based approach, with escalation in the number of antiemetics based on a preoperative estimate of PONV risk. These recommendations have been translated into performance metrics, most notably by the Multicenter Perioperative Outcomes Group (MPOG) through the Anesthesiology Performance Improvement and Reporting Exchange (ASPIRE), which currently includes over 25 children's hospitals. Despite widespread dissemination of this guideline-defined quality metric for PONV prophylaxis, the association between adherence to such a risk-based approach and clinical outcomes remains unknown. Further, while the consensus guidelines are based on a comprehensive collation of evidence surrounding PONV risk prediction and prophylaxis, this evidence base is not robust. Most notably, risk prediction in pediatric PONV is sparsely studied, and available risk models utilize suboptimal methods that leave them prone to misclassification and bias. Thus, it is important to objectively estimate how published national guidelines, based on the current landscape of pediatric PONV risk prediction, are associated with clinical outcomes.

KEYWORDS   postoperative nausea and vomiting; pediatric anesthesioloy; MPOG

Introduction

Postoperative nausea and vomiting (PONV) is a frequent complication following general anesthesia in pediatric patients, with a reported incidence often exceeding that in the adult population.[1,2] Further, PONV is associated with a number of negative patient outcomes including dehydration, pulmonary aspiration, delayed mobilization, risk for wound dehiscence, delayed discharge and unanticipated hospital admission.[2,3]

Given its high incidence and associated costly outcomes, a systematic approach to reducing PONV through appropriate prophylaxis represents a key opportunity for optimization of patient care. Consensus guidelines for prophylaxis against PONV in pediatric patients recommend a risk-based approach, with escalation in the number of antiemetics based on a preoperative estimate of PONV risk.[4] These recommendations have been translated into performance metrics, most notably by the Multicenter Perioperative Outcomes Group (MPOG) through the Anesthesiology Performance Improvement and Reporting Exchange (ASPIRE), which currently includes over 25 children's hospitals.[5]

However, despite widespread dissemination of this metric, the impact of this approach on clinical outcomes is not known. Further, while the consensus guidelines are based on a comprehensive collation of evidence surrounding PONV risk prediction and prophylaxis, this evidence base is not robust. Most notably, risk prediction in pediatric PONV is sparsely studied, and available risk models utilize suboptimal methods that leave them prone to misclassification and bias. Thus, it is important to objectively estimate how published national guidelines, based on the current landscape of pediatric PONV risk prediction, are associated with clinical outcomes. In a preliminary single center analysis (discussed in detail in later sections), there was evidence for a modest relative and absolute benefit of adherence to guideline directed prophylaxis although this was a heterogeneous effect that appeared to differ across the risk spectrum.[6] The study proposed herein will aim to more rigorously investigate the association between adherence to guideline recommended risk-based PONV prophylaxis and the incidence of PONV in pediatric patients.

More specifically, our primary research question is:

Does adherence to PONV prophylaxis guidelines (as defined by MPOG-ASPIRE process measure PONV-04) reduce the incidence of PONV (as defined by MPOG-ASPIRE outcome measure PONV-03) in pediatric patients undergoing general anesthesia?

It follows, then, that the primary aim of the proposed analysis is to estimate the causal effect of adherence to PONV prophylaxis guidelines on the incidence of PONV in the post-anesthesia recovery unit.

Methods

Study Design

This is a retrospective cohort study with IRB approval from the Duke University Health System (Pro00112464). All future documentation will adhere to reporting guidelines as outlined by the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement[7] as well as guidelines for the conduct and reporting of Bayesian analyses.[8,9] In an effort to promote open science and reproducibility, the majority of the data processing and analytic scripts used for this project will be maintained at: https://github.com/andrew10043/mpog_peds_ponv.

Study Population

We will include all patients age 3-17 undergoing general anesthesia at participating institutions from January 1st, 2015 through December 31st, 2023. Complete inclusion and exclusion criteria are outlined below:

- Inclusion criteria
  - Procedure date from 01/01/2018 - 12/31/2023
  - General anesthetic encounter ≥ 15 minutes
  - Age at time of procedure ≥ 3 years
- Exclusion criteria
  - Cardiothoracic surgery
  - Cardiac catheterization
  - Liver transplantation
  - Lung transplantation
  - American Society of Anesthesiologists (ASA) physical status classification ≥ 5
  - Postoperative ICU admission
  - Electroconvulsive therapy
  - Labor epidural encounters
  - Cesarean deliveries
  - Radiation oncology procedures
  - Radiology procedures
  - Non-operative cases
  - All additional cases marked as *excluded* from either PONV-04 or PONV-03 metrics
  - All cases with missing PONV-03 outcome data

Data Source

We will use data from the MPOG database only, without any linkage to additional data sources.

Primary Outcome

The primary outcome is PONV, as defined by the PONV-03 outcome metric. This is a dichotomous measure, with possible values of **passed** (indicating no PONV) or **failed** (indicating PONV). Specifically, a value of **failed** is assigned for all patients who either have documented nausea or emesis *or* receive a rescue antiemetic in the immediate postoperative period, defined as the period from PACU start through 6 hours after anesthesia end time.[5]

Full details for the PONV-03 metric can be found at: https://spec.mpog.org/Spec/Public/33.

Secondary Outcomes(s)

Given that the administration of rescue antiemetics in the PACU may occur for reasons other than evidence of active PONV (e.g., continued *prophylaxis* for a patient assumed to be at high risk, pre-discharge *"top-up"* for outpatient surgical cases, etc.), we plan to examine the primary outcome restricted to cases only with **documented** PONV. However, as institutional documentation practices also may vary substantially, this secondary outcome metric will only be considered following a review of its use and incidence patterns in the multicenter cohort. In tandem with this secondary outcome, we will also examine the distribution of antiemetic administration timing, as this may further inform the patterns of antiemetic rescue vs. continued PACU *prophylaxis* in certain cases.

Exposure Variable

The exposure variable of interest is compliance with PONV prophylaxis guidelines as defined by the PONV-04 measure. A patient will be considered to have received **compliant prophylaxis** if he or she receives *at least* the recommended number of prophylactic antiemetic agents, with **noncompliant prophylaxis** representing any number of agents less than the recommended value. The PONV-04 metric considers compliance as: ≥ 1 agent for patients with 1 risk factor, ≥ 2 agents for patients with 2 risk factors, and ≥ 3 agents for patients with ≥ 3 risk factors.[5]

Risk factors are defined as follows:

- Age ≥ 3 years
- Female patient ≥ 12 years
- History of PONV in the patient, parent or sibling
- High risk surgery (strabismus, adenotonsillectomy, tympanoplasty)
- Inhaled anesthetic duration ≥ 30 minutes
- Administration of long acting opioids

To be considered as a valid antiemetic agent towards metric compliance, each agent must come from a different pharmacologic class:

- 5-HT3 receptor antagonists
- Anticholinergic agents
- Antihistamines
- Butyrophenones
- NK-1 receptor agonists
- Phenothiazines
- Steroids
- Prokinetic agents
- Propofol (infusion only)

Full details for the PONV-04 metric can be found at: https://spec.mpog.org/Spec/Public/48

Additionally, there is a small discrepancy between the definition of compliant prophylaxis in the published consensus guidelines and the implementation of these guidelines in MPOG-ASPIRE. The written guidelines[4] recommend that patients with 1-2 risk factors receive two agents while the MPOG interpretation of this metric separates this group further into 1 risk factor (1 agent recommended) and 2 risk factors (2 agents recommended). In our single center work this did not lead to an appreciable difference in outcomes, but we will plan to include a sensitivity analysis using this alternate definition of exposure.

Covariates

Covariates are outlined in the tables below. Each name is linked to its respective MPOG phenotype document, if applicable, for further details. All continuous variables will be preserved without categorization (with exceptions for the MPOG-defined dichotomous risk factors).

Patient Level Variables

Each patient is identified using a unique MPOG ID. Unless necessary due to small sample size, we will retain all original race categories as reported in the database.

Table 1: Basic Patient Level Variables (excluding PONV-related concepts)

Variable	Type	Definition / Description
MPOG Patient ID	Character string	Unique patient identifier
Age	Continuous	Patient age in years
Race	Nominal	Race or Ethnicity of the patient
Sex	Nominal	Documented sex of the patient at the time of the case
ASA Physical Status Classification	Ordinal	ASA classification

Center Level Variables

Each institution will be represented by a unique ID. Medical school and pediatric hospital affiliation status, along with bed size, will be used as the primary center level variables for modeling purposes.

Table 2: Center Level Variables

Variable	Type	Definition / Description
Institution	Character string	Unique institution identifier
Medical School Affiliation	Nominal	Indicates whether the case occurred at an institution affiliated with a medical school
Pediatric Hospital Affiliation	Nominal	Indicates whether the case occurred at a Children's Hospital affiliated with an active MPOG institution
Hospital Bed Size	Ordinal	Categorical value indicating the the bed-size for each hospital

Procedure Level Variables

Surgical procedures will be classified by surgical service (e.g., general surgery, neurosurgery, ENT, etc.) tags, with procedure text used to clarify any discrepancies and to further exclude specific procedure types. The *Anesthesia Technique* variables and *Airway Type* will be used to categorize cases based on airway management and anesthetic approach (e.g., GETA with epidural, LMA with block, etc.).

Table 3: Procedure Level Variables

Variable	Type	Definition / Description
Year of Procedure	Ordinal	Based on surgery start date/time.
Case Duration	Continuous	Duration of time a patient is anesthetized (in minutes)
Procedure	Character string	Textual description of the procedure
Procedure Type: Adenotonsillectomy	Nominal	Determines whether the case includes Tonsilectomy, Adenoidectomy, Adenotonsillectomy (or control of postop tonsil bleed)
Procedure Type: Strabismus	Nominal	Determines whether a case was a Strabismus procedure
Procedure Type: Tympanoplasty	Nominal	Determines if a case includes tympanoplasty
Surgical Service	Nominal	Surgical service assigned to the case
Postoperative Destination	Nominal	Postoperative destination classification (e.g., transported to ICU, PACU, etc.)
Airway Type	Nominal	Returns the airway type assigned to a case based on collation mapping of EMR data (e.g., ETT, LMA, mask, MAC-sedation, etc.)
Anesthesia Technique: General	Nominal	This phenotype attempts to classify each case in terms of general anesthesia technique (e.g., "no", "General-ETT", "General-LMA", etc.)
Anesthesia Technique: Neuraxial	Nominal	Identifies the use of neuraxial anesthesia and the type, if used
Anesthesia Technique: Peripheral Nerve Block	Nominal	Identifies the use of a peripheral nerve block and the type, if used
Anesthesia Technique: Sedation	Nominal	Indicator variable for MAC cases with sedation
Halogenated Anesthetic Gas Use	Nominal	Specifies if (and when) halogenated anesthetic gases were administered during a case
Nitrous Oxide Use	Nominal	Specifies if (and when) nitrous oxide was administered during a case

Clinician Level Variables

The *primary provider* phenotype will be used to provide a unique attending ID for multilevel modeling purposes. It will also be used to classify the in-room clinician for cases as either *resident* or *non-resident* (CRNA or AA).

Variable	Type	Definition / Description
Primary Provider	Nominal	Determines the anesthesia attending and non-attending anesthesia providers signed in for the longest duration of a case

PONV Related Variables

Variables outlined in the table below will be sourced from the PONV-03 and PONV-04 data files and many do not have specific MPOG phenotypes.

Table 5: PONV Related Variables

Variable	Type	Definition / Description
PONV-03 Result	Nominal	Primary outcome, as described in previous sections.
PONV-04 Result	Nominal	Primary exposure, as described in previous sections.
PONV-03 Result Reason	Nominal	Explanatory variable for PONV-03 result (e.g., reasons for exclusion or failure)
PONV-04 Result Reason	Nominal	Explanatory variable for PONV-04 result (e.g., reasons for exclusion or failure)
Prophylactic Agents	Character string	Character vector of all prophylactic antiemetic agents administered.
Risk Factor: Age ≥ 3	Nominal	Indicator variable for patients with age ≥ 3 years
Risk Factor: Female ≥ 12	Nominal	Indicator variable for female patients ≥ 12 years old
Risk Factor: History	Nominal	Indicator variable for patients with personal or family history of PONV
Risk Factor: Volatile Duration	Nominal	Indicator variable for patients with volatile anesthetic exposure ≥ 30 minutes
Risk Factor: Opioid 3	Nominal	Indicator variable for patients who are administered long-acting opioids
Risk Factor: Procedure	Nominal	Indicator variable for patients undergoing a high risk procedure (adenotonsillectomy, tympanoplasty, strabismus)

Statistical Analysis

Overview and Approach

As we aim to estimate the causal effect of compliant prophylaxis on the incidence of PONV, there are several methodological challenges we must first consider. Chief among these is the presence of confounding, particularly confounding by indication, that will severely bias the estimate of interest if not accounted for. Propensity score (PS) weighting is one of many causal inference tools that is capable of controlling for *measured* confounding in the estimation of the causal effect of interest. Several weighting approaches have been described, two of which we will explore in the proposed analysis: weighting using (1) inverse

probability of treatment weights (IPTW) and (2) overlap weights (OW). These weights differ in their approach to the construction of *pseudo-populations* and, in turn, represent approaches to two different causal estimands. Weighting using IPTW, which assigns weights of $\frac{1}{PS}$ to treated patients and $\frac{1}{1-PS}$ to untreated patients, targets the *average treatment effect* (ATE) - the estimand most often used in randomized trials. Alternatively, weighting using overlap weights, which assigns weights of $1 - PS$ for treated patients and PS to untreated patients, targets the ATE in the *overlap population* - the group of patients with similar propensities to receive or not receive the exposure (i.e., those with *clinical equipoise*).

The typical approach to confounder adjustment using propensity score weighting involves the sequential fitting of two frequentist statistical models. A design stage logistic regression model which is used to generate point estimates of propensity scores for each patient, which are then used to weight observations in the subsequent outcome model. Given the many benefits of Bayesian inference, an attractive alternative to this frequentist approach is to combine propensity score weighting with Bayesian outcome modeling. Among other potential advantages, this approach would allow for the propagation of uncertainty in the estimation of the propensity score model through to the outcome model. This represents an area of considerable debate, and while several approaches have been explored, *Liao* and *Zigler*[10] outlined a *pseudo-Bayesian* two-stage approach in which a propensity score model is estimated, and the outcome model is then fit many times using sampled weights from the posterior distribution of the propensity score model. Alternatively, the outcome model can iteratively sample from the posterior distribution of weights at each step in the MCMC algorithm, reducing the need to fit many thousands of outcome models. Unfortunately, even the latter approach becomes computationally difficult with larger samples sizes, particularly those present in large multicenter databases.

More recently, *Nafa* and *Heiss*[11] have outlined an approach aimed and circumventing this issue by using only the location and scale of the propensity score weights in the outcome model and propagating the "uncertainty by placing a prior on the scale component of the weights." Here, they parameterize propensity score weights as:

$$\tilde{w}_i = \kappa_i + \lambda_i \cdot \pi(\lambda)$$

where κ and λ represent the location (typically the mean), and the scale of the posterior distribution of propensity score weights, respectively. The $\pi(\lambda)$ term then represents a vector of length n sampled at each iteration of the MCMC algorithm from the defined prior on the scale component. This approach both allows for consideration of uncertainty in the propensity score weights while both regularizing their variance and not requiring direct use of the (very large) matrix of propensity scores generated from the design stage model for studies with large N. The authors recommend a "weakly to moderately informative" prior distribution for $\pi(\lambda)$, for example *Exponential* (4).

Propensity Score Model

Propensity score estimation in the setting of clustered data represents represents a unique challenge. First, we note that the clustering structure of the data for our particular question is *nested*, in that we assume that patients are fully nested within anesthesiologists (clinicians) who are in turn fully nested within hospitals. Second, we must define the nature of the clustering as either *central* (i.e., the mechanism of treatment allocation varies across clusters) or *incidental* (i.e., the mechanism of treatment allocation is consistent across clusters).[12] In the setting of anesthesiologists nested within hospitals, there is almost certainly some aspect of central clustering whereby certain characteristics (both measured and unmeasured) differentially influence the provision of compliant prophylaxis across clusters. Third, we must consider the relative number and sizes of the clusters. In our context, we will have upwards of >30 hospital clusters, each

with likely several thousand patients. However, clinician-level clusters may be more variable, with the possibility of many smaller clusters. This may pose a problem when estimating certain types of models, as discussed shortly. Generally speaking, three main approaches to estimating propensity scores using clustered data are described:

In a **single-level** propensity score model, the clustered structure is ignored, and the propensity score is estimated as:

$$logit(p_{ich}) = \beta_0 + \beta_1 X_{ich} + \beta_2 K_{ch} + \beta_3 Z_h$$

where, β_0 is a fixed intercept, and β_1, β_2, and $\beta3$ are fixed coefficients for individual-level (X_{ich}), clinician-level (K_{ch}) and hospital-level (Z_h) covariates. Implicit in this model structure is the assumption that no confounding persists following conditioning on observed covariates (X_{ich}, K_{ch}, and Z_h). However this is unlikely, particularly in the setting of *central* clustering.[12]

In a **fixed effects** propensity score model, cluster-specific fixed intercepts are included:

$$logit(p_{ich}) = \beta_0 + \beta_{0c} + \beta_{0h} + \beta_1 X_{ich}$$

where, β_0 is a fixed population intercept, and β_{0c} and β_{0h} are fixed intercepts (i.e., coefficients for cluster-level indicator variables) for hospital and clinician-clusters. This can be extended to include interactions between cluster membership and individual-level covariates (or across-cluster interactions). These models do not explicitly include cluster-level covariates (K_{ch}, and Z_h) as the intercepts fully absorb the effects of both observed and unobserved covariates at the cluster level (e.g., all cluster-specific covariates in a given cluster will be identical). However, when extended to this three-level example, this model structure does not easily take into account the nested nature of the clustered data (e.g., clinicians within hospitals) and requires estimation of a very large number of parameters. Further, at the clinician level there will be many very small clusters, adding to the difficulty in fitting a fixed effect term at this level.

Finally, in a **random effects** propensity score model, cluster-specific random intercepts are included:

$$logit(p_{ich}) = \beta_{0ch} + \beta_1 X_{ich} + \beta_2 K_{ch} + \beta_3 Z_h$$
$$\beta_{0ch} = \delta_{00h} + U_{0ch}$$
$$\delta_{00h} = \gamma_{000} + V_{00h}$$

where, β_{0ch} is the hospital- and clinician-specific intercept that is composed of: the fixed intercept (γ_{000}), the hospital-level random intercept (V_{00h}) and the clinician-level random intercept (U_{0ch}), with the latter two assumed to follow ~ *Normal* $(0, \sigma^2)$. This can be extended to include random cluster-specific slopes. Here, the random intercepts capture the effects of unmeasured covariates at the cluster level. However, random effects models make the added assumption that the random effects are *uncorrelated* with any covariates (e.g., X_{ich}, K_{ch}, and Z_h) in the model. If there are unmeasured cluster-level covariates that are correlated with these covariates, this assumption will be violated and can lead to bias in estimation of propensity scores. Further, because of shrinkage of the random intercepts towards zero, there may be more residual imbalance *within* clusters, and models are more reliant on capturing and including cluster-level covariates.

A number of authors have explored the topic of propensity score model construction in clustered data, using simulations to assess various combinations of approaches with respect to their ability to generate

unbiased estimates. *Li* and colleagues showed that estimators ignoring clustering in *both* the propensity score model and outcome model had much larger bias than those considering clustering in at least one stage.[13] However, ignoring clustering in the outcome model appeared to be more detrimental than ignoring it in the propensity score model. Separately, *Fuentes* and colleagues extensively examined the properties of propensity score modeling approaches in multilevel data, including the application of various weighting techniques.[14] They showed that both fixed effects and *correctly specified* (i.e., inclusion of all relevant cluster-specific variables) random effects approaches to estimating propensity scores yielded unbiased results, however there is increased bias with a misspecified random effects model. With respect to covariate-by-cluster interactions (i.e., differential treatment allocation mechanisms across clusters), random slope/intercept models using overlap weights were most successful, however fixed effect approaches with overlap weights were also adequate when cluster size was large. Importantly, these simulations did not include correlated unmeasured cluster-level confounders, which would theoretically bias the random effects model.[14]

Taken together, evidence from simulation studies appears to favor a fixed effects approach to specification of the propensity score model, as this avoids the likely scenario in which unmeasured cluster-level confounders lead to a biased estimate when using random effects models. However, a strictly fixed effects approach in the proposed setting is relatively unattractive given the fully nested nature of the clusters and the large number of clinician-level clusters, many of which may be quite small. One alternative may be a hybrid approach, in which hospital is modeled as a fixed effect, with nested random effects for clinician within hospitals. This allows for the advantage of fixed effects at the hospital-level (where unobserved cluster-level confounding is arguably more likely) while respecting the nested nature of the data generating process. Such a parameterization is shown below:

$$logit\,(p_{ich}) = \beta_{0ch} + \beta_1 X_{ich} + \beta_2 K_{ch} + \beta_3 Z_h$$
$$\beta_{0ch} = \delta_{00h} + U_{0ch}$$
$$\delta_{00h} = \gamma_{000} + \theta_{00h}$$

where, β_{0ch} is the hospital- and clinician- specific intercept which is, in turn composed of the population fixed intercept (γ_{000}), hospital-specific fixed intercept (θ_{00h}), and clinician-specific random intercept (U_{0ch}). The remaining parameters are as previously described.

Importantly, this group is separately submitting a PCRC proposal for a descriptive analysis of PONV prophylaxis compliance in which the relationships between individial-, clinician-, and hospital-level variables with compliant prophylaxis allocation will be extensively explored. Results from this modeling will be illuminating with respect to the ideal propensity score modeling approach and will help guide our eventual selection.

Variable selection for inclusion in the propensity score model (and outcome model) will be guided by construction of a causal directed acyclic graph (DAG) outlining the proposed relationships between the exposure, outcome, and covariates of interest (Figure 1). While the DAG below represents a starting point, we will use information gathered from the afformentioned descriptive analysis to understand more about variables associated with PONV prophylaxis compliance and further expand this figure. Prior to propensity score weighting, covariates of interest will be compared across cohorts (compliant vs. non-compliant prophylaxis) using standardized mean differences.
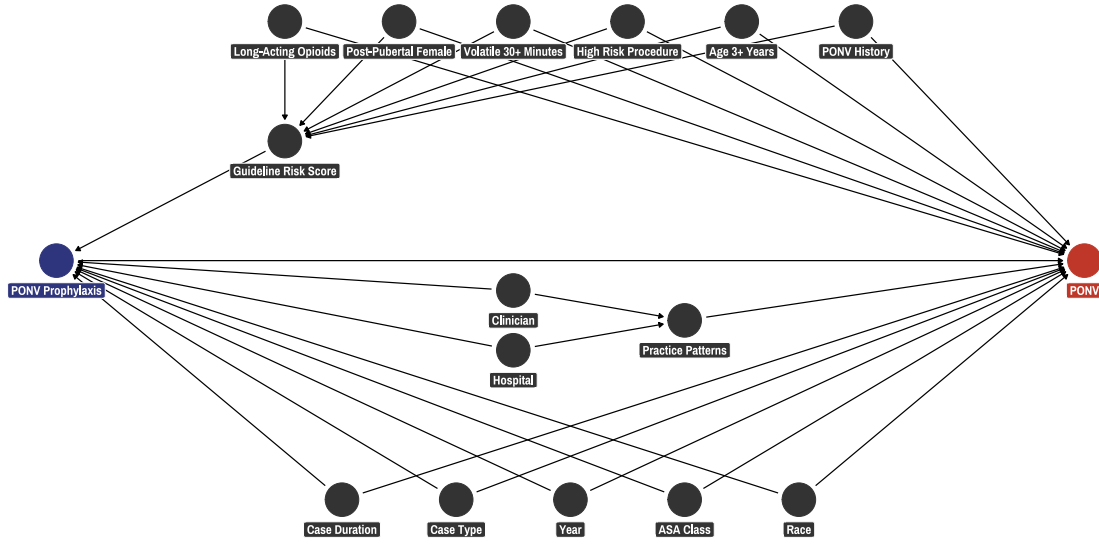
Figure 1: DAG Showing proposed relationships amongst the exposure, outcome and covariates of tnterest

Assuming the hybrid approach described above, the propensity score model will be a multilevel, Bayesian Bernoulli model with fixed hospital level effects and random intercepts for clinicians within hospitals. Further addition of covariate-cluster interactions (either interaction terms between hospital-specific intercept and patient-level covariates, or random slopes for clinicians) will be considered based on initial descriptive analyses.

$$\textbf{Outcome: } compliant_{ich} \sim Bernoulli\,(p_{ich})$$
$$\textbf{Patient Level: } \text{logit}(p_{ich}) = \beta_{0ch} + \beta_1 x_{1i} + \cdots + \beta_j x_{ji}$$
$$\textbf{Clinician Level: } \beta_{0ch} = \delta_{00h} + U_{0ch}$$
$$\textbf{Hospital Level: } \delta_{00h} = \gamma_{000} + \theta_{00h}$$

with,

$$U_{0c} \sim Normal\,(0, \sigma_c^2)$$

where, β_{0ch} represents the clinician and hospital specific intercept, composed of a fixed population intercept (γ_{000}), fixed hospital-level intercept (θ_{00h}) and clinician-level random effects (U_{0ch}). $\beta_1 \dots \beta_j$ represent fixed effects coefficients for all included patient-level covariates. Continuous variables (e.g., age, duration of anesthesia, etc.) will be modeled using thin plate splines.

Priors for the propensity score model will be relatively uninformative:

$$\beta_j \sim Normal\,(0, 1.5)$$
$$\theta_{00h} \sim Normal\,(0, 1.5)$$
$$\gamma_{000} \sim Normal\,(0, 1.5)$$
$$\sigma_c \sim Half\text{-}Cauchy\,(0, 1)$$

Following estimation of the propensity score model, posterior draws will be transformed to two sets of weights for each patient: (1) IPTW and (2) OW as previously outlined. Center (mean) and scale (standard deviation) of these distributions of weights will be calculated for use in the outcome model. Balance of covariates of interest will be assessed following weighting (in randomly selected draws from the posterior) using standardized mean differences. Distributions of propensity score weights will also be inspected to ensure adequate overlap, and to identify extreme values.

Outcome Model

The primary outcome model will be a multilevel, propensity score weighted, Bayesian Bernoulli model with random intercepts and slopes for the coefficient of interest ($\beta_{compliant}$) at the hospital- and clinician-level. The full model can be represented as:

$$\textbf{Outcome: } PONV_{ich} \sim Bernoulli\,(p_{ich})^{\tilde{w}_i}$$
$$\textbf{Patient Level: } \text{logit}(p_{ich}) = \beta_{0ch} + \beta_{1ch}x_{1i} + \beta_2 x_{2i} + \cdots + \beta_j x_{ji}$$
$$\textbf{Clinician Level: } \beta_{0ch} = \delta_{00h} + U_{0ch}$$
$$\beta_{1ch} = \delta_{10h} + U_{1ch}$$
$$\textbf{Hospital Level: } \delta_{00h} = \gamma_{000} + V_{00h}$$
$$\delta_{10h} = \gamma_{100} + V_{10h}$$

with,

$$\tilde{w}_i = \kappa_i + \lambda_i \cdot \pi(\lambda)$$

where, β_{0ch} represents the clinician and hospital specific intercept, β_{1ch} represents the clinician and hospital specific treatment effect (with x_{1i} = compliance status), and $\beta_2 \ldots \beta_j$ represent additional fixed effects coefficients. U_{0ch} and V_{00h} represent clinician-, and hospital-level random intercepts, respectively, while U_{1ch} and V_{10h} represent clinician- and hospital-level random slopes for the treatment effect. Finally, γ_{000} is the fixed intercept and γ_{100} is the fixed effect estimate for the treatment and $\tilde{w}_i$ is the propensity score weight term. The distribution and correlation structure between random intercepts and slopes at both the hospital and clinician level are described as:

$$\begin{pmatrix} U_{0c} \\ U_{1c} \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \ \Sigma_c \right]$$

$$\begin{pmatrix} V_{0h} \\ V_{1h} \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \ \Sigma_h \right]$$

with,

$$\Sigma_c = \begin{pmatrix} \tau_{00} & 0 \\ 0 & \tau_{10} \end{pmatrix} R_c \begin{pmatrix} \tau_{00} & 0 \\ 0 & \tau_{10} \end{pmatrix}$$

$$\Sigma_h = \begin{pmatrix} \phi_{00} & 0 \\ 0 & \phi_{10} \end{pmatrix} R_h \begin{pmatrix} \phi_{00} & 0 \\ 0 & \phi_{10} \end{pmatrix}$$

Finally, priors for all parameters are outlined to complete the model:

$$\beta_j \sim Normal\,(0,1)$$
$$\gamma_{000} \sim Normal\,(-1.5, 0.75)$$
$$\gamma_{100} \sim Normal\,(0,1)$$
$$\tau_{00} \sim Half\text{-}Cauchy\,(0,1)$$
$$\tau_{01} \sim Half\text{-}Cauchy\,(0,1)$$
$$\phi_{00} \sim Half\text{-}Cauchy\,(0,1)$$
$$\phi_{01} \sim Half\text{-}Cauchy\,(0,1)$$
$$R_c \sim LKJ\,(2)$$
$$R_h \sim LKJ\,(2)$$
$$\pi(\lambda) \sim Exponential\,(4)$$

Prior Distributions

The full set of prior distributions is delineated in the previous section and selection of these distributions is based primarily on baseline knowledge of the incidence of PONV in the pediatric population (including from prior single center work) and the use of prior predictive simulation. For the fixed intercept we will use *Normal* $(-1.5, 0.75)$, which implies the distribution of PONV probabilities shown in Figure 2 for the average patient receiving non-compliant prophylaxis and with covariates at the mean or in the reference class.
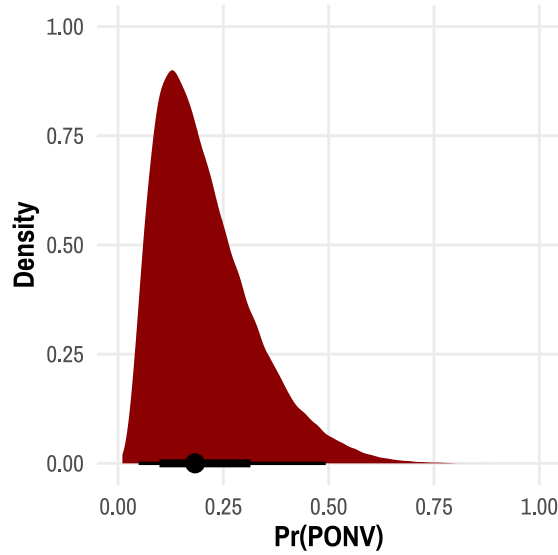

Figure 2: Prior distribution for the fixed intercept on the probability scale.

For the fixed coefficient of interest (γ_{100}) our primary prior will be relatively conservative and centered on no effect: *Normal* $(0,1)$. Simulated distributions using this prior are shown below for (A) γ_{100} on the log odds scale, (B) OR, and (C) the conditional absolute risk reduction (ARR) for patients with all covariates at the reference class. These simulations do not include random effects.
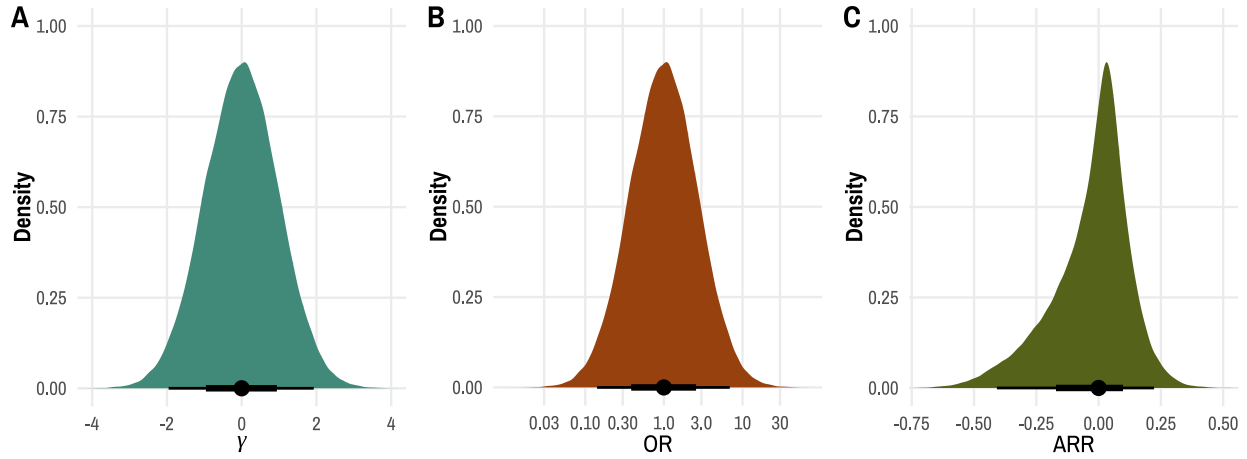
Figure 3: Prior distributions for the effect of interest on various scales.

In addition to this main set of β priors, we will conduct sensitivity analyses using two alternative sets of priors: (1) *brms* default priors (improper flat priors over the reals)[15] and a weakly informative Student's *t* distribution (df = 3, $\mu = 0$, $\sigma = 1$).

Prior distributions for both the standard deviations (*Half-Cauchy* $(0, 1)$) and the correlation matrices (*LKJ* (2)) of random effects will be relatively broad. These distributions are outlined in the plots below.
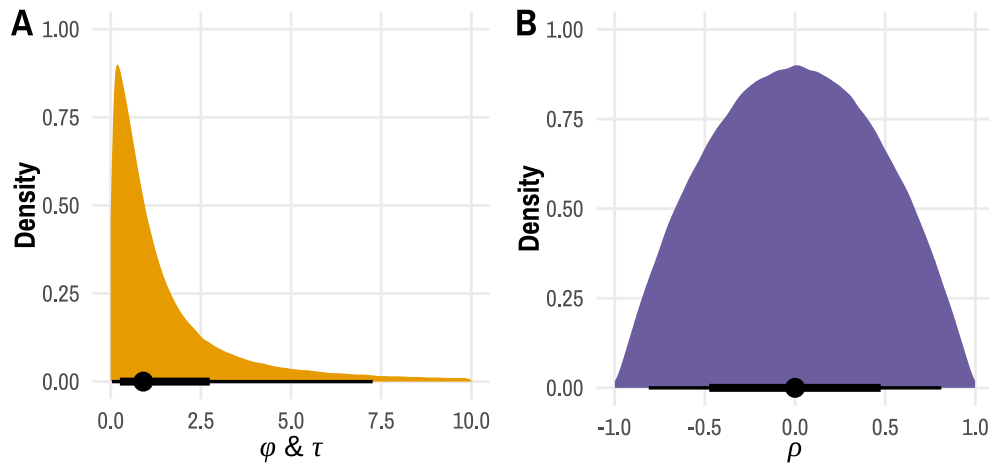

Figure 4: Prior distributions for random effects parameters.

Further Models

In an effort to fully evaluate the robustness of our estimated effects to analytic choices, we will include a standard multilevel, multivariable adjusted Bayesian Bernoulli model (e.g., an unweighted, adjusted model) for comparison to the results obtained through our pseduo-Bayesian propensity score weighted approach.

Additional Details

For the purposes of modeling, all continuous numeric variables will be centered and scaled to a mean of 0 and standard deviation of 1. Models will be fit using the Markov chain Monte Carlo (MCMC) method, specifically the No-U-Turn Sampler (NUTS) implementation of a dynamic Hamiltonian Monte Carlo algo-

rithm via the *brms* package.[15] As described in more detail below, models will be estimated 20 times - once for each multiply imputed dataset, with posterior distributions subsequently pooled for further analysis. Each model will be estimated with four chains in parallel for 8,000 iterations, with the first 3,000 iterations per chain discarded after the warmup stage for a total of 20,000 post-warmup draws. Model convergence will be evaluated using Gelman-Rubin $\hat{R}$ values, MCMC trace plots, effective sample size ratios and Monte-Carlo standard errors. Results will be presented as visual representations of distributions (e.g., draws from the posterior) when feasible, as this allows for the most complete conveyance of information. When summary data are used, a measure of central tendency (most often the median) and uncertainty (50% and 89% quantile credible intervals) will be reported, along with various probabilities of effect (e.g., probability of benefit) based on clinically significant thresholds. Results will be presented on both the relative (e.g., odds ratio) and absolute (e.g., absolute risk reduction) scales, and we will explore conditional measures of effect (e.g, conditional probabilities of benefit) for groups or individual patients. Statements of significance and p-values will not be reported.

Sensitivity Analyses

As previously described, we will perform several sensitivity analyses using alternative prior specifications. Additionally, we will examine the implications of the differential definition of compliant prophylaxis as written in the consensus guidelines and those published by MPOG as well as the outcome of documented PONV alone (without consideration of resuce antiemetic agents). Finally, we will explore the heterogeneity of PONV and the relationship between prophylaxis and PONV across centers to understand more about the surrogate nature of the PONV-03 outcome and any practice/documentation patterns that exist across institutions.

Power Analysis

In the context of the proposed Bayesian analysis we have used a simulation-based approach to assess the impact on sample size on the anticipated precision of our estimates. A simulated multicenter data set of varying sample sizes $[N]$ was constructed based on our preliminary single center analysis. Model structure is similar to that outlined in the above sections. The simulated fixed effect of compliant PONV prophylaxis was defined on the log-odds scale as γ_{100}= -0.105 (equivalent to an odds ratio of 0.9), while the simulated fixed intercept was defined on the log-odds scale as γ_{000}= -2.0 (equivalent to a probability of approximately 0.12). A total of 30 hospitals $[H]$ were simulated, each with a pre-determined proportion of the overall sample N. The total number of clinicians $[C]$ at each hospital was randomly generated by drawing from *Binomial*(100, 0.1) with assignment to each hospital applied after sorting both hospitals and clinician sizes to allow for realistic and proportional "staffing ratios." Probability of compliant prophylaxis (the exposure of interest) was determined for each patient assuming a population-level probability of 0.5 (log odds = 0) with both hospital and clinician level random intercepts drawn from *Normal* $(0, \sigma^2)$. For the purposes of this simulation we did not explicitly include any additional patient-level predictors of treatment allocation (e.g., the design stage model). The probability of receipt of compliant prophylaxis was calculated for each patient, and exposure allocation was determined by drawing from *Bernoulli* $(p_{compliant})$. Clinician- and hospital-level random intercepts (U_{0ch} and V_{00h}) and slopes (U_{1ch} and V_{10h}) were generated by drawing from the multivariate normal distribution. A formulaic representation of the data generating process is outlined below.

$$\textbf{Outcome: } PONV_{ich} \sim Bernoulli\,(p_{ich})$$

$$\textbf{Patient Level: } \mathrm{logit}(p_{ich}) = \beta_{0ch} + \beta_{1ch}x_{1i}$$

$$\textbf{Clinician Level: } \beta_{0ch} = \delta_{00h} + U_{0ch}$$

$$\beta_{1ch} = \delta_{10h} + U_{1ch}$$

$$\textbf{Hospital Level: } \delta_{00h} = \gamma_{000} + V_{00h}$$

$$\delta_{10h} = \gamma_{100} + V_{10h}$$

with,

$$x_{1i} \sim Bernoulli\,(q_i)$$

$$q_i = \theta_0 + \theta_c + \theta_h$$

$$\theta_c \sim Normal\,(0, \sigma_c^2)$$

$$\theta_h \sim Normal\,(0, \sigma_h^2)$$

and,

$$\begin{pmatrix} U_{0c} \\ U_{1c} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix},\ \Sigma_c \right] \qquad \begin{pmatrix} V_{0h} \\ V_{1h} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix},\ \Sigma_h \right]$$

and,

$$\Sigma_c = \begin{pmatrix} \tau_{00} & 0 \\ 0 & \tau_{10} \end{pmatrix} R_c \begin{pmatrix} \tau_{00} & 0 \\ 0 & \tau_{10} \end{pmatrix} \qquad \Sigma_h = \begin{pmatrix} \phi_{00} & 0 \\ 0 & \phi_{10} \end{pmatrix} R_h \begin{pmatrix} \phi_{00} & 0 \\ 0 & \phi_{10} \end{pmatrix}$$

and,

$$\theta_0 = 0$$

$$\sigma_c = 0.1$$

$$\sigma_h = 0.1$$

$$\gamma_{000} = -2$$

$$\gamma_{100} = -0.105$$

$$\tau_{00} = 0.03$$

$$\tau_{01} = 0.03$$

$$\phi_{00} = 0.03$$

$$\phi_{00} = 0.03$$

$$R_c = \begin{pmatrix} 1 & -0.2 \\ -0.2 & 1 \end{pmatrix} \quad R_h = \begin{pmatrix} 1 & -0.2 \\ -0.2 & 1 \end{pmatrix}$$

for $i = 1, \ldots, n_{ch}$, $c = 1, \ldots, C_h$, and $h = 1, \ldots, H$. We then iteratively varied the sample size ($N = 1000, 10{,}000$, $100{,}000$), and for each value of N fit 50 Bayesian outcome models with the structure and priors as outlined in the outcome modeling section. The basic results of the stimulation are shown in the plots below. Figure 5 shows posterior distributions of the odds ratio for the fixed effect estimates (γ_{100}) accompanied by their 95% credible intervals.
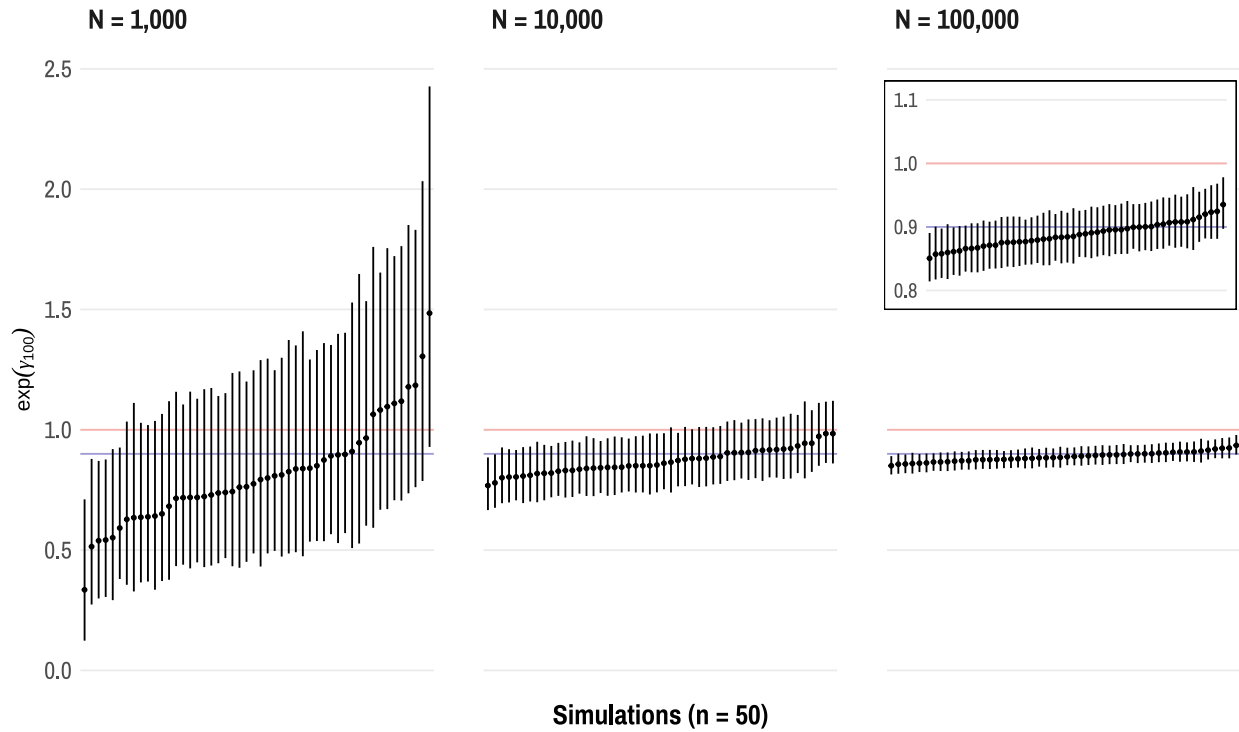
Figure 5: Precision simulation results. Each facet represents a different total sample size (N). For each facet (N), 50 simulations were performed, and each is represented by the posterior median odds ratio (point) and 95% credible intervals (lines), ordered by median OR. The red line marks OR = 1 (no effect) while the blue line marks OR = 0.9 (the true simulated effect). The N = 100,000 facet contains an inset to more easily display the precision of these estimates around the true simulated value.

In these simulations we find that the precision in our effect estimate of interest rapidly increases as sample size moves from 1,000 to 100,000 at which point all 95% credible intervals no longer include OR = 1. As an alternative visual representation of the precision of these estimates we can examine the distribution of their widths (on the OR scale). This is shown in Figure 6 below.
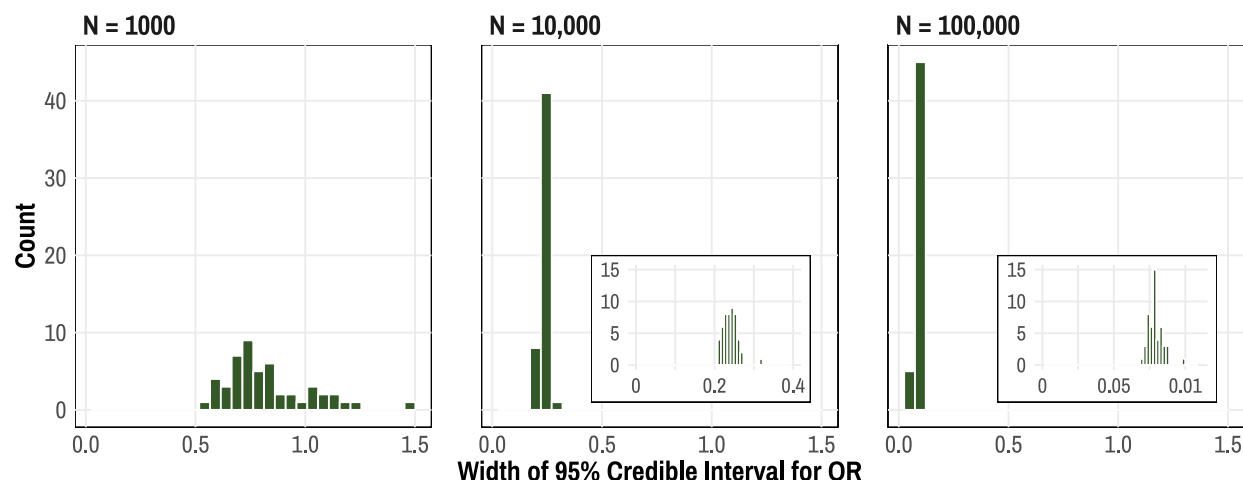
Figure 6: Precision simulation results. Again, each facet represents a different total sample size (N). Here we show the distribution of the widths of the 95% credible interval for the OR, an alternative metric for assessing the precision of our estimates. Facets for N = 10,000 and N = 100,000 contain insets to more clearly show the distribution of widths for these cases.

Examining the inset of the N = 100,000 simulation, we see that all 95% credible intervals for the OR of interest were < 0.1 in width, representing very precise estimates. Given the excellent results obtained from the N = 100,000 simulation, further increases in sample size were not pursued. For reference, each 100,000 sample model in this simulation took approximately ~23 minutes to run.

Handling of Missing Data

While full Bayesian imputation (e.g., imputation of missing values *during* model fitting) is an attractive approach to missing data in this setting, *Stan* (the software behind the *brms* package) does not currently support imputation of discrete parameters. Given that we anticipate missing data to not be confined only to our continuous covariates of interest, we will use an alternative approach to imputation. Missing covariate data will be imputed using multiple imputation with chained equations via predictive mean matching (numeric continuous variables), logistic regression (dichotomous categorical variables), or polytomous regression (multigroup categorical variables) within the *mice* package.[16] The imputation model will be fit 20 times, generating 20 imputed datasets. Both propensity score and outcome models will be estimated once for each imputation dataset, and the posterior distributions will be pooled, similar to the approach used by *Kuck et al.*[17] Missing outcome data will not be imputed; patients with missing PONV-03 values will be excluded.

Software

All analyses, data visualization, and reporting will be performed using R (version 4.3.0; available at https://www.r-project.org/) using the *brms*[15], *tidybayes*[18], *bayesplot*[19], *bayestestR*[20], *mice*[16], *tidyr*[21], *dplyr*[21], and *ggplot2*[21] packages.

Major Threats to Inference & Mitigation Strategies

In this retrospective analysis of PONV prophylaxis, the major threat to inference is the presence of confounding by indication given the non-randomized allocation of prophylaxis. As previously discussed, we will use a pseudo-Bayesian approach to propensity score weighting to attempt to account for *measured* confounding. Importantly, the presence of unmeasured (or unaccounted for) confounders is always plau-

sible in such an analysis. By using a thoughtfully created causal DAG (Figure 1) with expert collaborator input we aim to most appropriately account for confounding while acknowledging that truly unmeasured confounders may still exist.

Preliminary Data Analyses

We recently conducted a single-center retrospective study evaluating the association between compliant (termed "adequate" in this manuscript) prophylaxis (as defined by PONV-04) and PONV (as defined by PONV-03) in pediatric patients undergoing general anesthesia at Duke University Hospital. This study analyzed 14,747 patients across four years (2018-2021) and used propensity-score weighted Bayesian binomial models to estimate the association of interest. In this study we used an approach similar to that outlined by *Liao* and *Zigler*[10] for the pseudo-Bayesian implementation of propensity score weighting. Given the relatively smaller sample size, it was computationally feasible to iterate through the posterior distribution of the propensity score model while fitting the outcome model, thus incorporating the uncertainty in propensity score estimation into the outcome effect estimates. The overall rate of adequate prophylaxis in the cohort was 39% and the overall incidence of PONV was 10.6%. Receipt of adequate prophylaxis varied considerably across risk groups and several variables were poorly balanced across prophylaxis groups. In our primary (weighted and multivariable adjusted) outcome model there was modest evidence for a relative benefit of adequate prophylaxis, with a median odds ratio of 0.82 (95% credible interval 0.66, 1.02; Figure 7).
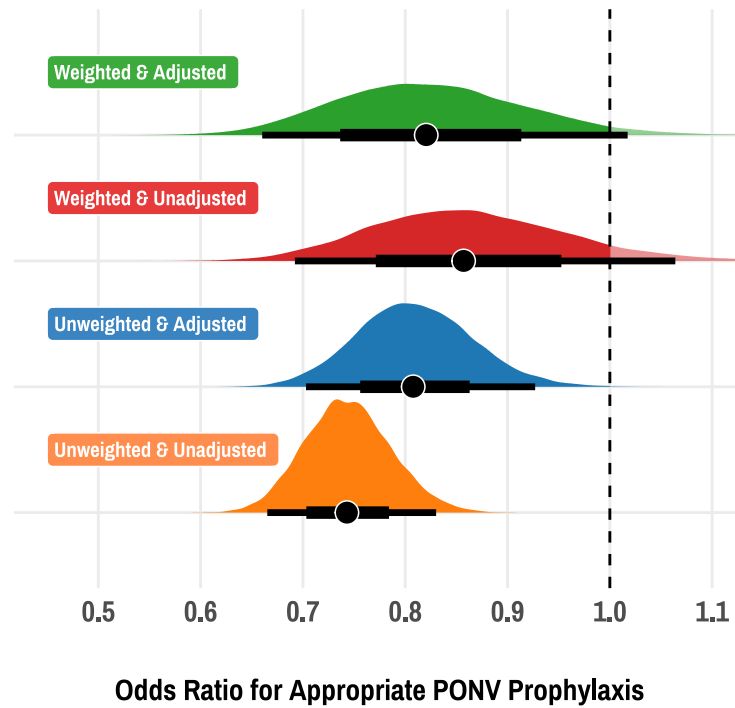
Figure 7: Posterior distributions for the weighted and unweighted estimates for the odds ratio (OR) for the association between adequate prophylaxis and PONV from both unadjusted and adjusted models. Dashed vertical line represents an OR of 1, with the portion of each distribution falling above this point represented with a reduction in opacity. Black points represent the median of the posterior distribution. Inner and outer slab intervals represent 66% and 95% credible intervals, respectively.

On the absolute scale, there was a median marginal weighted absolute risk reduction of 1.3% (95% credible interval -0.1%, 3.1%). Patient-specific, covariate-adjusted ARR was heterogeneous across the study population (Figure 8). Here, each estimate of ARR is conditioned on the individual patient's characteristics and thus is reflective of a patient-specific estimate of benefit. This plot highlights the magnitude and variability of absolute benefit across the range of patients represented in the study population.
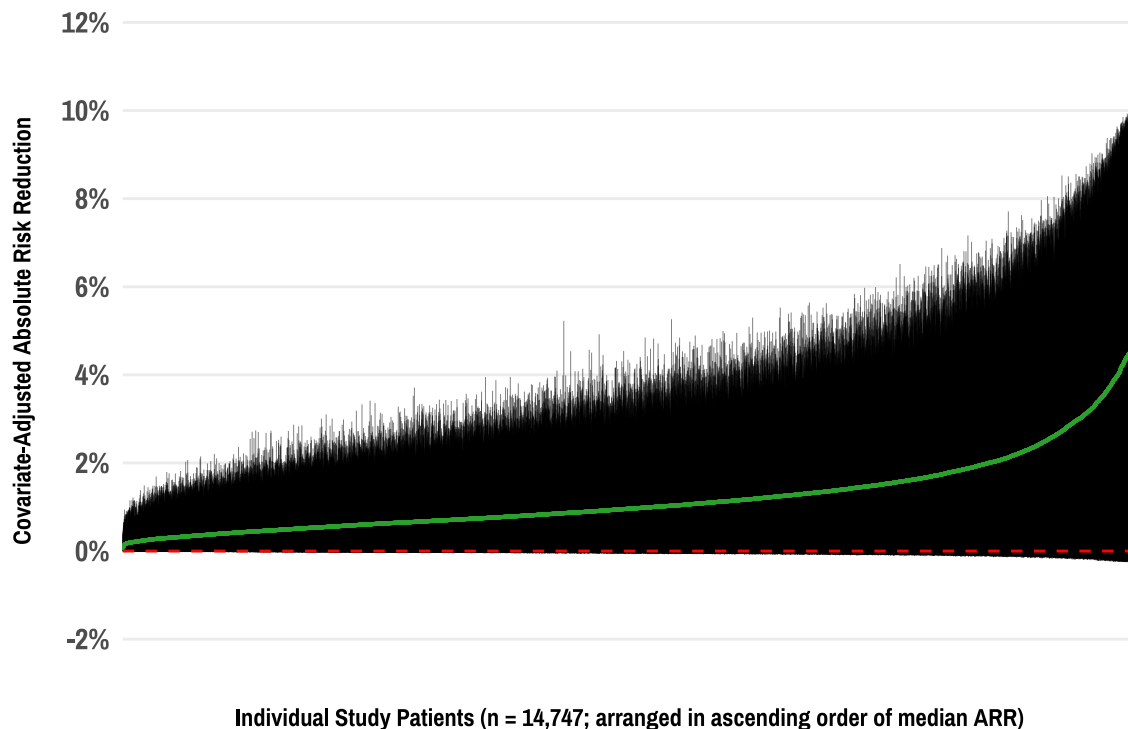
Figure 8: Predicted covariate-adjusted absolute risk reduction (ARR) for each of the 14,747 study patients. Each patient's predicted distribution of ARR is estimated using 2,000 draws from the weighted posterior distribution of the adjusted model. Green points represent the posterior median of the patient-specific distribution and black intervals represent 95% credible intervals for the posterior median. Patient data is ordered on the x-axis based on ascending median ARR. Red dashed line marks ARR = 0; positive ARR values represent lower risk with adequate prophylaxis.

In an interaction model examining the association of adequate prophylaxis across the MPOG-defined risk spectrum, there was evidence for a differential association, with some benefit in lower risk patients but attenuation of this effect in high-risk groups (Figure 9). In *unweighted* estimates there was modest evidence for an interaction between sum of risk factors and the association of adequate prophylaxis with PONV as shown by a reduced incidence in patients with 1-2 risk factors (conditional probability of benefit 0.96 and 0.95 respectively) but increased risk in patients with 3+ risk factors receiving adequate prophylaxis (conditional probability of benefit 0, 0.01, and 0.03 for 3, 4 and 5 risk factors, respectively). This phenomenon was attenuated by weighting, but evidence for differential effects remained, with persistent benefit in 1-2 risk factors (conditional probability of benefit 0.90 and 0.94, respectively) but near equalization of risk in 3+ risk factors (conditional probability of benefit 0.51, 0.57, and 0.57, for 3, 4 and 5 risk factors, respectively).

This phenomenon, and particularly its attenuation with weighting, is possibly consistent with several points: (1) dichotomous risk factor summation as a means to estimate preoperative risk for PONV ignores the possibility of differential effects of individual risk score components, (2) substantial important prognostic information exists beyond these commonly-cited individual risk factors, and (3) discrepancy between clinician-estimated preoperative risk and *a posteriori* determined risk in relation to volatile anesthetic duration. PONV risk at a given sum of risk factors is not homogeneous, but rather is determined by the unique composition of identified risk factors and other important prognostic attributes. It is possible that these differences are identified or inferred by clinicians, prompting use of more antiemetics in patients who, despite having a given risk score, are in fact at higher risk than anticipated.
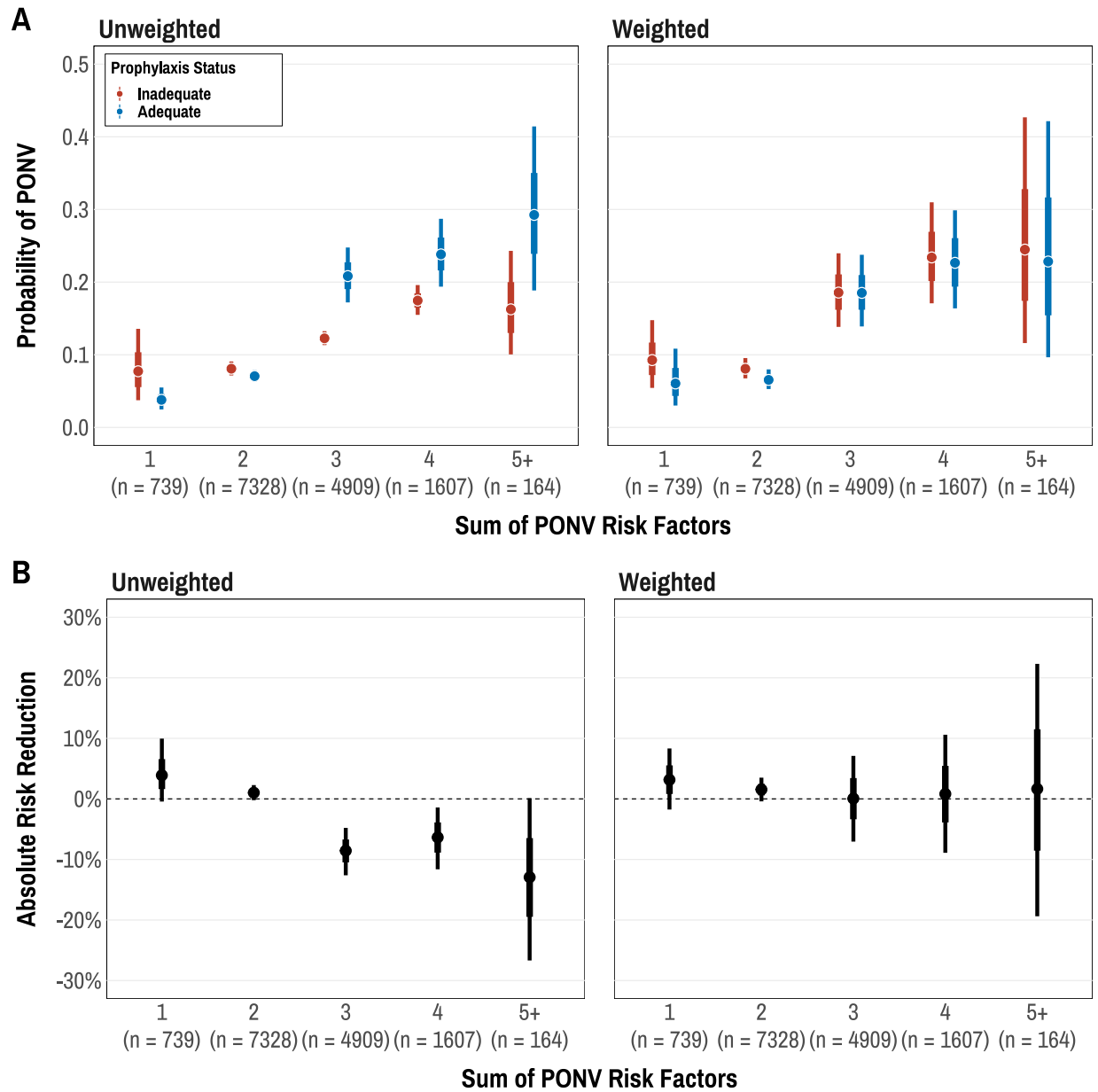
Figure 9: Probability of PONV (A), and absolute risk reduction of adequate prophylaxis (B) stratified by appropriateness of prophylaxis and sum of MPOG PONV risk factors. Unweighted (left plot) and weighted (right plot) estimates are shown, and are generated using 16,000 draws from each model's respective posterior distribution. Points represent the median of the posterior distribution and inner & outer intervals represent 66% and 95% credible intervals for the posterior median, respectively. Note the differential association between appropriate prophylaxis and PONV risk across the spectrum of risk scores in the unweighted estimates, with subsequent attenuation of this phenomenon following propensity score weighting. Black dashed line in (B) marks ARR = 0. See text for definition and values of the conditional probabilities of benefit at each level of risk summation.

Areas for Discussion & Known Limitations

Variability in reporting and objective documentation of nausea and/or vomiting compared to the administration of rescue antiemetic agents is a limitation of the primary outcome for the proposed study. As discussed briefly in earlier sections, while the majority of antiemetic administration in the PACU is likely to be true *rescue* and thus a valid surrogate endpoint for PONV, there is almost certainly some proportion of antiemetic administration that is not explicitly linked to the occurence of PONV. For example, the continuation of *prophylaxis* in the PACU for a patient deemed to be *high risk* be the perioperative team may lead to antiemetic administration in a patient without symptoms and missclassification of the outcome of interest. Similarly, a patient having outpatient surgery who is due to be discharged may receive a *"top-up"* of an antiemetic prior to discontinuation of their peripheral IV in hopes of suppressing post-discharge nausea and vomiting despite not currently having symptoms. Again, this would lead to misclassification of the outcome of interest.

While data contained within the MPOG database is not capable of resolving this potential limitation (something only possible prospective collection of objective measures of nausea and vomiting), we plan to examine both the heterogeneity of PONV incidence across centers (e.g., distribution of random intercepts) and the heterogeneity of the association between compliant prophylaxis and outcome (e.g., distribution of random slopes) to understand the potential magnitude of this biased outcome metric. Further, pending initial exploratory data analysis and documentation consistency, we will also tentatively plan to examine the outcome of PONV as *reported*, irrespective of antiemetic administration, as a more "definitive" measure of true PONV.

References

1. Kovac AL. Postoperative nausea and vomiting in pediatric patients. Pediatric Drugs 2021;23:11–37. Available at: <Go to ISI>://WOS:000584565500001 https://link.springer.com/article/10.1007/s40272-020-00424-0.

2. Morrison C, Wilmshurst S. Postoperative vomiting in children. Bja Education 2019;19:329–33. Available at: <Go to ISI>://WOS:000504314000003 https://www.bjaed.org/article/S2058-5349(19)30108-8/pdf.

3. Blacoe DA, Cunning E, Bell G. Paediatric day-case surgery: An audit of unplanned hospital admission royal hospital for sick children, glasgow. Anaesthesia 2008;63:610–5. Available at: https://www.ncbi.nlm.nih.gov/pubmed/18477272.

4. Gan TJ, Belani KG, Bergese S, Chung F, Diemunsch P, Habib AS, Jin Z, Kovac AL, Meyer TA, Urman RD, Apfel CC, Ayad S, Beagley L, Candiotti K, Englesakis M, Hedrick TL, Kranke P, Lee S, Lipman D, Minkowitz HS, Morton J, Philip BK. Fourth consensus guidelines for the management of postoperative nausea and vomiting. Anesth Analg 2020;131:411–48. Available at: https://www.ncbi.nlm.nih.gov/pubmed/32467512.

5. Multicenter perioperative outcomes group ASPIRE measures. 2022. Available at: https://spec.mpog.org/Measures/Public.

6. Andrew BY, Habib AS, Taicher BM. The association of guideline directed prophylaxis with postoperative nausea and vomiting in pediatric patients: A single-center, retrospective cohort study. Anesth Analg (in press) 2023.

7.   Elm E von, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP, Initiative S. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: Guidelines for reporting observational studies. J Clin Epidemiol 2008;61:344–9. Available at: https://www.ncbi.nlm.nih.gov/pubmed/18313558.

8.   Sung L, Hayden J, Greenberg ML, Koren G, Feldman BM, Tomlinson GA. Seven items were identified for inclusion when reporting a bayesian analysis of a clinical study. J Clin Epidemiol 2005;58:261–8. Available at: https://www.ncbi.nlm.nih.gov/pubmed/15718115.

9.   Kruschke JK. Bayesian analysis reporting guidelines. Nat Hum Behav 2021;5:1282–91. Available at: https://www.ncbi.nlm.nih.gov/pubmed/34400814.

10.   Liao SX, Zigler CM. Uncertainty in the design stage of two-stage bayesian propensity score analysis. Stat Med 2020;39:2265–90. Available at: https://www.ncbi.nlm.nih.gov/pubmed/32449222.

11.   Nafa AJ, Heiss A. Taking uncertainty seriously: Bayesian marginal structural models for causal inference in political science. 2023. Available at: https://github.com/ajnafa/Latent-Bayesian-MSM.

12.   Chang TH, Stuart EA. Propensity score methods for observational studies with clustered data: A review. Stat Med 2022;41:3612–26. Available at: https://www.ncbi.nlm.nih.gov/pubmed/35603766.

13.   Li F, Zaslavsky AM, Landrum MB. Propensity score weighting with multilevel data. Stat Med 2013;32:3373–87. Available at: https://www.ncbi.nlm.nih.gov/pubmed/23526267.

14.   Fuentes A, Ludtke O, Robitzsch A. Causal inference with multilevel data: A comparison of different propensity score weighting approaches. Multivariate Behav Res 2022;57:916–39. Available at: https://www.ncbi.nlm.nih.gov/pubmed/34128730.

15.   Bürkner P-C. Brms: An r package for bayesian multilevel models using stan. Journal of Statistical Software 2017;80.

16.   Zhang Z. Multiple imputation with multivariate imputation by chained equation (MICE) package. Ann Transl Med 2016;4:30. Available at: https://www.ncbi.nlm.nih.gov/pubmed/26889483.

17.   Kuck K, Naik BI, Domino KB, Posner KL, Saager L, Stuart AR, Johnson KB, Alpert SB, Durieux ME, Sinha AK, Brummett CM, Aziz MF, Cummings KC, Gaudet JG, Kurz A, Rijsdijk M, Wanderer JP, Pace NL, Multicenter Perioperative Outcomes Group Enhanced Observation Study Investigator Group for the Multicenter Perioperative Outcomes Group Enhanced Observation Study Collaborator G. Prolonged opioid use and pain outcome and associated factors after surgery under general anesthesia: A prospective cohort association multicenter study. Anesthesiology 2023;138:462–76. Available at: https://www.ncbi.nlm.nih.gov/pubmed/36692360.

18.   Kay M. Tidybayes: Tidy data and geoms for bayesian models. 2023. Available at: http://mjskay.github.io/tidybayes/.

19.   Gabry J, Mahr T. Bayesplot: Plotting for bayesian models. 2022. Available at: https://mc-stan.org/bayesplot/.

20.   Makowski D, Ben-Shachar M, Lüdecke D. bayestestR: Describing effects and their uncertainty, existence and significance within the bayesian framework. Journal of Open Source Software 2019;4.

21.   Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen T, Miller E, Bache S, Müller K, Ooms J, Robinson D, Seidel D, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H. Welcome to the tidyverse. Journal of Open Source Software 2019;4.