Associations between enteropathogens detected in the environment and child growth and enteric infections: an individual participant data meta-analysis

Andrew Mertens, Jack Colford, Oliver Cumming, Joe Brown, Jill Stewart, David Holcomb, Drew Capone, Jackie Knee, Tom Clasen, Heather Reese, Amy Pickering, Clair Null, Steve Luby, Jessica Grembi, Ben Arnold, Audrie Lin, Jade Benjamin-Chung, Laura Kwong, Lauren Steinbaum, Ali Boehm, Kara Nelson, Erica Fuhrmeister, Mahbubur Rahman, Sammy Njenga, Rassul Nala, Ruwan Thilakaratne, Ayse Ercumen (middle order not finalized)

## Summary

## Methods

We examined associations between prevalence of pathogens and MST markers in the environment and child health outcomes.

For binary outcomes, estimated prevalence ratios using modified Poisson regressions.1 For abundance outcomes, we used linear regressions to estimate differences in log-transformed gene copies and negative binomial regressions to estimate count differences in soil-transmitted helminth (STH) egg counts. Because of repeated sampling or clustered designs in some studies, we used the Huber Sandwich Estimator to calculate robust standard errors.2 For abundance measures, we imputed values for samples below the limit of quantification (LOQ) and the limit of detection (LOD). For samples below the LOD, we used half the LOD and for samples below the LOQ, we used the midpoint between the LOD and LOQ for the imputation. To avoid analyzing targets where most of the data were imputed, we only included targets where at least 50% of samples were within the quantifiable range in our analysis.

All analyses were adjusted for potential confounders. Covariates were prescreened using likelihood ratio tests, and only variables associated with the outcome with a p-value < 0.2 were included in the model for each outcome. We included the following variables in the prescreening set if they were measured within an included study: child age, child sex, maternal age, household food security status, number of people in the household, age and education of primary caregiver in the household, asset-based household wealth, number of rooms, construction materials (walls, floor, roof), access to electricity, land ownership and if anyone in the household works in agriculture. For rare binary outcomes, we only included one independent variable (including the treatment arm) per 10 positive samples, or per 10 negative samples if <50% of samples were negative. We also did not estimate prevalence ratios for any targets with fewer than two positive or negative values per study arm for a given sample type. Targets with <10 negatives were also excluded from the composite “any pathogen” and “any MST marker” variables to avoid generating sparse cells; targets with <10 positives were included in the composite variables.

Given the heterogeneity in study settings (e.g., local WASH conditions, climate, urbanization, population density, region-specific infectious disease patterns, intervention designs), we reported individual study-specific estimates for all analyses. For targets where data were available from four or more studies, we tested for heterogeneity in estimates using Cochran’s Q-test.3 If there was no significant heterogeneity (p-value>0.2), we pooled estimates using fixed-effects models. If there was evidence for heterogeneity but there was qualitative support for combining studies, we pooled estimates using random-effects models.

Aim 2: Primary outcomes for this aim include: Prevalence of diarrhea and length-for-age z-scores (LAZ). Secondary outcomes include z-scores for weight-for-age (WAZ), weight-for-length (WLZ), head circumference and middle-upper-arm-circumference, prevalence of stunting, wasting and underweight, prevalence and intensity of infection with specific enteropathogens, and prevalence of respiratory infections. We will address Aim 2 by estimating PRs and PDs for the binary health outcomes and mean differences for the continuous health outcomes for individuals with vs. without exposure to pathogens and MST markers in environmental samples. The primary outcomes of Aim 1 (prevalence of any enteropathogen, and any general, human or animal MST markers in environmental samples) will be used as the exposure variables for this aim. We will estimate the associations between these exposure variables and the primary and secondary outcomes of Aim 2. We will also estimate associations between the prevalence/abundance of a specific pathogen in environmental samples and the prevalence/abundance of the same pathogen measured in child stool. For the LAZ, stunting and head circumference outcomes, we will consider all environmental samples collected over the child’s lifetime prior to the anthropometry measurement. For the other outcomes, we will only consider environmental samples collected up to three months before the measurement of the health outcome. The Aim 2 analyses will be conducted by sample type (e.g., water, hands, soil) and will pool data from all study arms.

All analyses with be conducted using generalized linear models, adjusting for potential confounders. We will conduct unadjusted analyzes as a sensitivity analysis. The analysis for Aim 2 will adjust for study arm. In addition, all adjusted analyses will control for child age, sex, and pre-screened covariates found to be significantly related to the outcome in bivariate likelihood ratio tests (full list of covariates above). We will use the likelihood ratio test to assess the association between each outcome and each covariate and will include covariates with a p-value<0.2 in the analysis.

-note we drop missing continious but add a missigness category for categorical covariates

All analyses were adjusted for potential confounders. Covariates were prescreened using likelihood ratio tests, and only variables associated with the outcome with a p-value < 0.2 were included in the model for each outcome. We included the following variables in the prescreening set if they were measured within an included study: child age, child sex, maternal age, household food security status, number of people in the household, age and education of primary caregiver in the household, asset-based household wealth, number of rooms, construction materials (walls, floor, roof), access to electricity, land ownership and if anyone in the household works in agriculture. For rare binary outcomes, we only included one independent variable (including the treatment arm) per 10 positive samples, or per 10 negative samples if <50% of samples were negative. We therefore did not estimate prevalence ratios for any targets with fewer than 10 positive or negative values for a given sample type, or with fewer than two positive or negative values per study arm for a given sample type. Targets with <10 negatives were also excluded from the composite “any pathogen” and “any MST marker” variables to avoid generating sparse cells; targets with <10 positives were included in the composite variables.

### Overall summary of results:

Most study-specific estimates are null, with inconsistent direction of effects in significant associations. Estimates pooled over multiple studies were also null, except for a small and marginally significant association between any pathogen in any sample and lower child height-for-age Z-scores (which is significant without adjustment for confounders).

### Overall notes on data availability:

* Odagiri et al. 2016 only measured weight, so we only have WAZ, and Reese only measures/shared height, so we only have HAZ.
* The tables at the bottom of the report show the number of samples and number of health outcomes by study the column for both positive sample and diarrhea measure is likely the limiting factor for sparse analyses.

### Notes on analysis

* The analysis included baseline (pre-intervention) measurements.
* All primary estimates are adjusted for intervention arm and child and household covariates.
* Only child health measurements taken after environmental samples were used.
* Diarrhea measurements must have occured after environmental samples, but within 4 months of environmental sample collection.
* Environmental samples were matched to the most proximate child health outcome, without using multiple measurements. For example, environmental samples at baseline were matched to child anthropometry and midline, but not endline.

### Notes on time ordering of environmental samples and child health outcomes, and data merging by study

#### WASH Benefits Bangladesh

* Endline (year 2) anthropometry and diarrhea was used for Kwong et al. 2021 (STH samples)
* World Bank substudy diarrhea and anthropometry was used for Boehm et al. 2016
* R01 substudy diarrhea and anthropometry was used for Fuhrmeister et al. 2020. The substudy was conducted over 8 rounds taken around 3 months apart, with environmental sampling occurring in rounds 3 and 4. Environmental samples were merged to diarrhea from the subsequent round and anthropometry from the main trial endline (year two) sampling.

#### WASH Benefits Kenya

* Endline (year 2) anthropometry and diarrhea was used.

#### Mapsan

* The Mapsan trial had three sampling rounds, baseline, midline, and endline, each 12 months
* The Mapsan trial environmental sampling data is divided into three studies which had differences in samples, microbial targets, and sampling times, Holcomb et al 2020 (baseline and midline), Capone et al 2021 (baseline and endline), and Capone et al 2021 in prep. (baseline and midline).
* Diarrhea was used from concurrent rounds, while anthropometry was used from subsequent rounds, except for endline environmental samples, where concurrent anthropometry was used.

#### Odisha

* Environmental samples were shared already merged with child health data, but samples outside of the specified time range for diarrhea or taken before environmental

#### Gram Vikas

* Sampling rounds were approximately 4 months apart, so anthropometry data was taken from subsequent round, and diarrhea data was taken from either the current or subsequent round, based on which sample was taken after but closer to the environmental sampling, and within 4 months.

## Results

Adjusted diarrheal disease prevalence ratio for any pathogen presence in any environmental sample: 1.11 (95% CI: 0.94, 1.31)

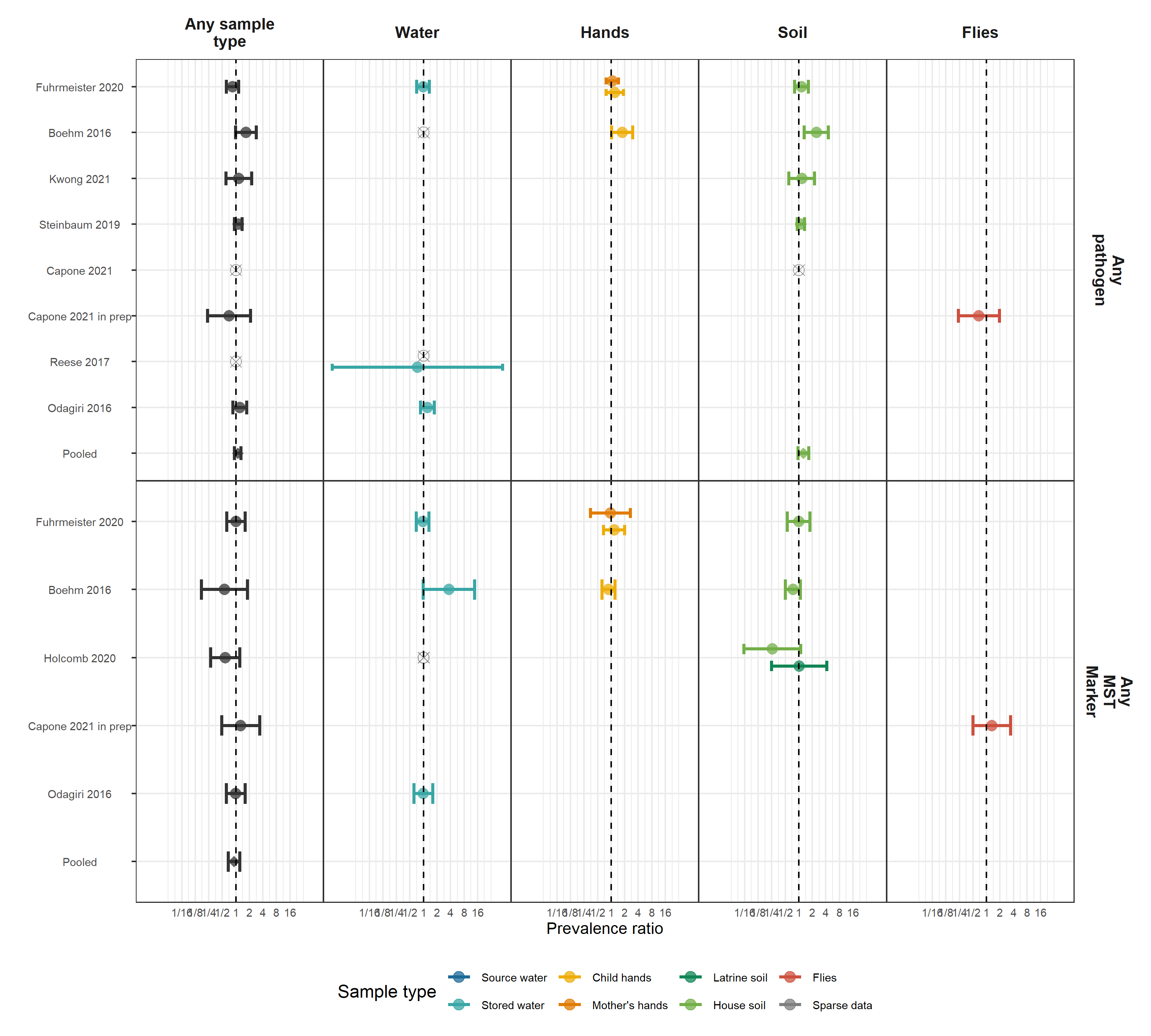
Adjusted HAZ difference for any pathogen presence in any environmental sample:

-0.08 (95% CI: -0.16, -0.01)

# Primary figures

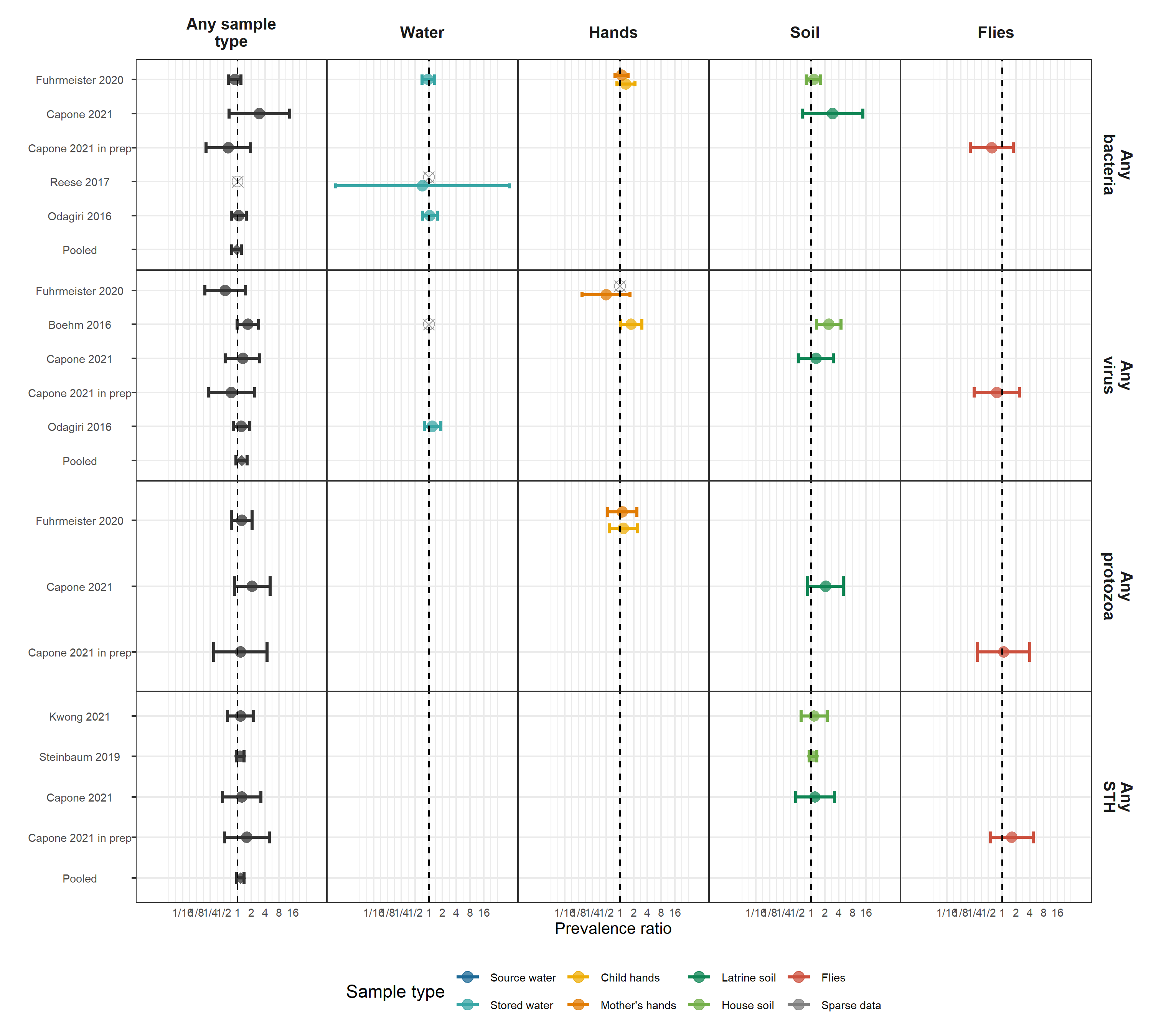
## Diarrhea

#### Adjusted associations between diarrhea and any pathogen or MST marker

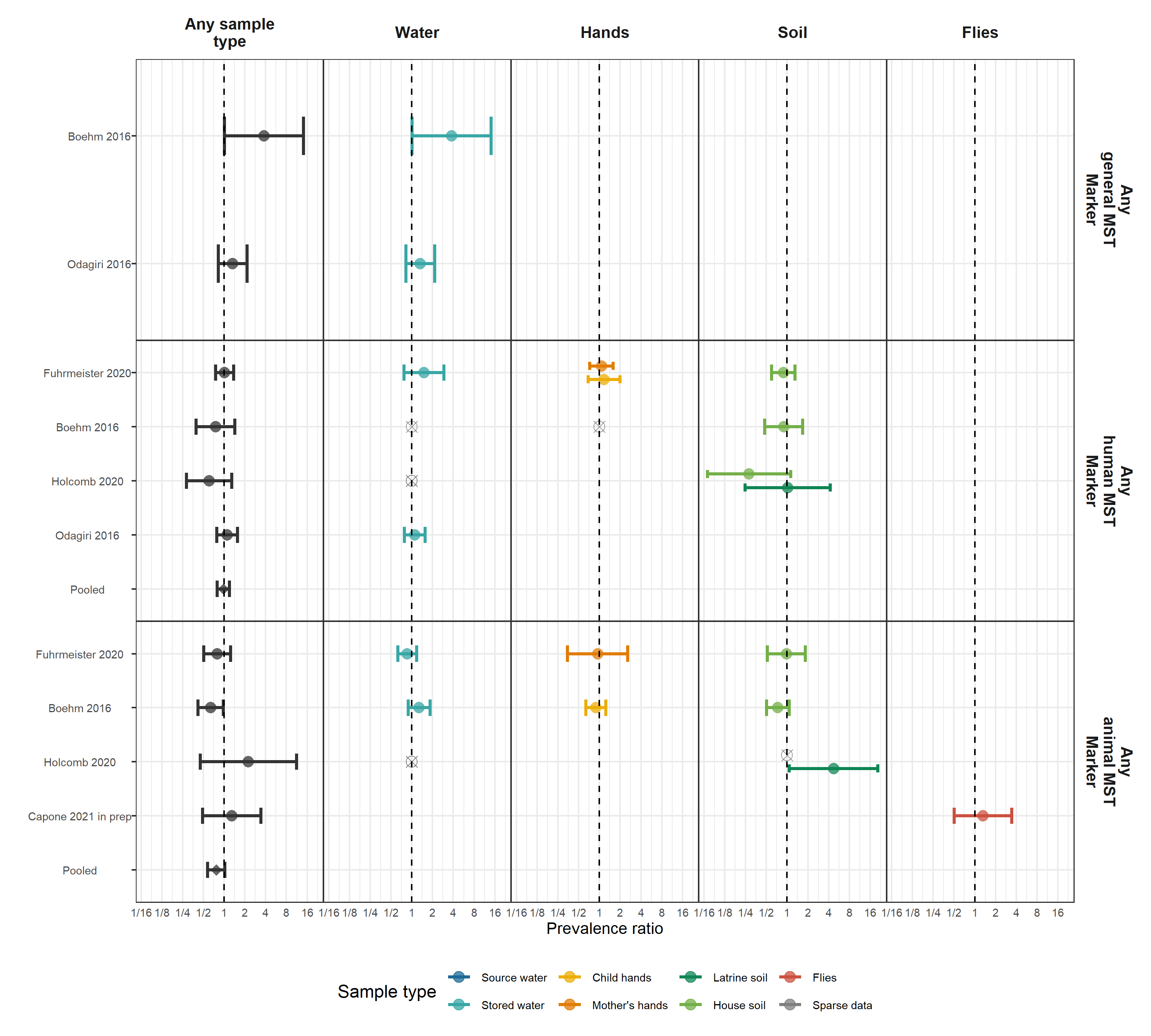


**Figure 1.** Forest plots of associations between child diarrheal disease and the prevalence of any enteropathogen or any MST markers in different types of environmental samples. Pooled estimates are presented when there are four or more study-specific estimates for a specific sample type and target combination and are denoted with diamond-shaped points. Grey crossed points denote data that were too sparse to estimate a prevalence ratio (i.e., <10 positive or negative observations). Samples of the same type from different locations (source vs. stored water, flies in kitchen vs. latrine, soil from courtyard vs. latrine) or different individuals (child vs. mother’s hands) are plotted separately. All estimates are adjusted for potential confounders.

#### Adjusted associations between diarrhea and types of pathogens

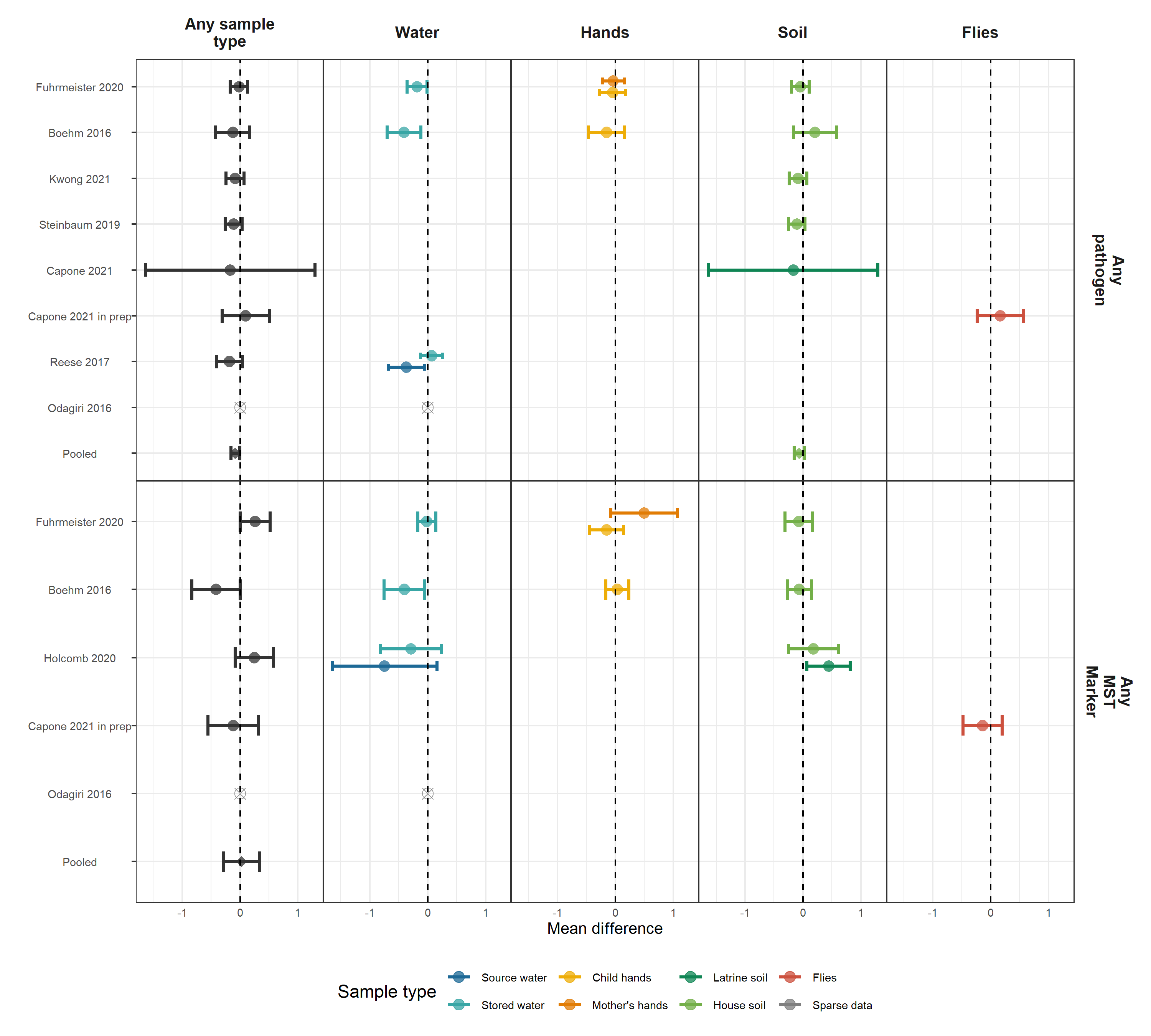


#### Adjusted associations between diarrhea and types of MST markers

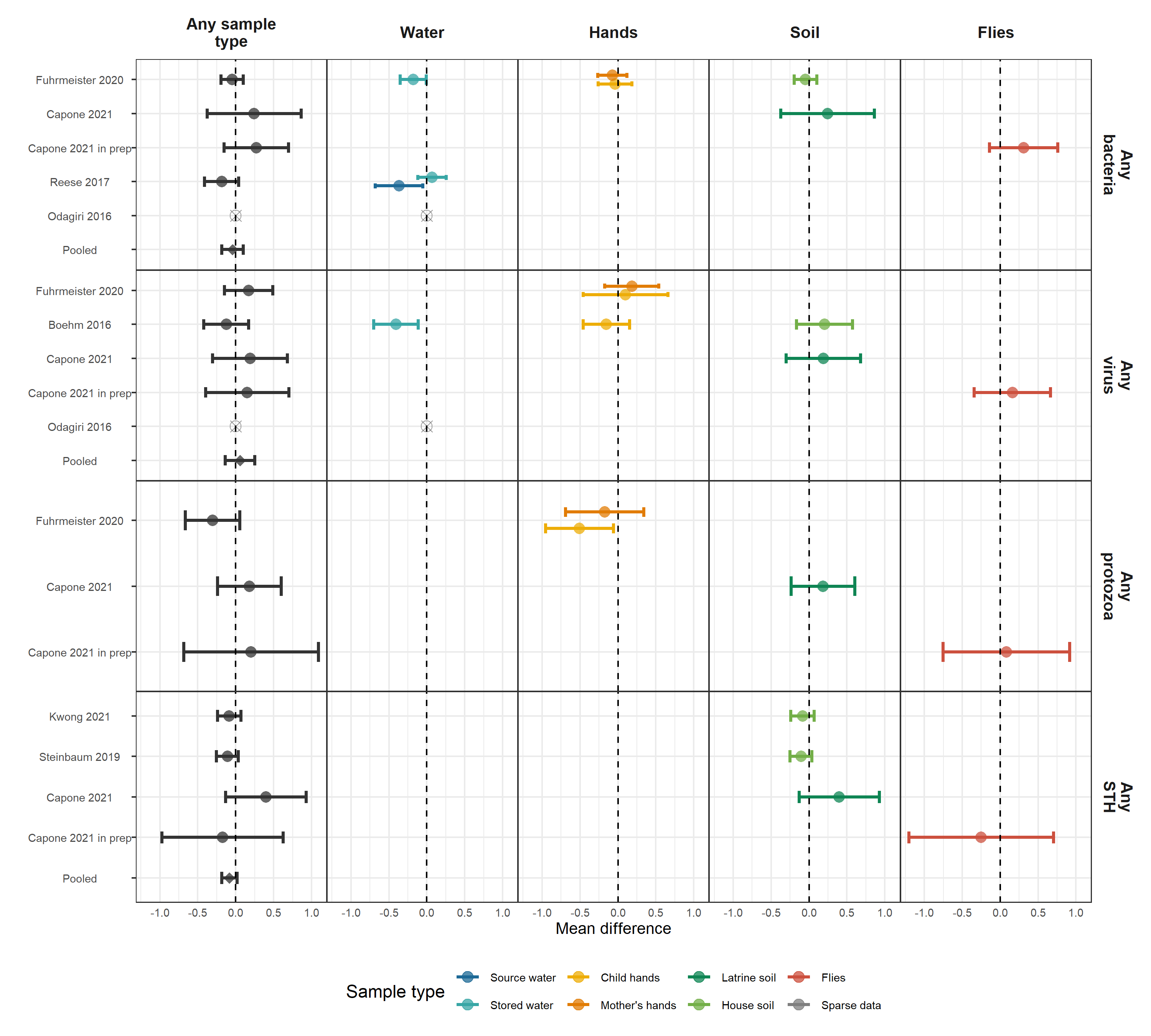


## Height-for-age Z-scores

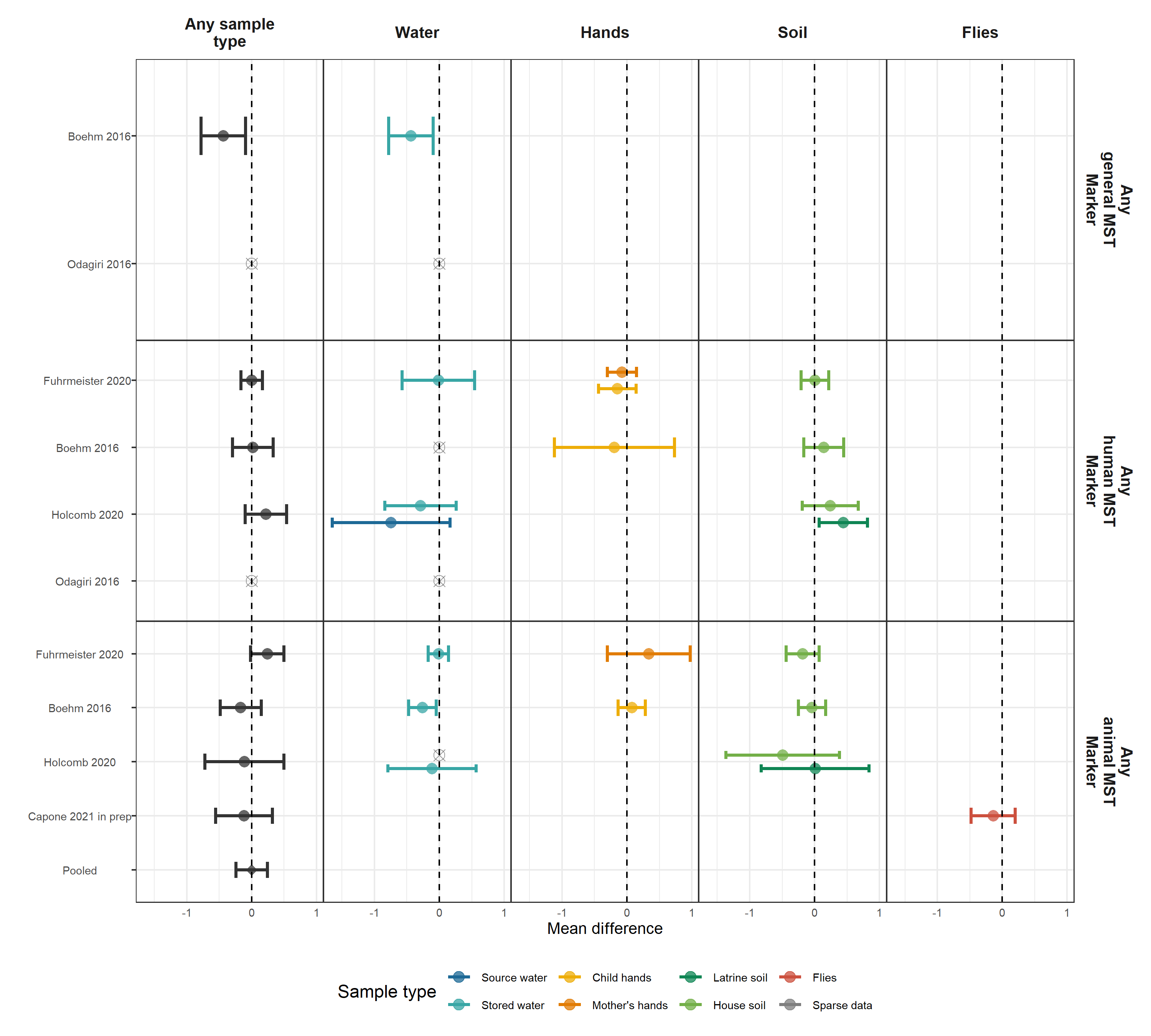
#### Adjusted associations between HAZ and any pathogen or MST marker



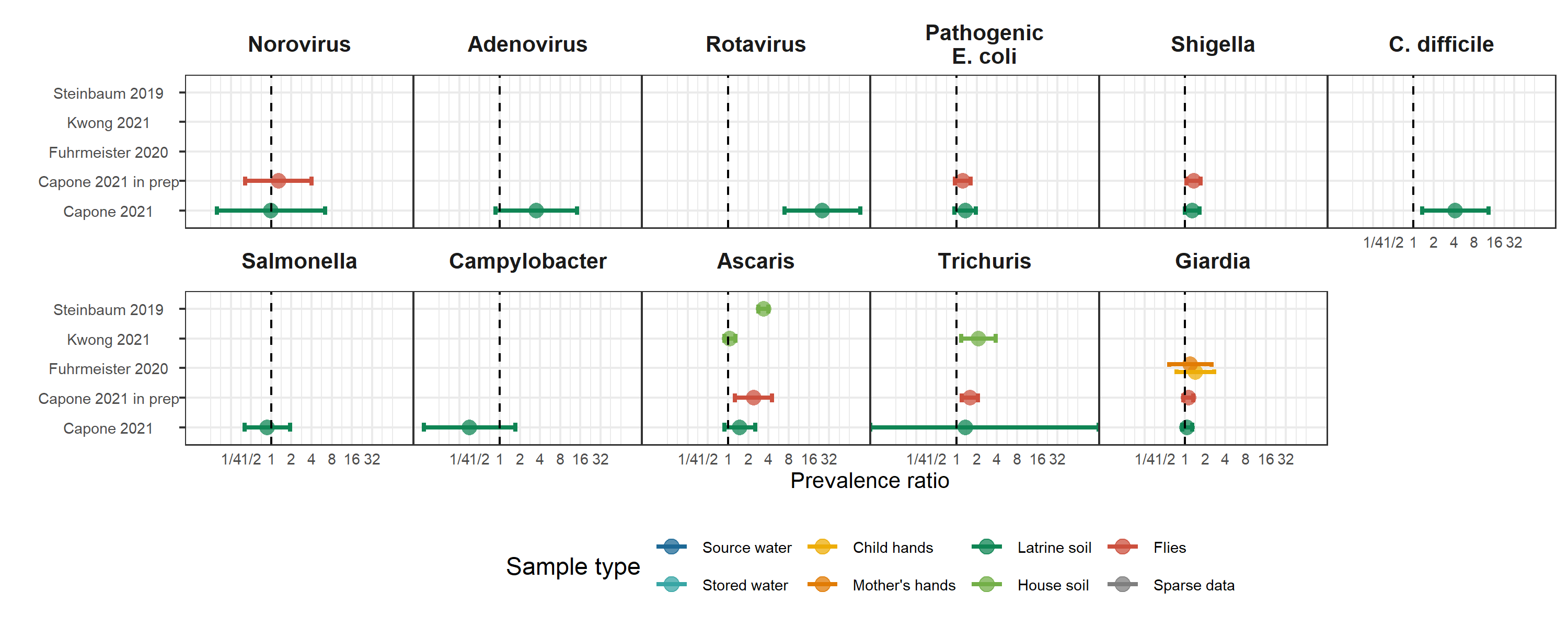
#### Adjusted associations between HAZ and types of pathogens



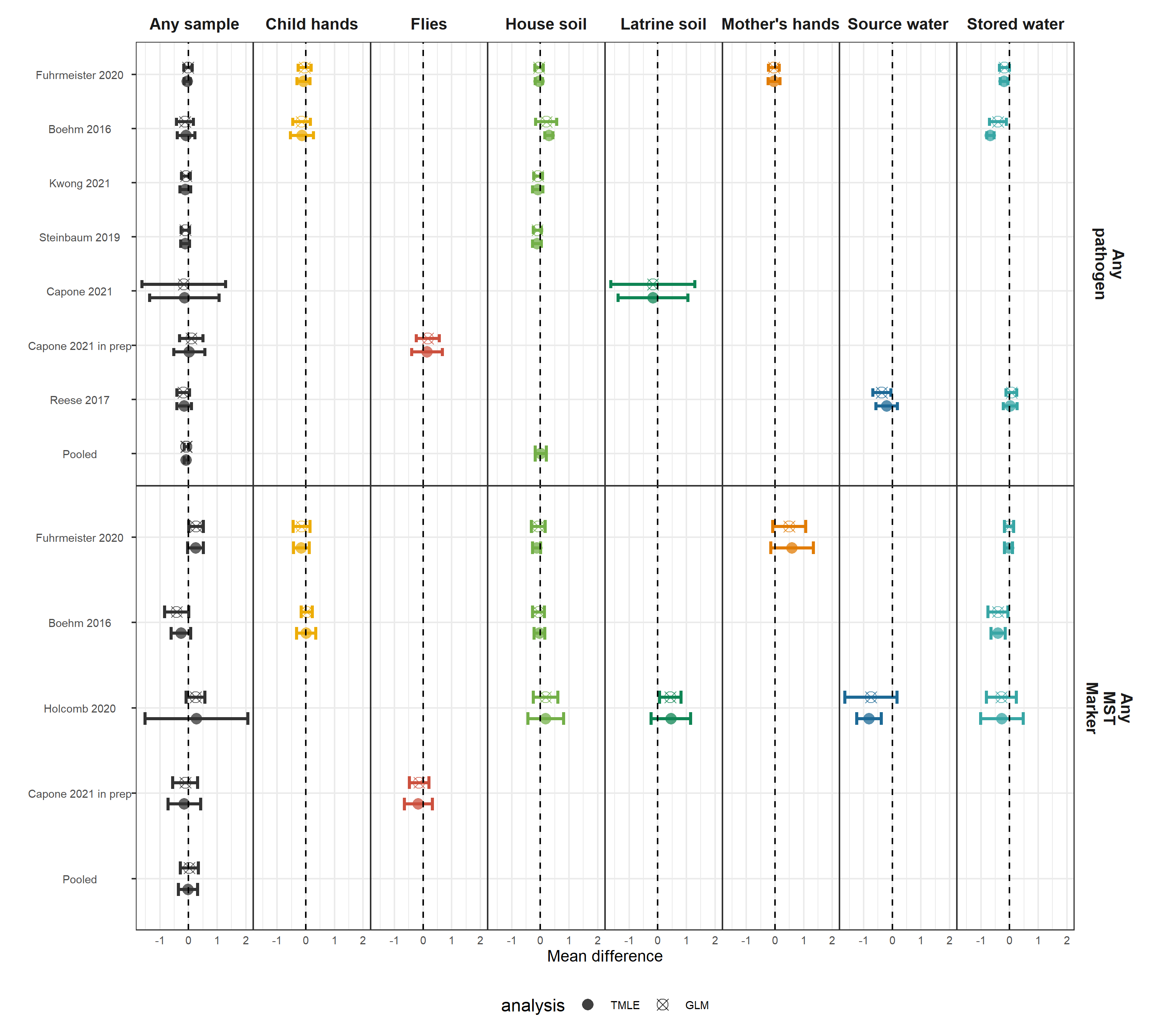
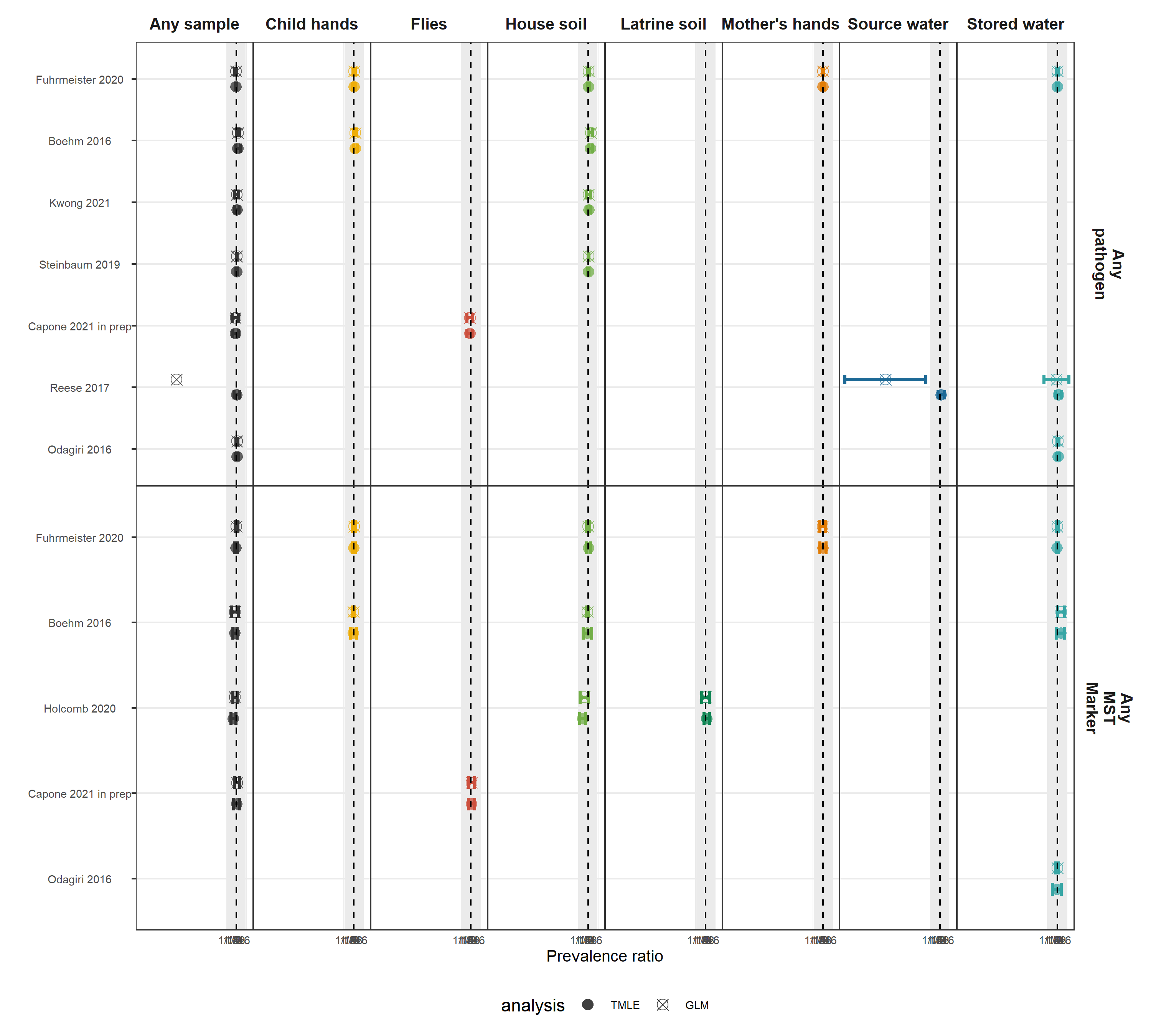
#### Adjusted associations between HAZ and types of MST markers



#### Adjusted associations between pathogen-specific presence in environmental samples and pathogen-specific infections

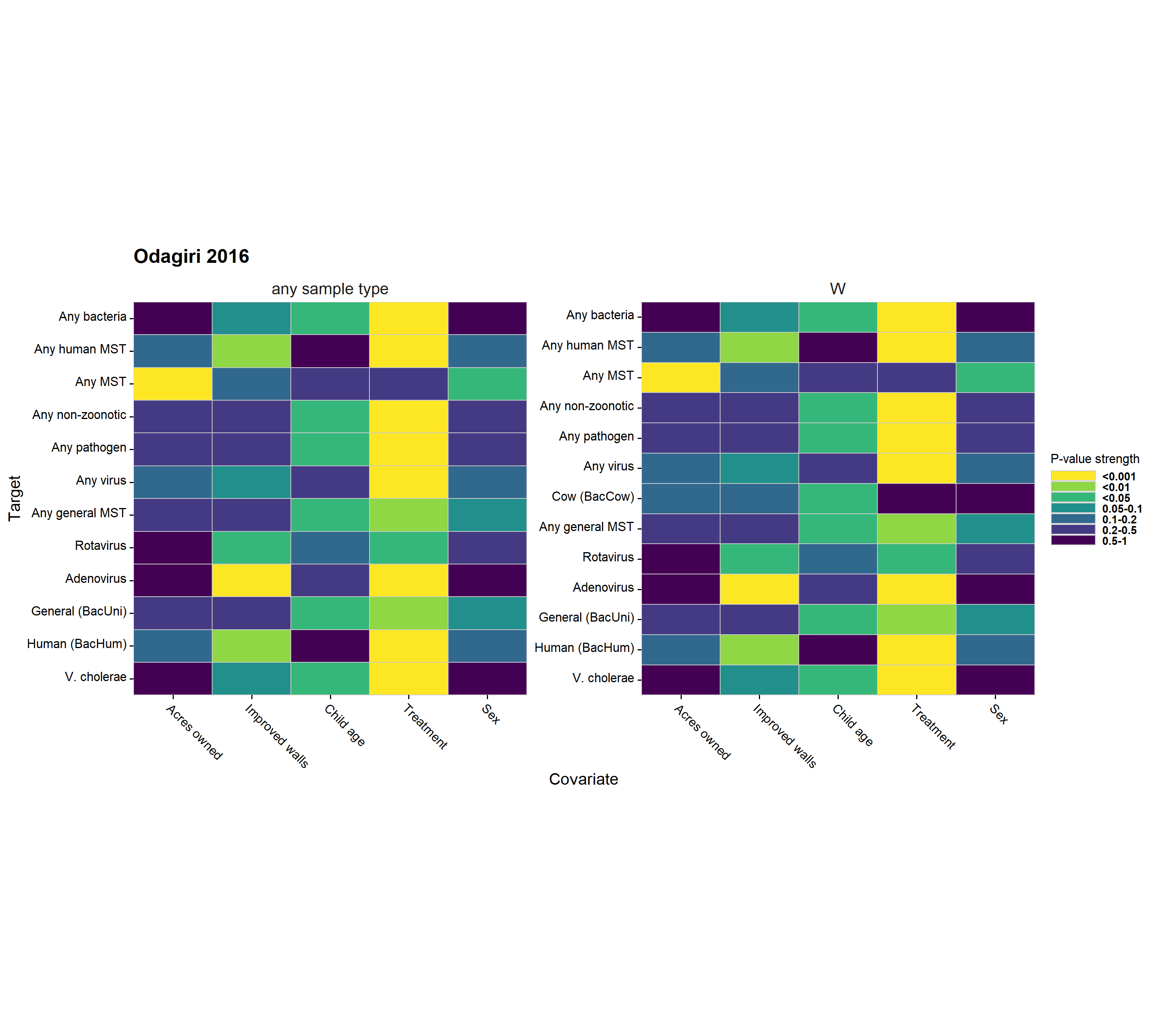
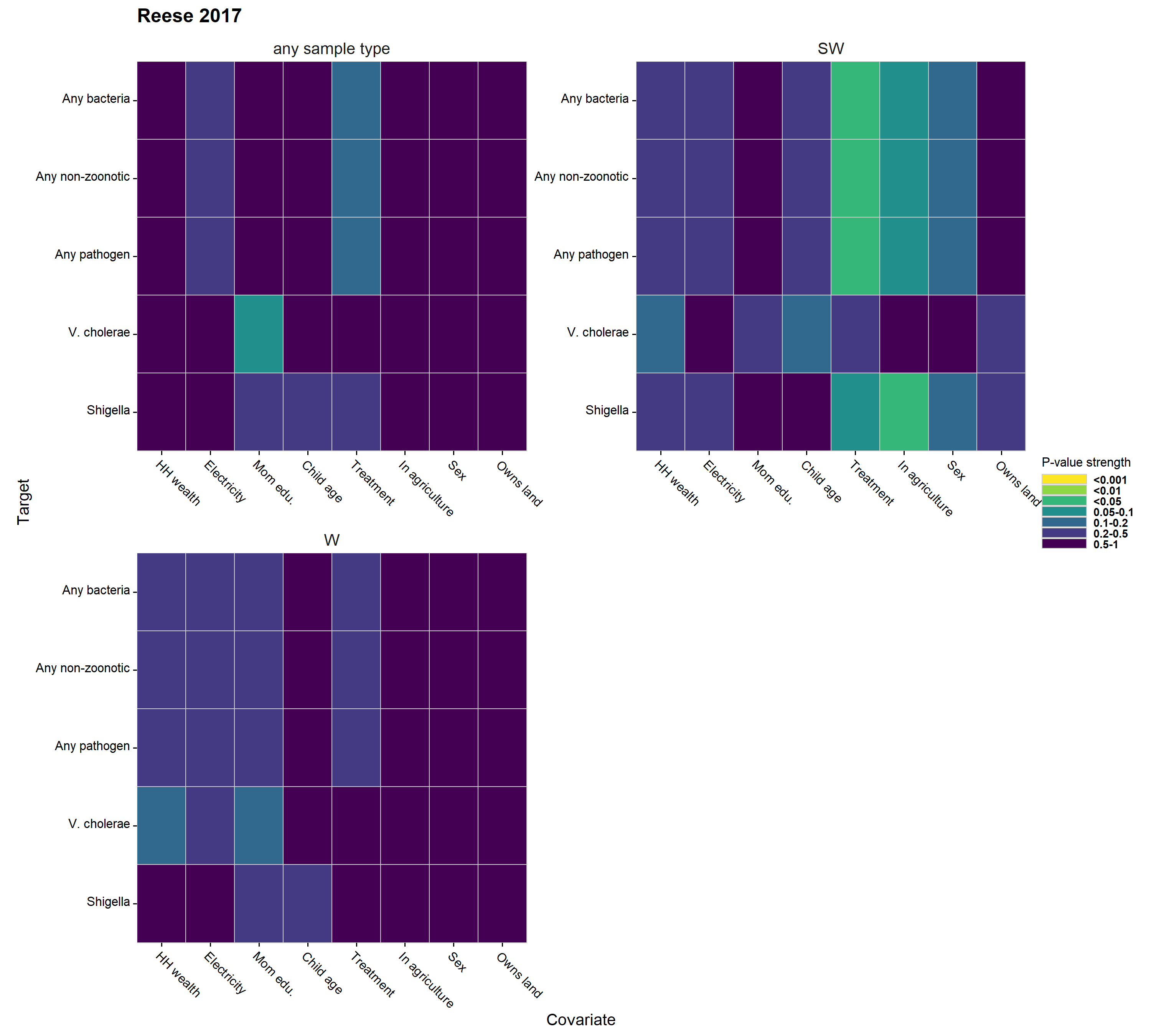
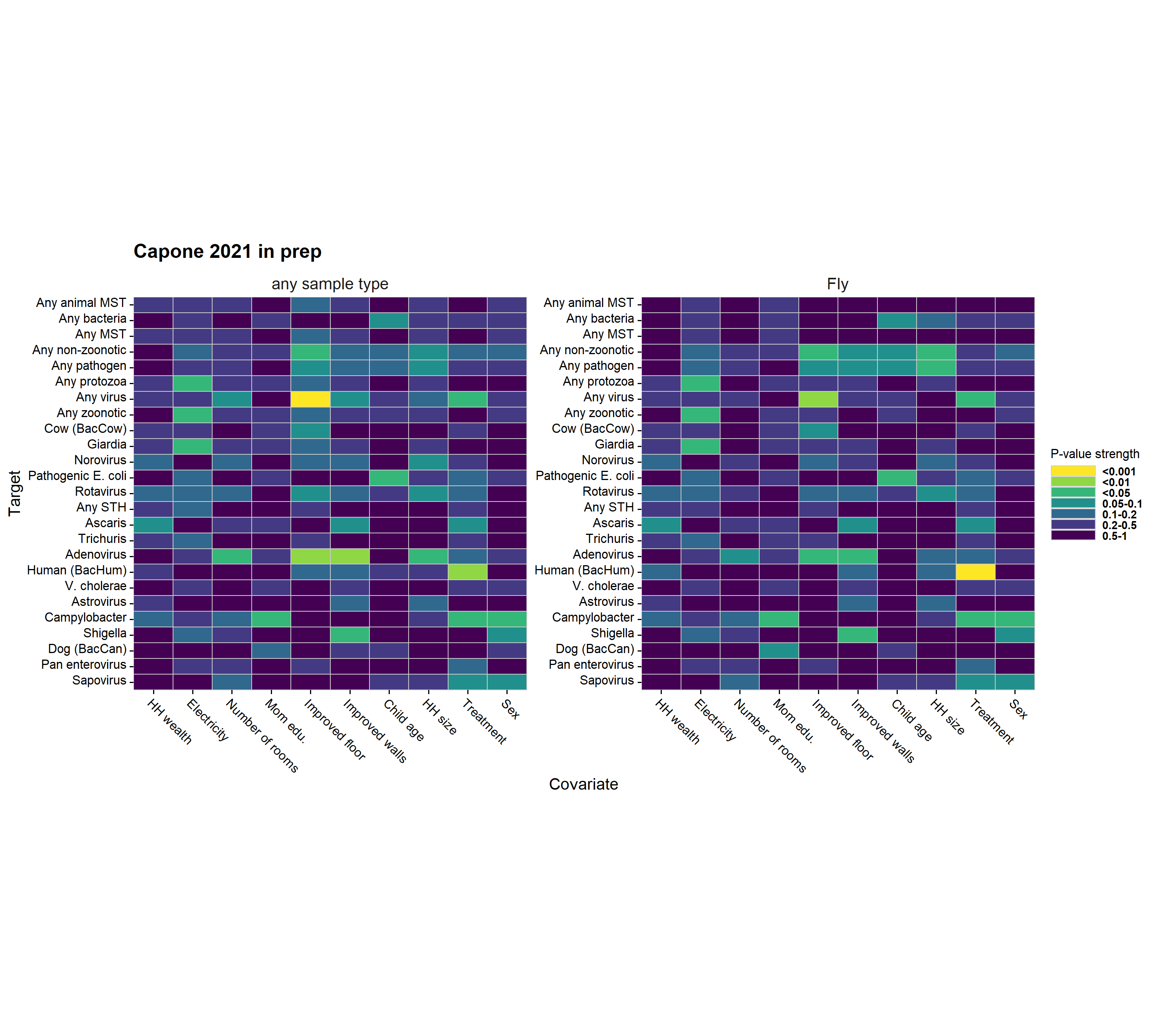
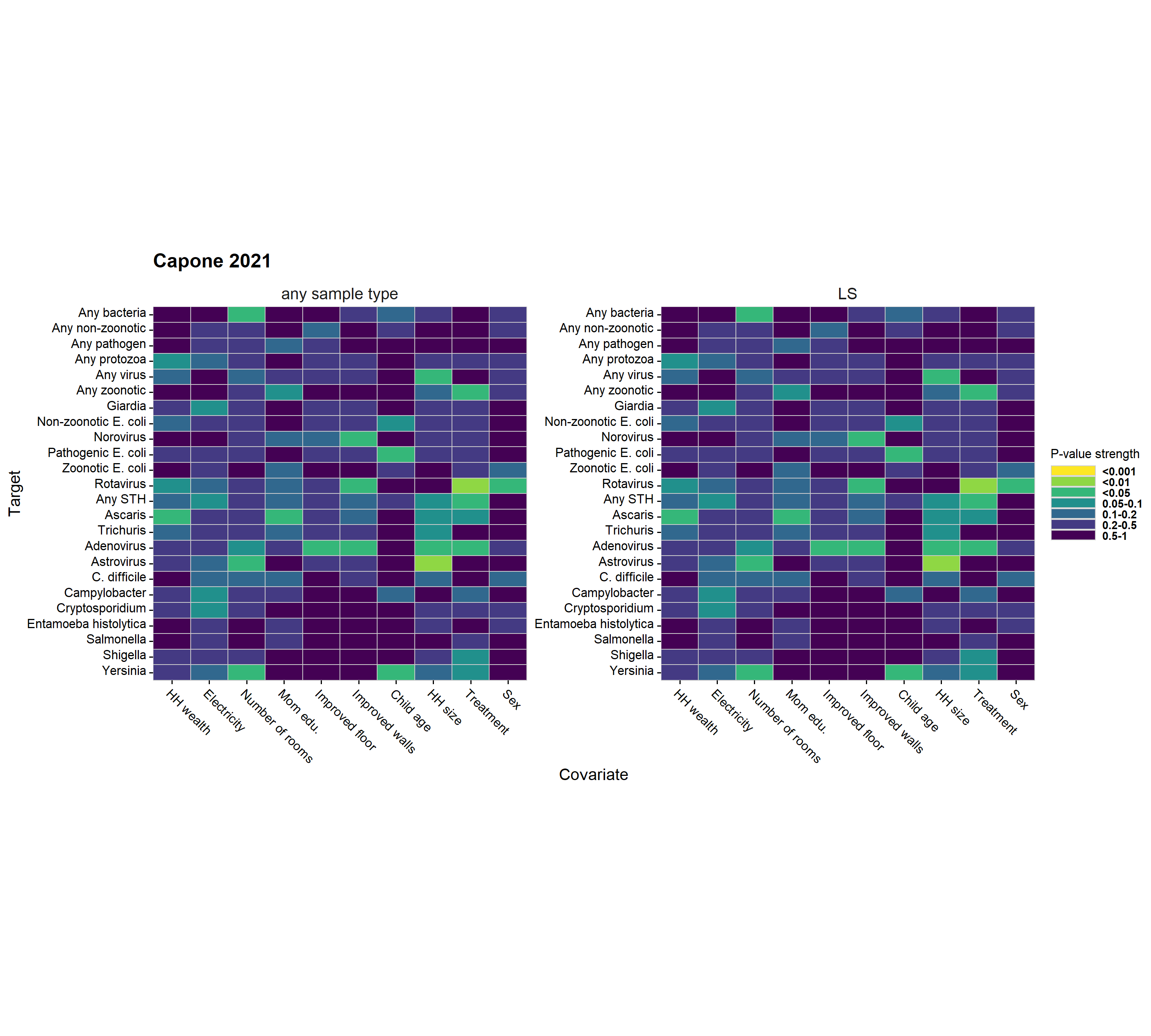
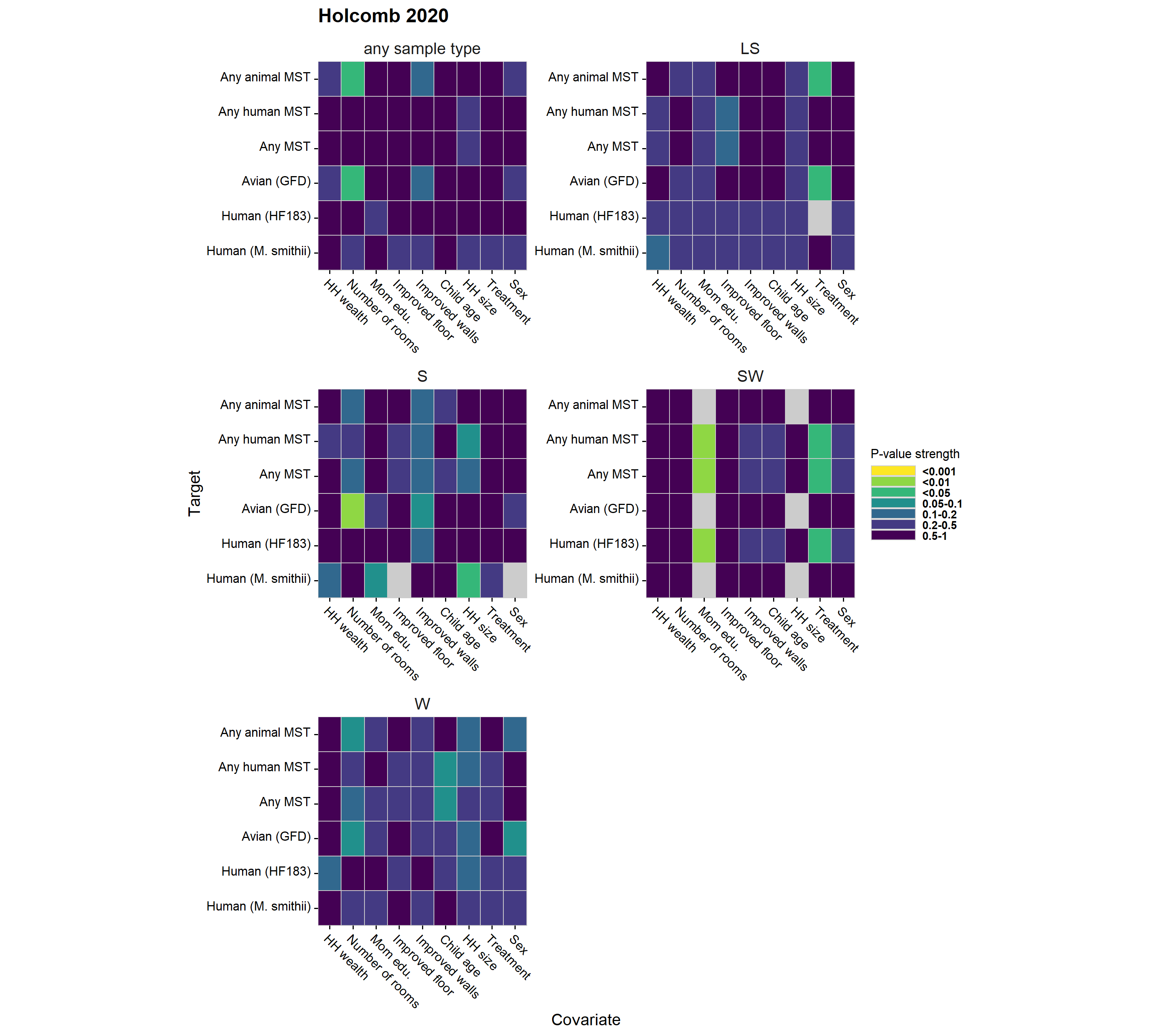
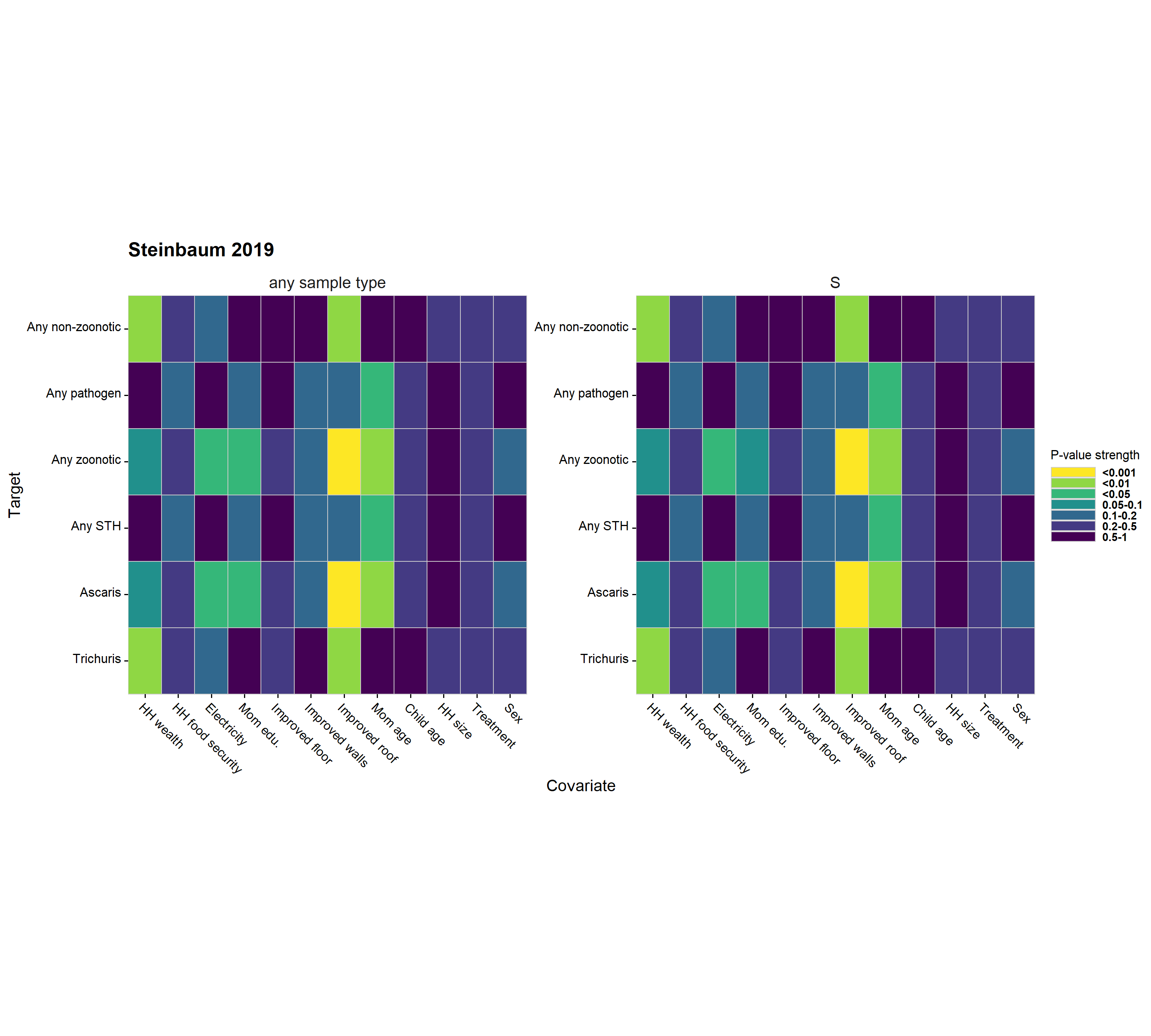
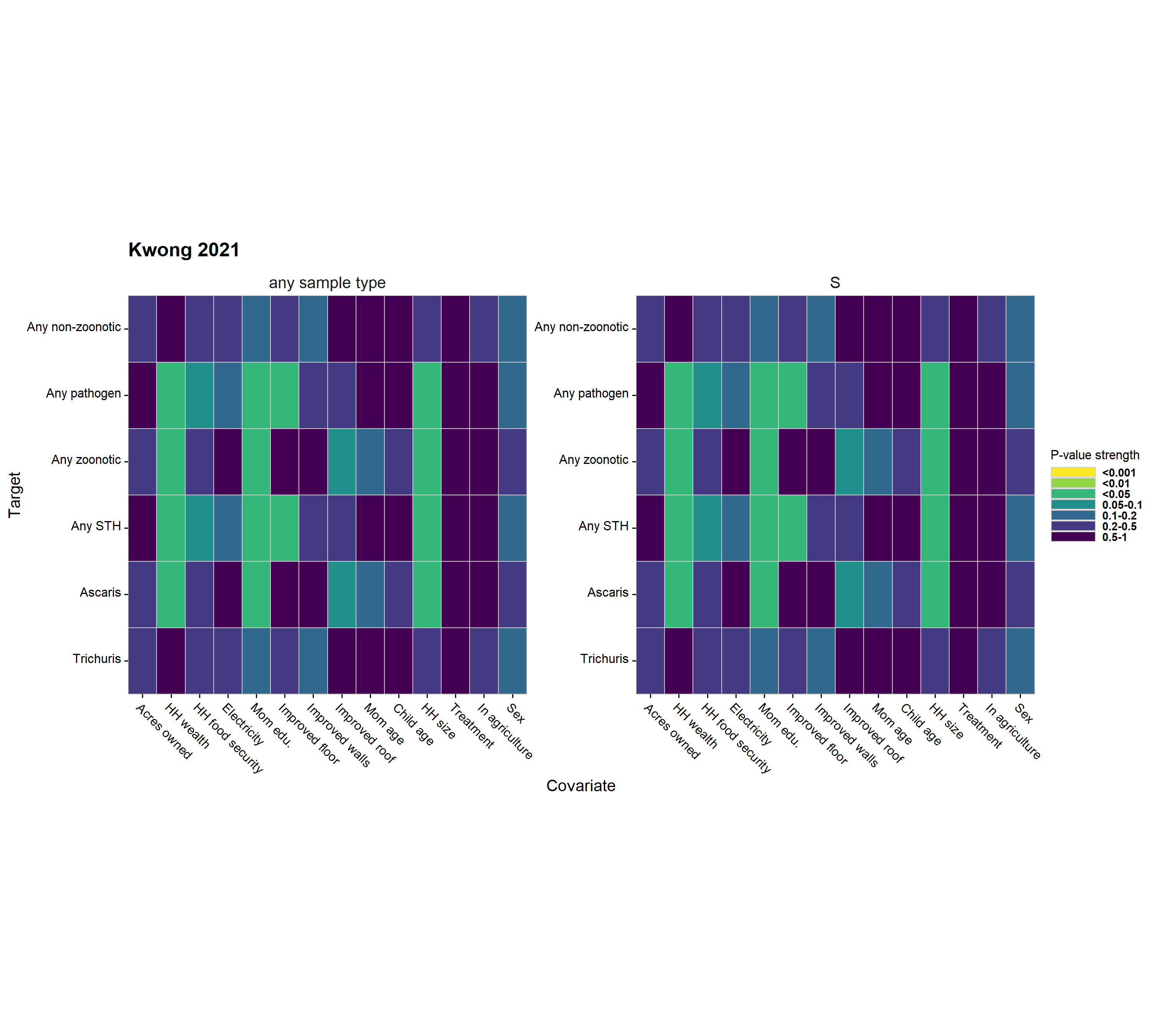
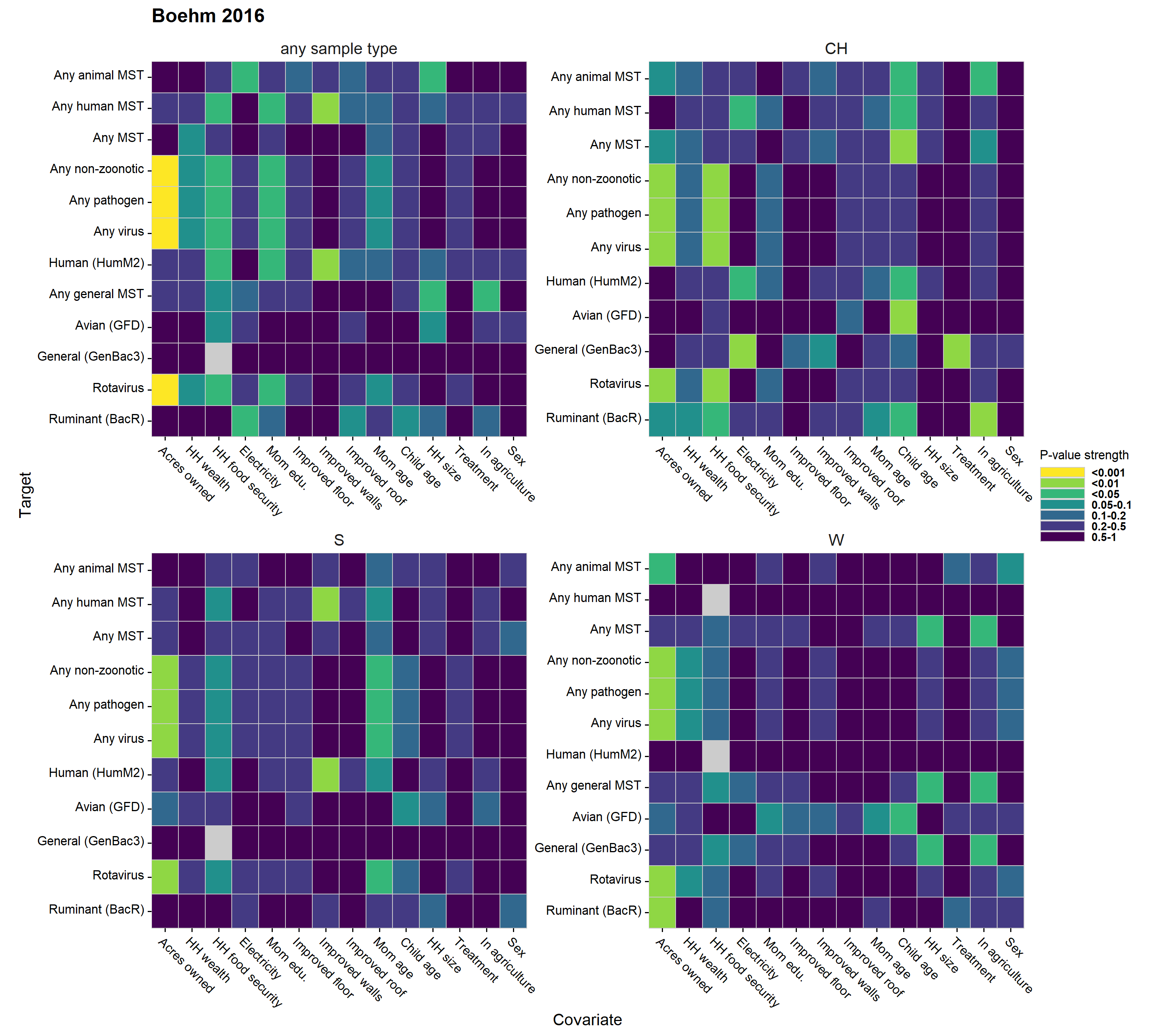
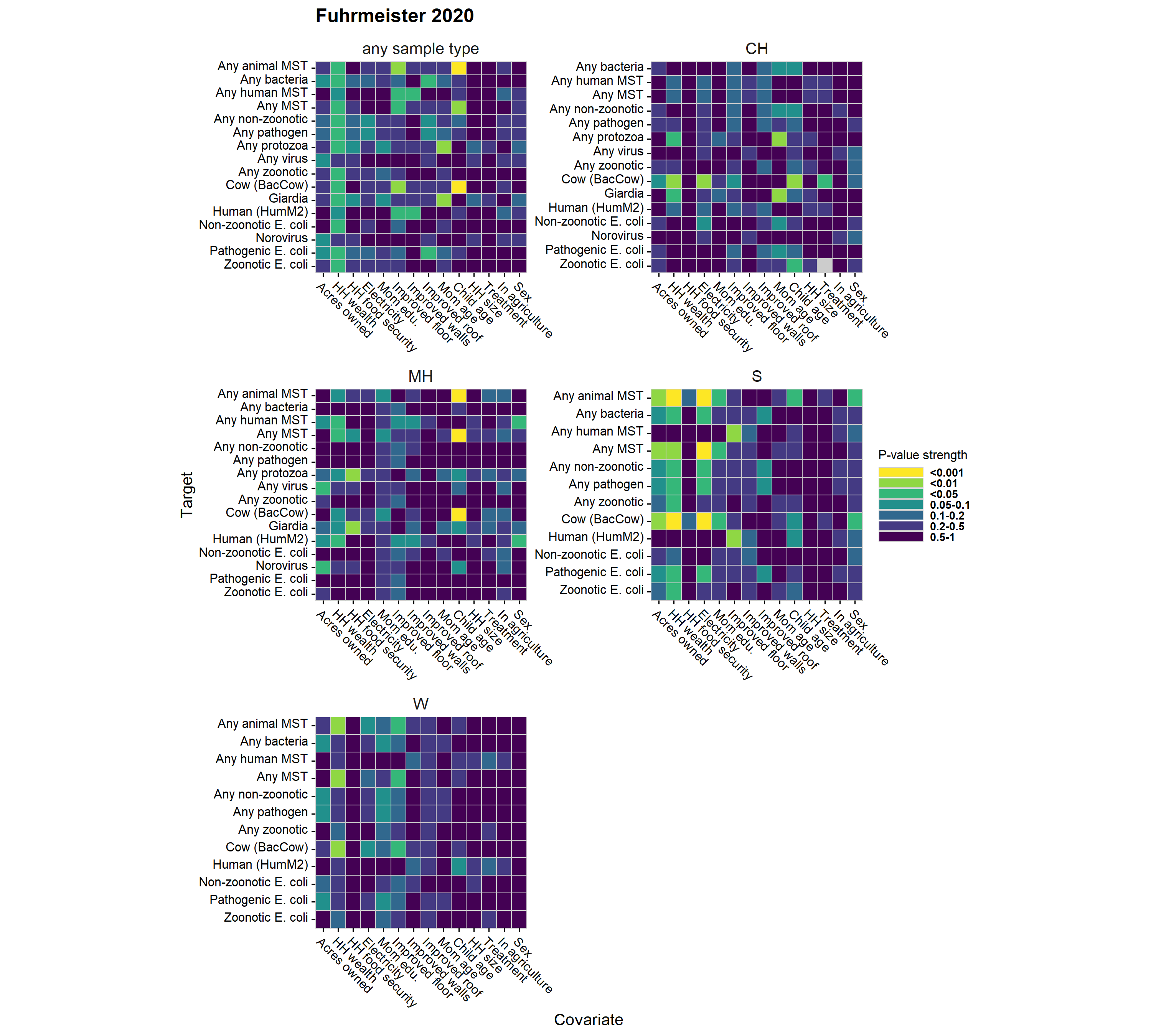


#### TMLE comparison



#### Covariate tables

#NOTE: Why do some Holcomb samples have so few covariates? I think the compound-averaged covariates aren’t being properly merged in. Need to check if there are more Holcomb covariates somewhere



# Tables

## Data availability tables

### Any sample type, any pathogen

| **study** | **sample** | **target** | **N samples** | **N pos. samples** | **N diar. meas.** | **N diar. pos.** | **N sample. and diar. pos.** | **N haz** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Fuhrmeister 2020 | any sample type | Any pathogen | 1,643 | 1,180 | 1,590 | 188 | 128 | 857 |
| Boehm 2016 | any sample type | Any pathogen | 497 | 34 | 412 | 99 | 11 | 411 |
| Kwong 2021 | any sample type | Any pathogen | 2,543 | 1,901 | 703 | 43 | 29 | 758 |
| Steinbaum 2019 | any sample type | Any pathogen | 2,234 | 414 | 1,874 | 485 | 97 | 1,761 |
| Capone 2021 | any sample type | Any pathogen | 566 | 531 | 167 | 21 | 20 | 253 |
| Capone 2021 in prep | any sample type | Any pathogen | 487 | 285 | 195 | 19 | 10 | 213 |
| Reese 2017 | any sample type | Any pathogen | 1,044 | 274 | 84 | 9 | 4 | 578 |
| Odagiri 2016 | any sample type | Any pathogen | 4,825 | 3,787 | 2,038 | 188 | 117 | 0 |

### Any sample type, any MST

| **study** | **sample** | **target** | **N samples** | **N pos. samples** | **N diar. meas.** | **N diar. pos.** | **N sample. and diar. pos.** | **N haz** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Fuhrmeister 2020 | any sample type | Any MST | 1,636 | 1,533 | 1,583 | 188 | 175 | 850 |
| Boehm 2016 | any sample type | Any MST | 497 | 490 | 412 | 99 | 97 | 411 |
| Holcomb 2020 | any sample type | Any MST | 935 | 692 | 292 | 27 | 17 | 412 |
| Capone 2021 in prep | any sample type | Any MST | 487 | 266 | 195 | 19 | 11 | 213 |
| Odagiri 2016 | any sample type | Any MST | 4,825 | 4,672 | 2,038 | 188 | 174 | 0 |

## References

1. Zou, G. A modified poisson regression approach to prospective studies with binary data. *American Journal of Epidemiology* **159**, 702–706 (2004).

2. Freedman, D. A. On The So-Called ‘Huber Sandwich Estimator’ and ‘Robust Standard Errors’. *The American Statistician* **60**, 299–302 (2006).

3. Cochran, W. G. The Combination of Estimates from Different Experiments. *Biometrics* **10**, 101–129 (1954).