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

-Pathogen-specific significant associations -Overall associations are null -specific significant findings

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).

## Methods

We examined associations between prevalence of pathogens and MST markers in the environment and child health outcomes, including all-cause diarrheal disease, child growth, and pathogen-specific infections. The primary outcomes for all exposures were caregiver recall of diarrheal disease and child height-for-age Z-scores. For specific pathogen presences in the environment, primary outcomes also included the corresponding pathogen detection in child stool. Secondary outcomes include z-scores for weight-for-age (WAZ) and weight-for-length (WLZ) and prevalence of stunting, wasting and underweight. For the growth outcomes outcomes, we considered all environmental samples collected over the child’s lifetime prior to the anthropometry measurement. For the diarrheal disease and pathogen-specific infection outcomes, we will only consider environmental samples collected up to four months before the measurement of the health outcome. The analyses was conducted by sample type (e.g., water, hands, soil) and pooled across study types, and used data from all study arms.

For binary outcomes, we estimated prevalence ratios using modified Poisson regressions.1 For continuous outcomes (child anthropometry Z-scores), we used linear regressions to estimate adjusted mean differences. Because of repeated sampling or clustered designs in some studies, we used the Huber Sandwich Estimator to calculate robust standard errors.2 All analyses were adjusted for potential confounders. We included child age and asset-based household wealth as adjustment covariates for all adjusted estimates. Other 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. Within each study, we only estimated associations when there were at least 5 cases of the binary outcome in the rarest strata of the exposure.

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.

### 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. (Give specific numbers for these tables)

### 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: 1.09 (95% CI: 0.90, 1.34)

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

-0.08 (95% CI: -0.15, -0.01)

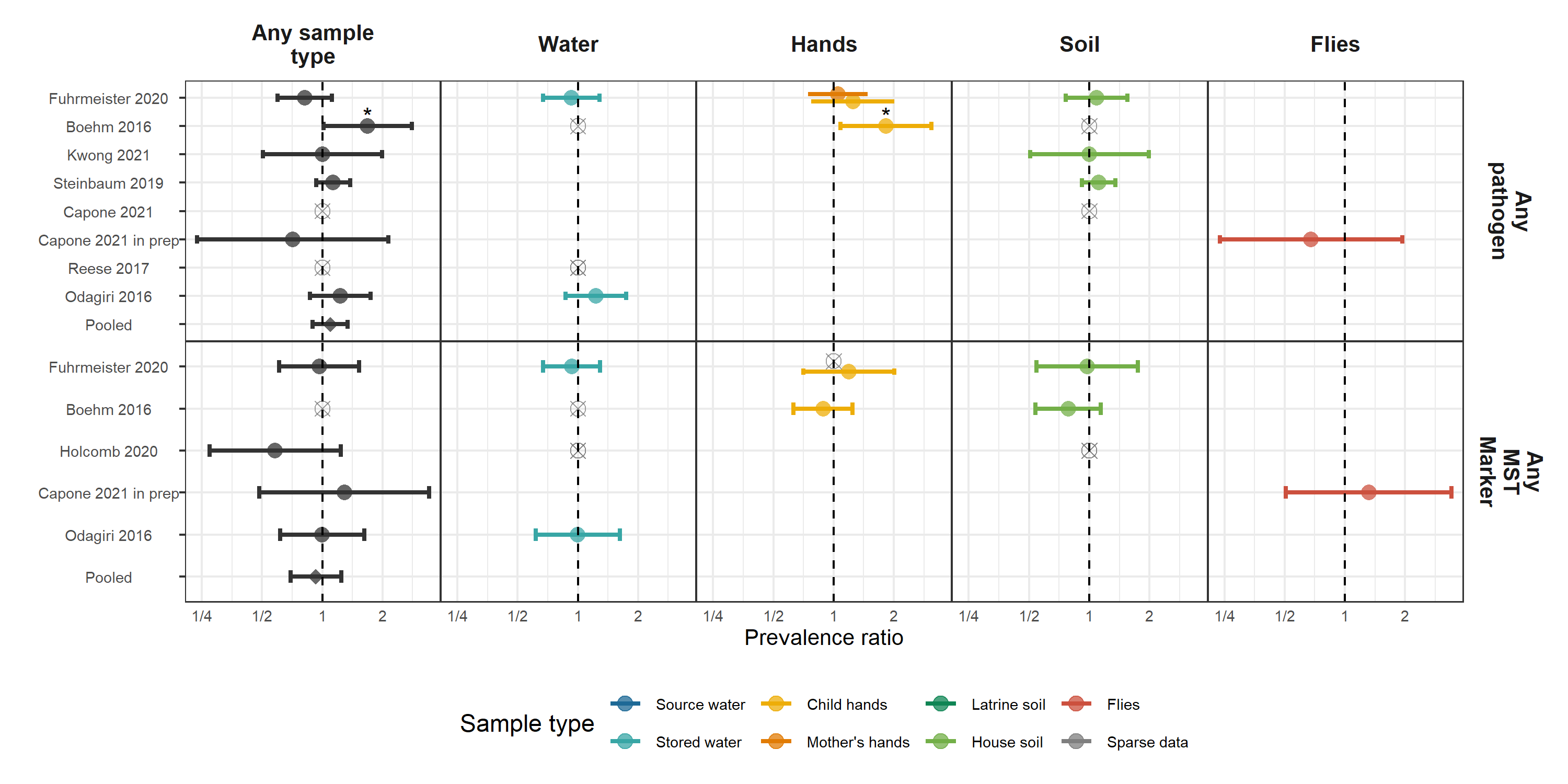
Interquartile range for the ratio of adjusted and unadjusted prevalence rations: [0.97, 1.05] Interquartile range for the difference of adjusted and unadjusted prevalence rations: [-0.014, 0.061]

## Limitations

* Unmeasured confounding
* Small sample size
* Timing of measurements

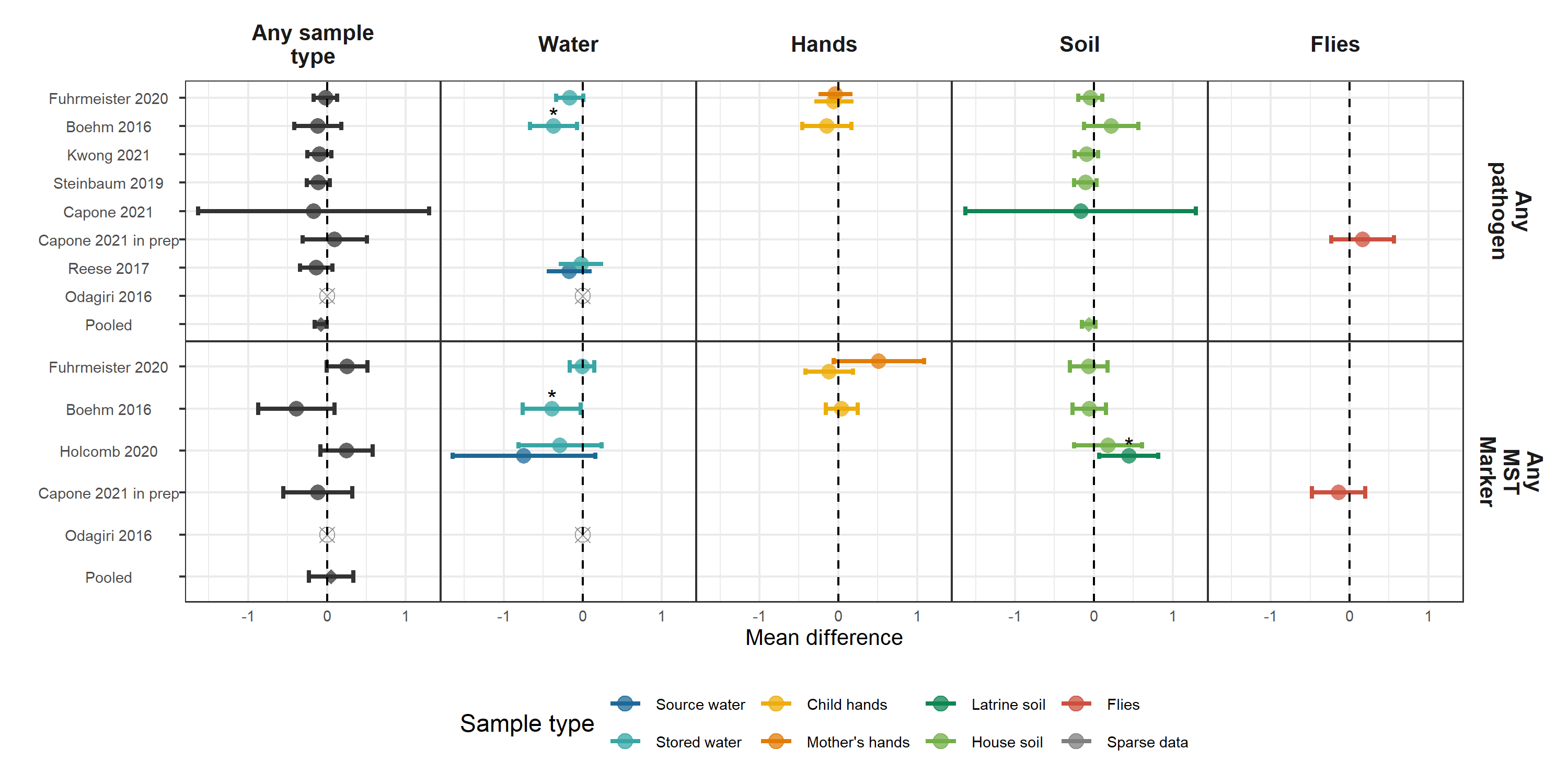
# Primary figures

#### 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. Asterisks above estimates denote statistical significance (\*= P-value < 0.05, \*\*= P-value < 0.01, \*\*\*= P-value < 0.001). All estimates are adjusted for potential confounders.

**Interpretation:** Presence of any pathogen or any mst marker in any environmental samples were not associated with diarrheal disease, except any pathogen presence located on child hands in Boehm et al. 2016.

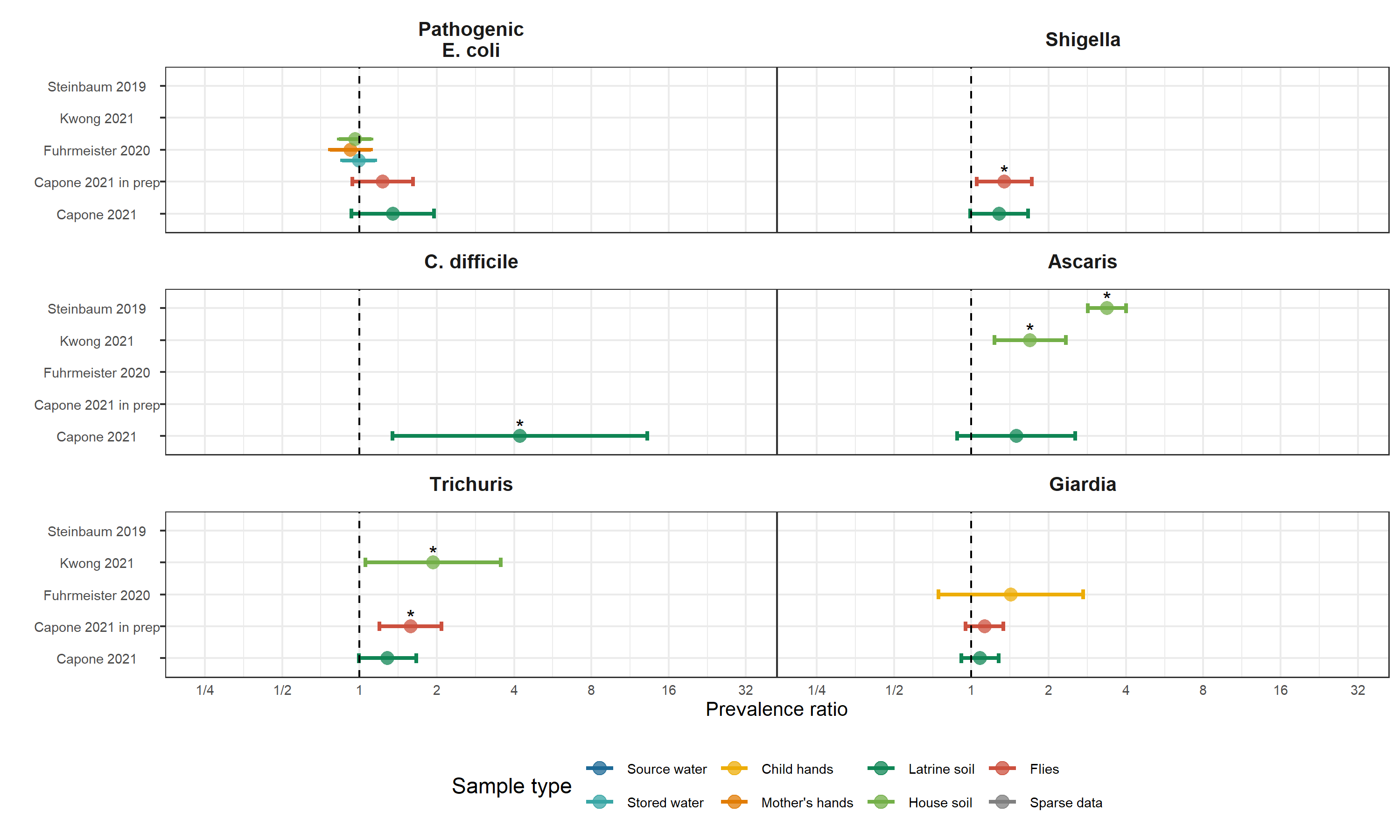


#### Adjusted associations between HAZ and any pathogen or MST marker

**Figure 2.** Forest plots of associations between child HAZ 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 mean difference. 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. Asterisks above estimates denote statistical significance (\*= P-value < 0.05, \*\*= P-value < 0.01, \*\*\*= P-value < 0.001). All estimates are adjusted for potential confounders.

**Interpretation:** Presence of any pathogen (but not any mst marker) in any environmental sample is significantly associated with lower HAZ when pooled across studies (Adjusted mean difference: -0.08 (95% CI: -0.15, -0.01)). This is driven primarily by the number of slightly harmful but insignificant effects rather than by any strong effect of any pathogen in specific studies or sample types. Nevertheless, water samples with any pathogen presence were significantly associated with lower mean HAZ in Boehm et al. 2016. Any MST presence in water was also significantly associated with lower mean HAZ in Boehm 2016, but was associated with higher mean HAZ in latrine soil samples in Holcomb et al. 2020.

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

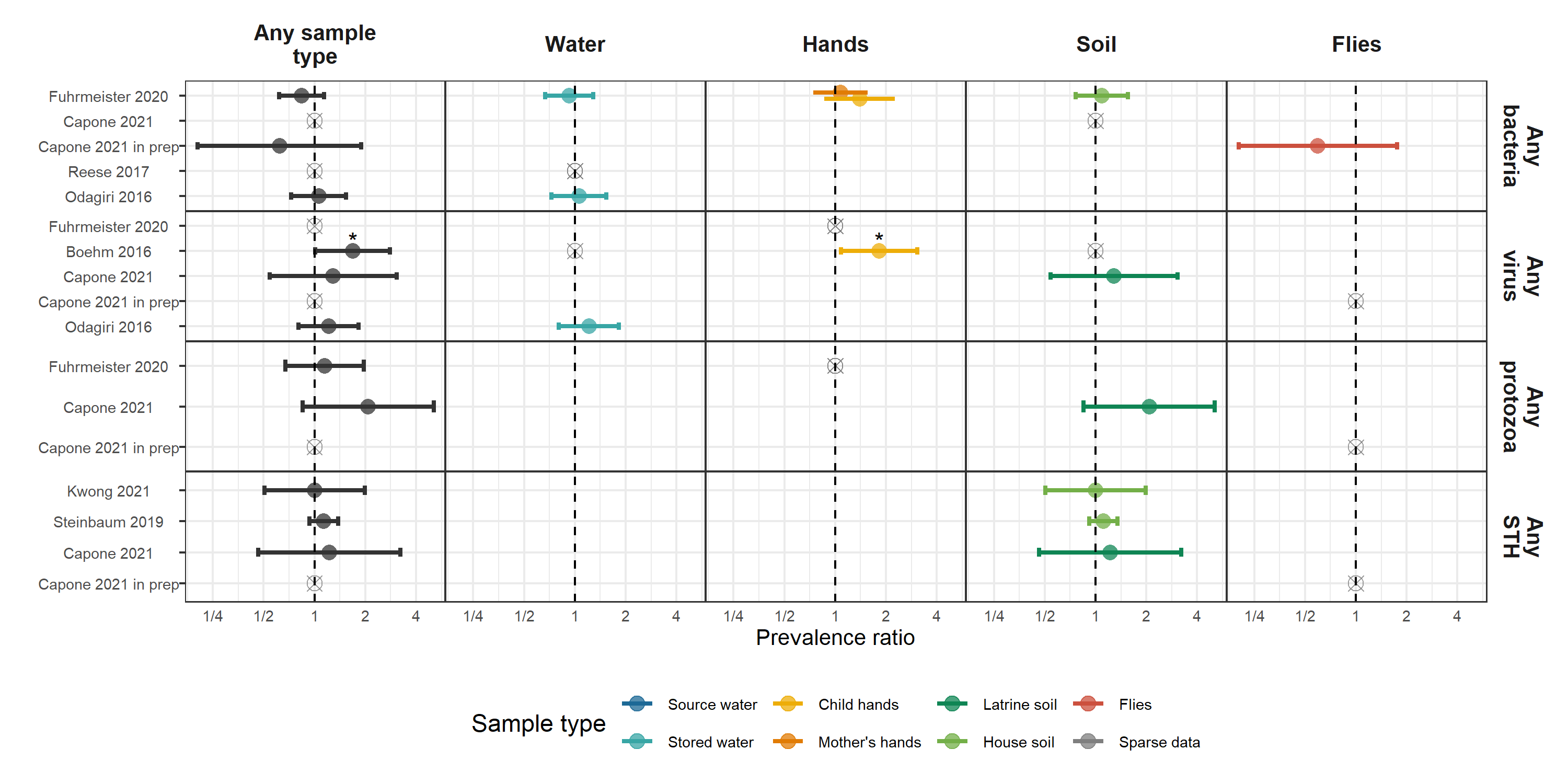


**Figure 3.** Forest plots of associations between specific pathogens in environmental samples and child infections with the same pathogens. Samples of the same type from different locations (source vs. stored water, flies in kitchen vs. latrine, soil from courtyard vs. latrine) or different individuals (child vs. mother’s hands) are plotted separately and denoted by different colors. Asterisks above estimates denote statistical significance (\*= P-value < 0.05, \*\*= P-value < 0.01, \*\*\*= P-value < 0.001). All estimates are adjusted for potential confounders.

**Interpretation:** There is a general trend of positive associations between specific pathogens in the compound environment and an increased risk of the same pathogen infecting the child living in the compound across different pathogens and sample types. *Giardia* and pathogenic *E. coli* were two pathogens without associations between environmental presence and child infection, but associations were significant or near significant for *Shigella*, *Ascaris*, and *Trichuris* contaminations and infections across multiple studies. *C. difficile* was only measured in latrine soil in Capone et al. 2021, but had the strongest association with *C. difficile* infections.

# Secondary figures

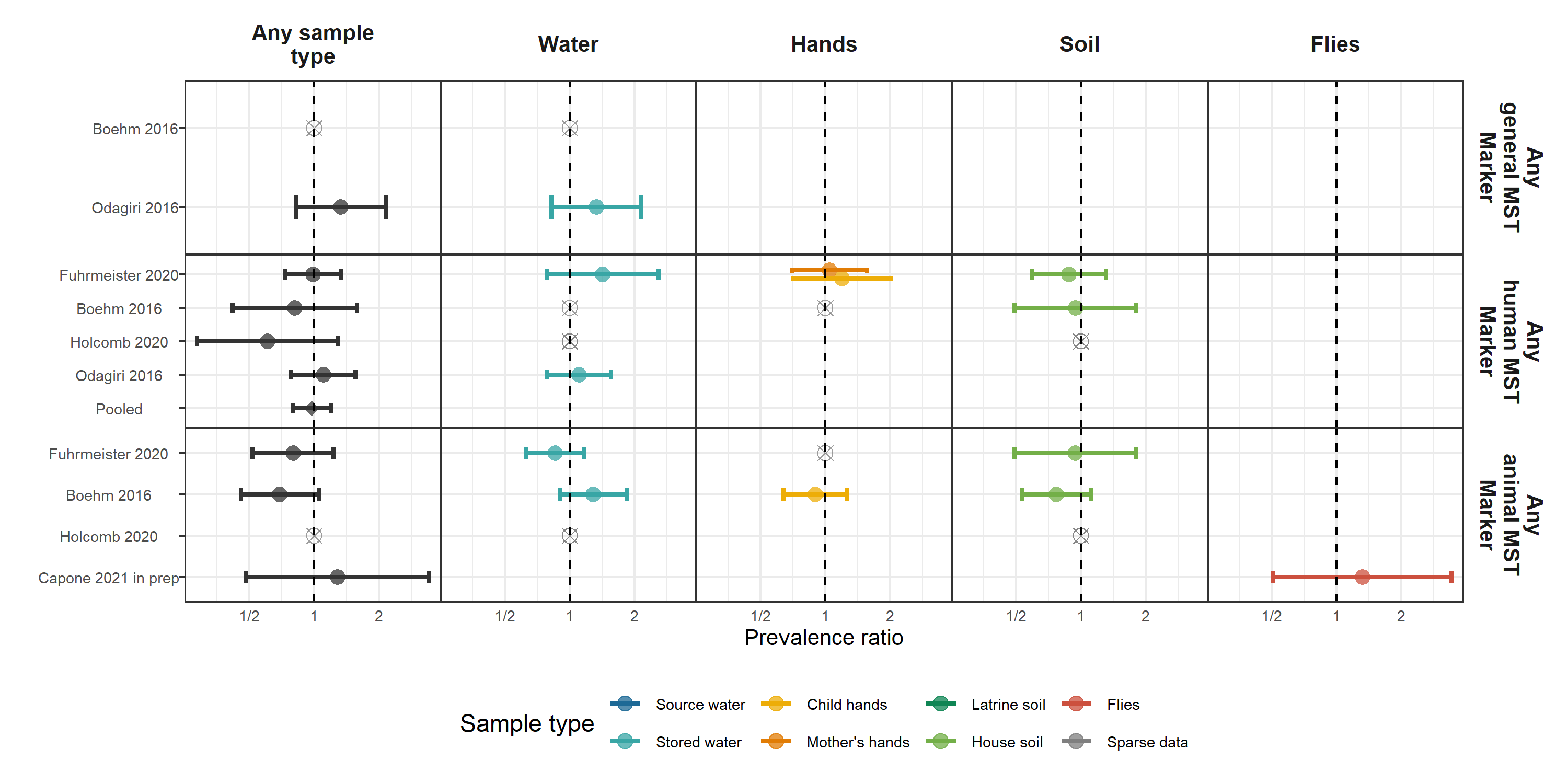
#### Adjusted associations between diarrhea and types of pathogens



**Figure 4.** Forest plots of associations between child diarrheal disease and the prevalence of any virus, any bacteria, any protozoa and any STH 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., <5 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. Asterisks above estimates denote statistical significance (\*= P-value < 0.05, \*\*= P-value < 0.01, \*\*\*= P-value < 0.001). All estimates are adjusted for potential confounders.

**Interpretation:** Presence of types of pathogens in environmental samples were not associated with diarrheal disease, except any viral pathogen presence located on child hands in Boehm et al. 2016.

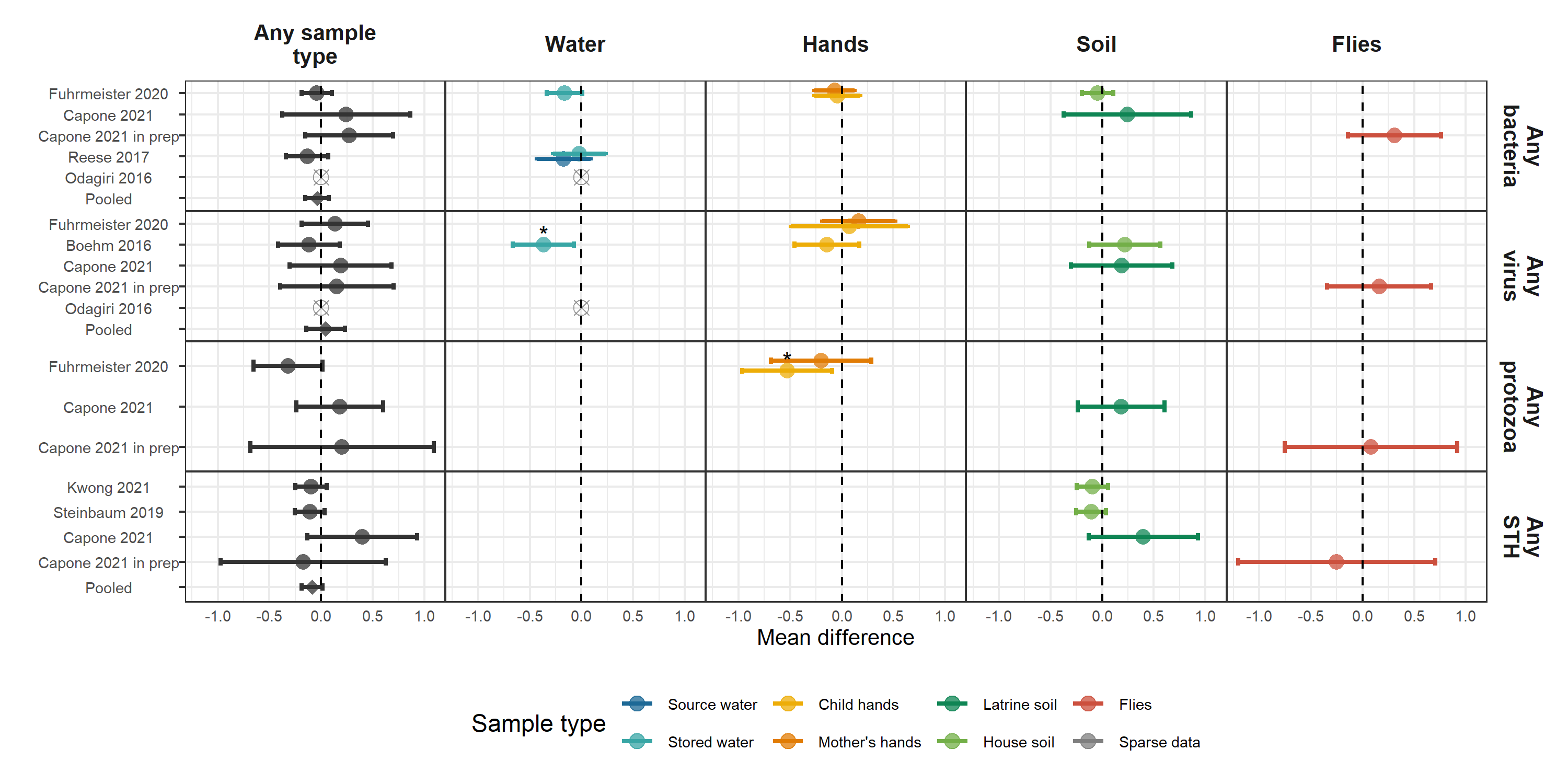
#### Adjusted associations between diarrhea and types of MST markers



**Figure 5.** Forest plots of associations between child diarrheal disease and the prevalence of any general, human, or animal MST 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., <5 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. Asterisks above estimates denote statistical significance (\*= P-value < 0.05, \*\*= P-value < 0.01, \*\*\*= P-value < 0.001). All estimates are adjusted for potential confounders.

**Interpretation:** No associations between specific groups of MST markers and child diarrheal disease in any sample type.

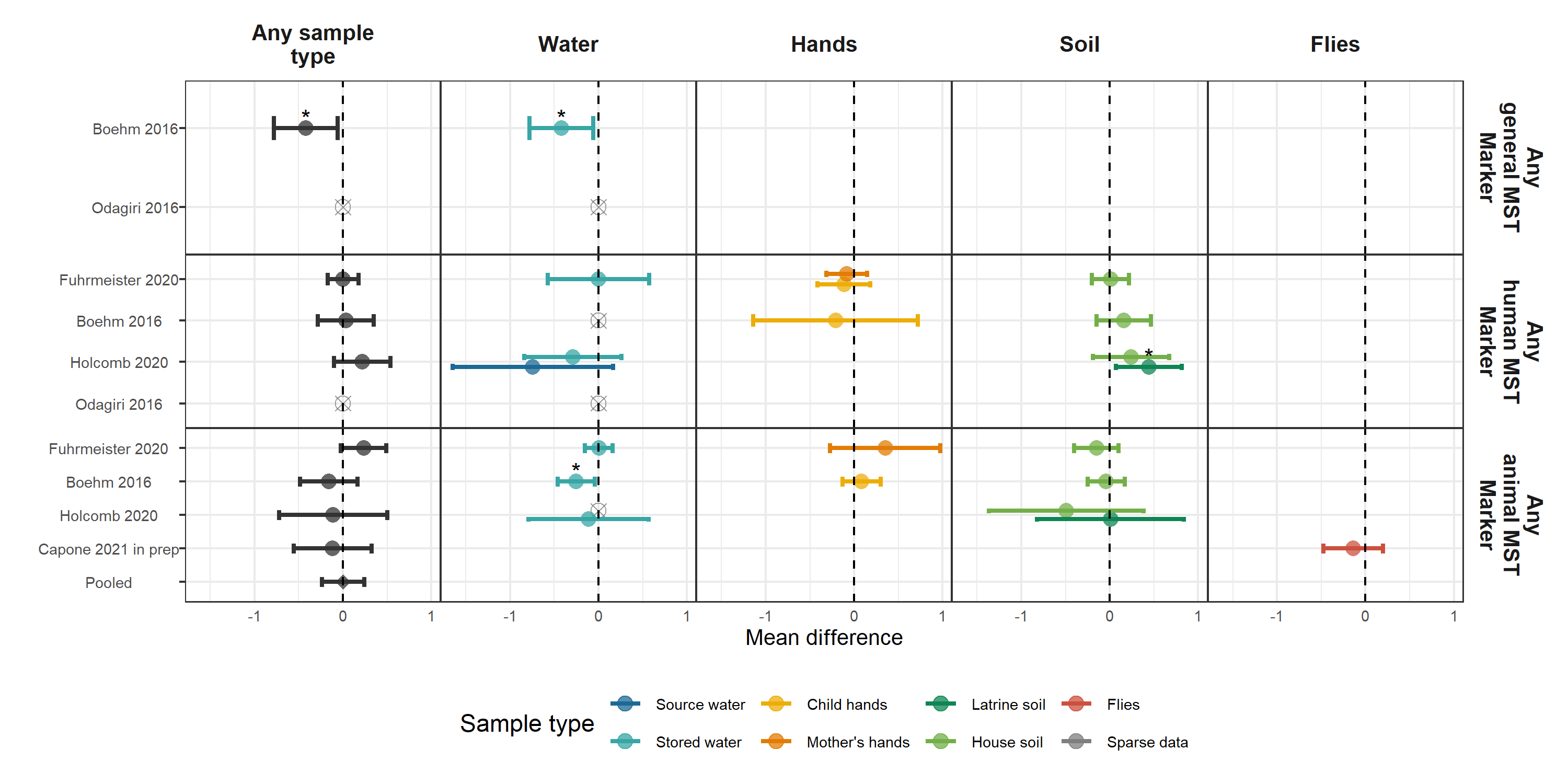
#### Adjusted associations between HAZ and types of pathogens



**Figure 6.** Forest plots of associations between child HAZ and the prevalence of groups of pathogens 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 mean difference. 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. Asterisks above estimates denote statistical significance (\*= P-value < 0.05, \*\*= P-value < 0.01, \*\*\*= P-value < 0.001). All estimates are adjusted for potential confounders.

**Interpretation:** When separated out by group of pathogen, pathogen presence in any environmental sample is no longer significantly associated with lower HAZ when pooled across studies. However, any virus presence in water was significantly associated with lower mean HAZ in Boehm et al. 2016 and any protozoa in water was significantly associated with lower mean HAZ in Furhmeister et al. 2020, both from the WASH Benefits trial.

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

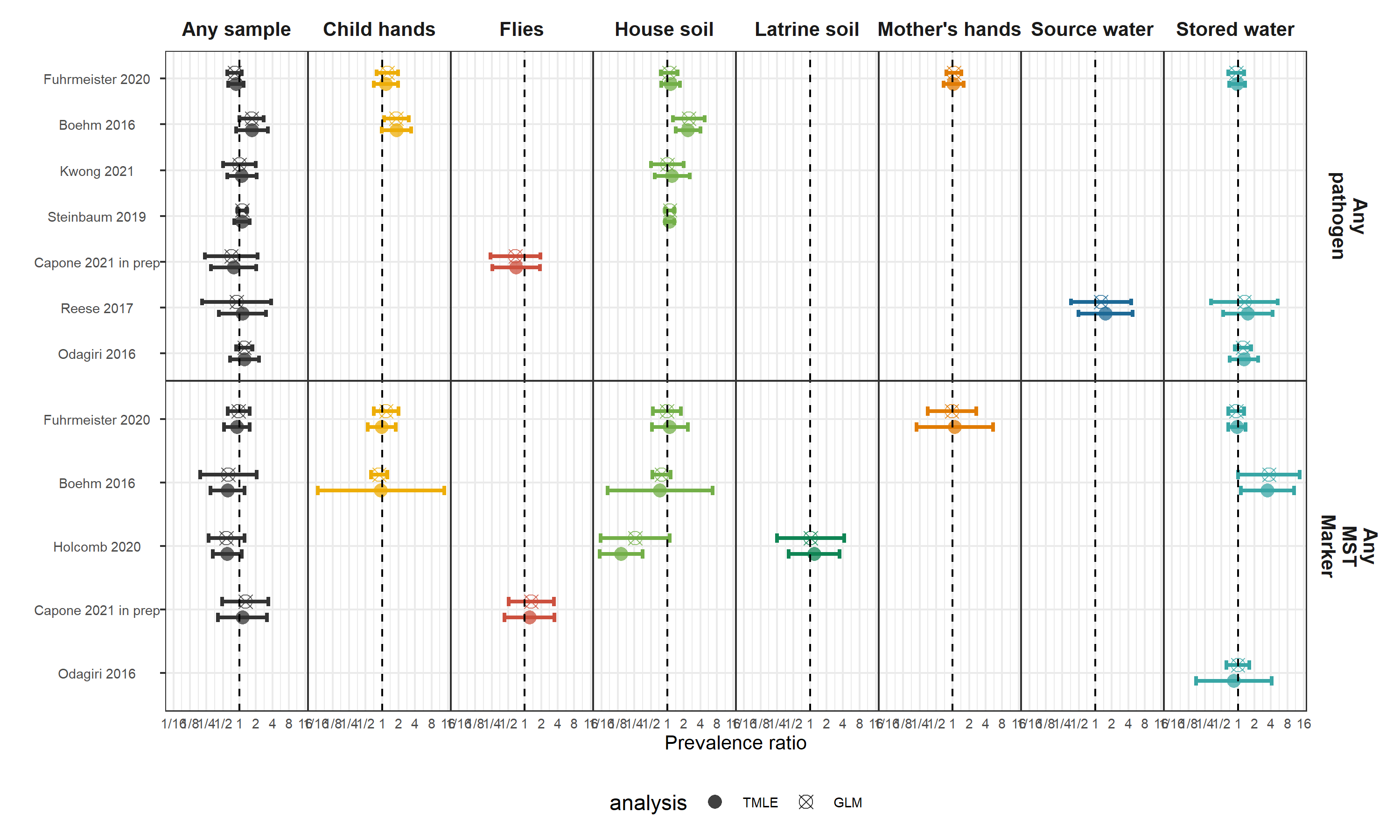


**Figure 7.** Forest plots of associations between child HAZ and the prevalence of groups of MST markers in different types of environmental samples. Pooled estimates are presented when there are four or more study-specific estimates for a specific sample type and target combination and are denoted with diamond-shaped points. Grey crossed points denote data that were too sparse to estimate a mean difference. 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. Asterisks above estimates denote statistical significance (*= P-value < 0.05,* ***= P-value < 0.01,*** = P-value < 0.001). All estimates are adjusted for potential confounders.

**Interpretation:** There is a general trend of presence of groups of MST markers in water samples being associated with lower mean HAZ.

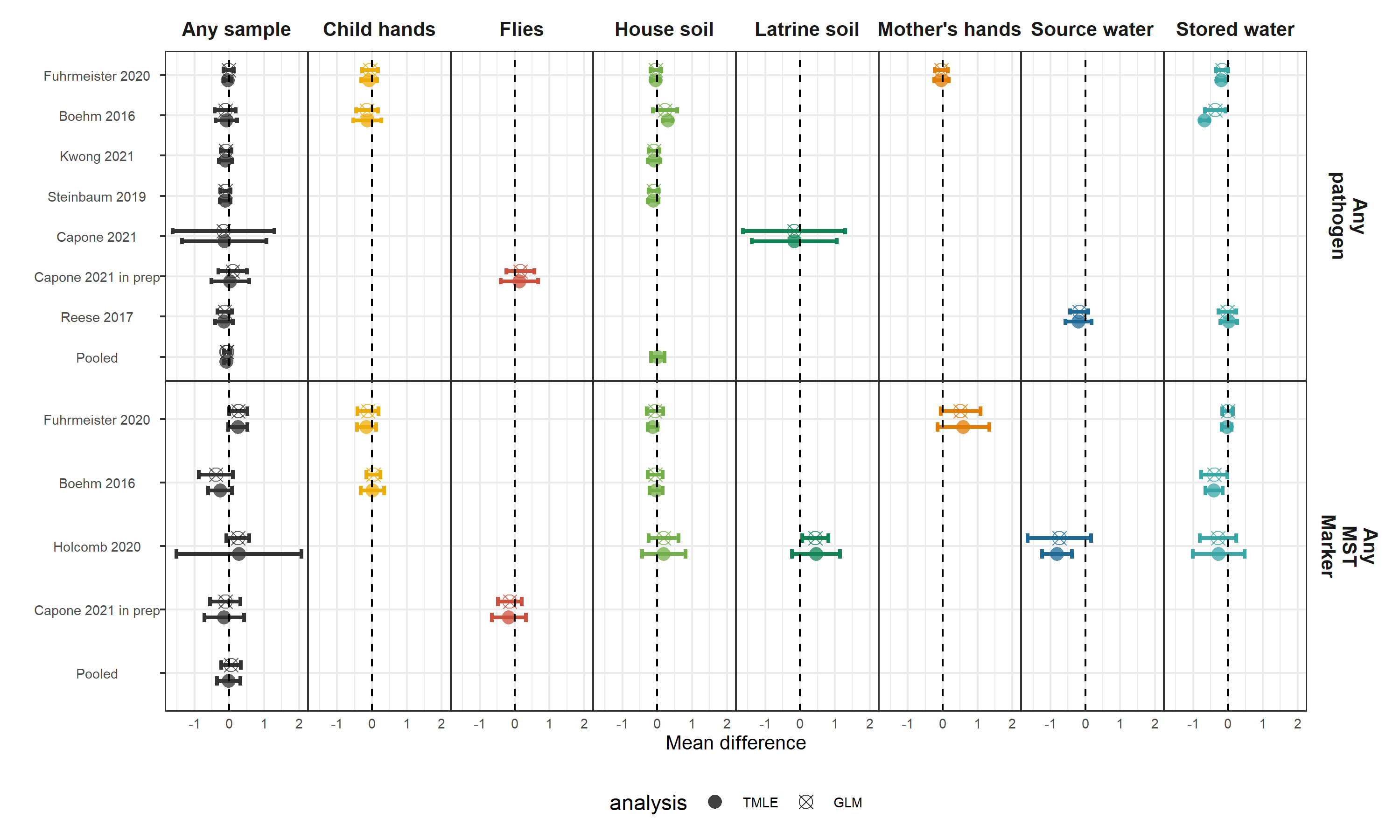
# Supplimentary figures

#### TMLE comparison



**Figure 8.** Comparison between associations estimated with generalized linear models (GLM) and machine-learning based targeted likelihood estimation models (TMLE) for the diarrhea outcome.

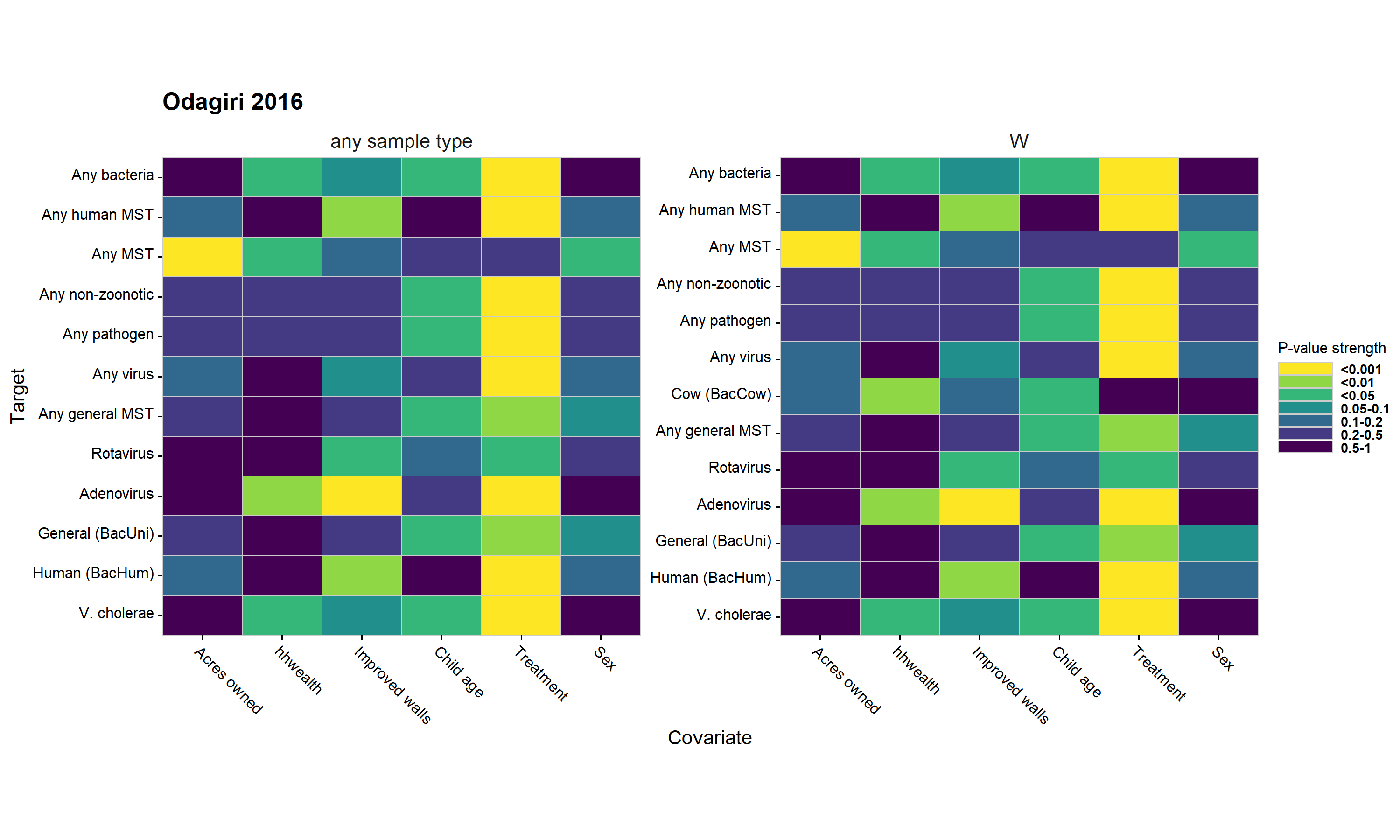
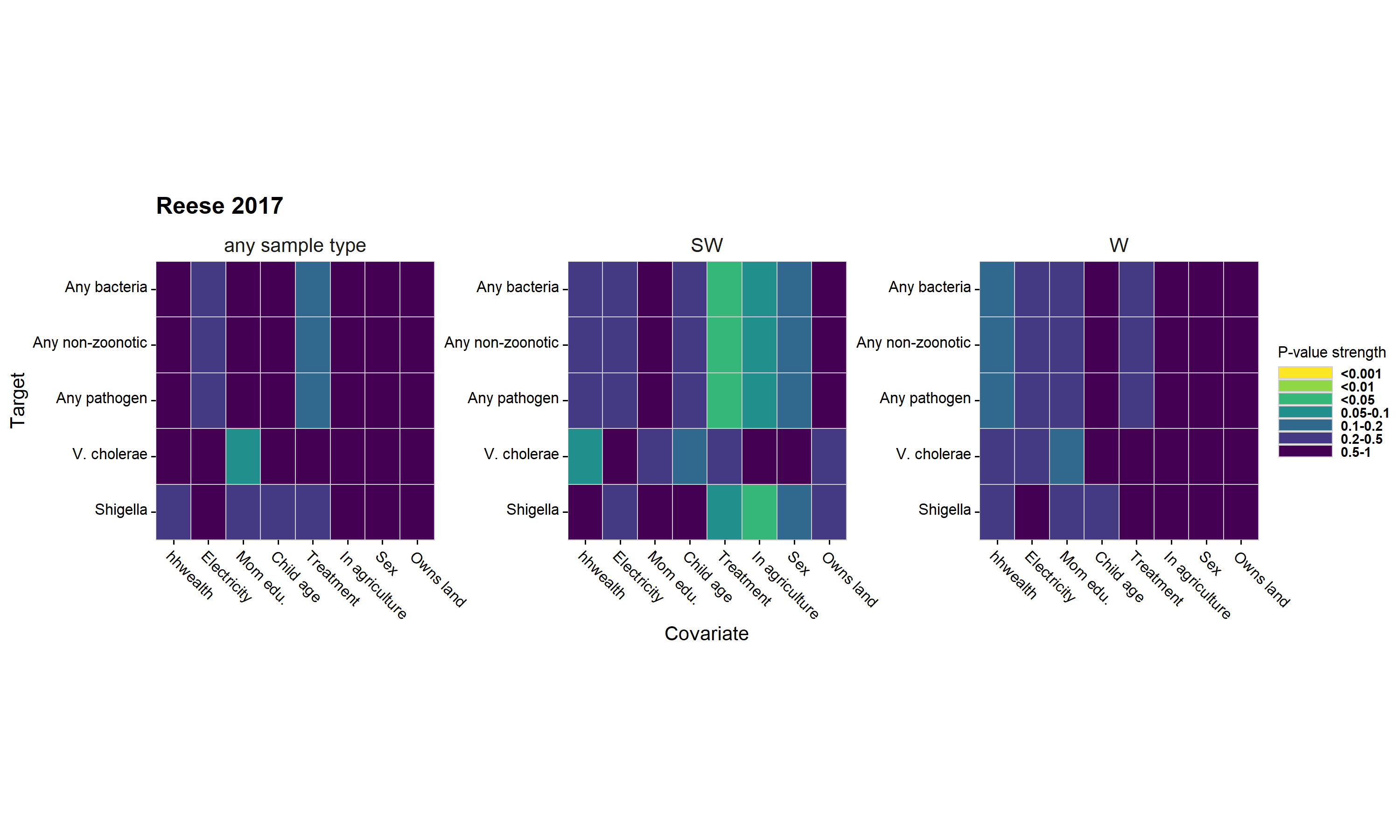
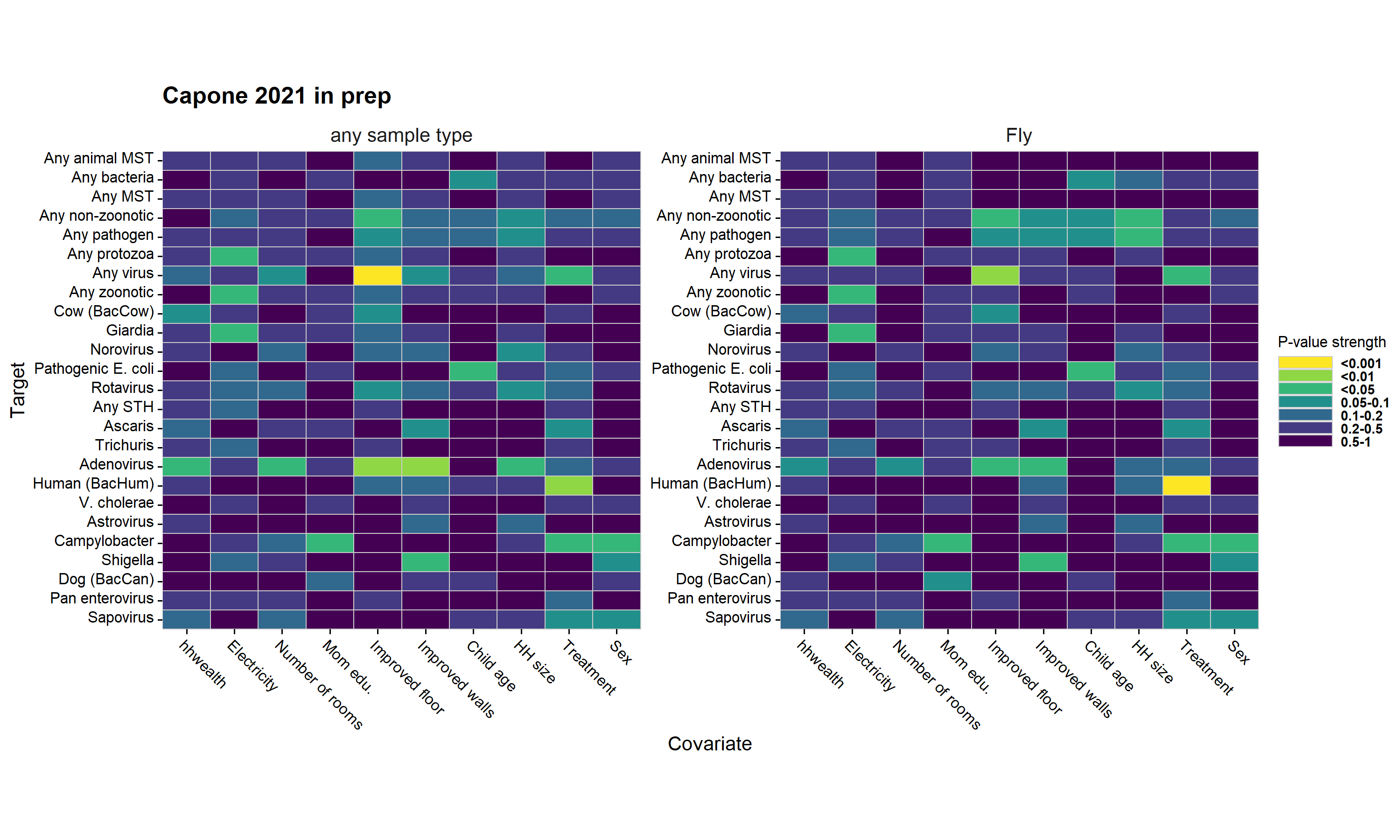
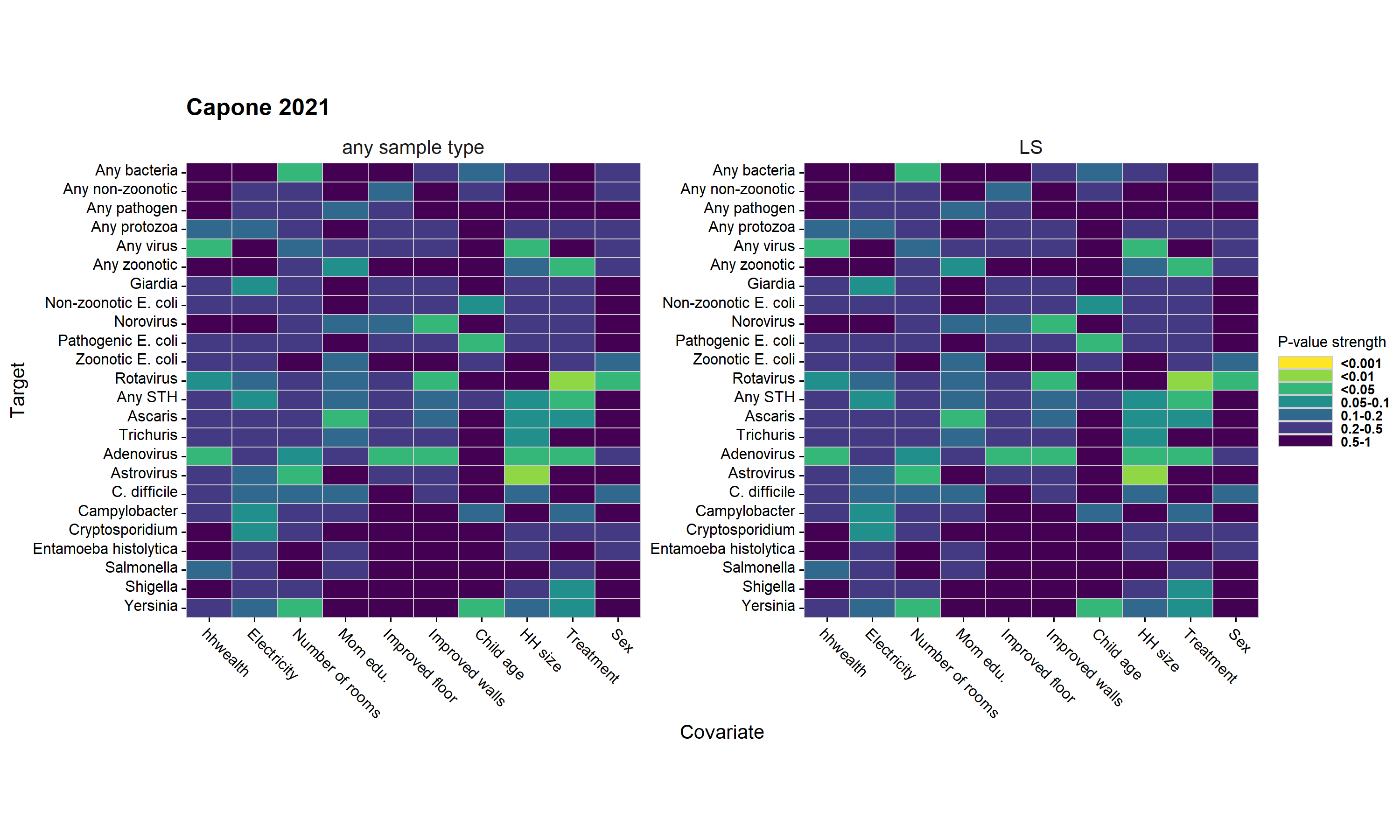
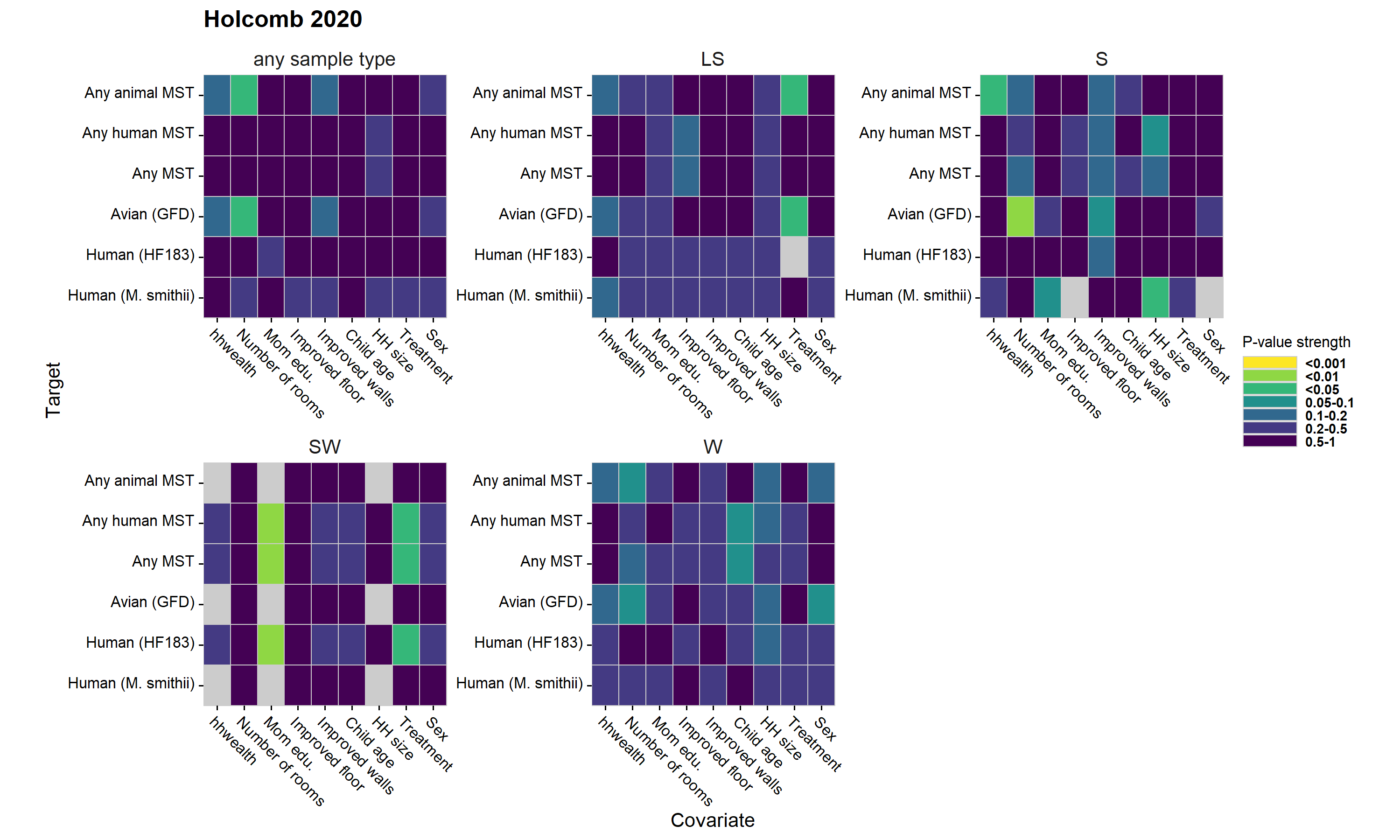
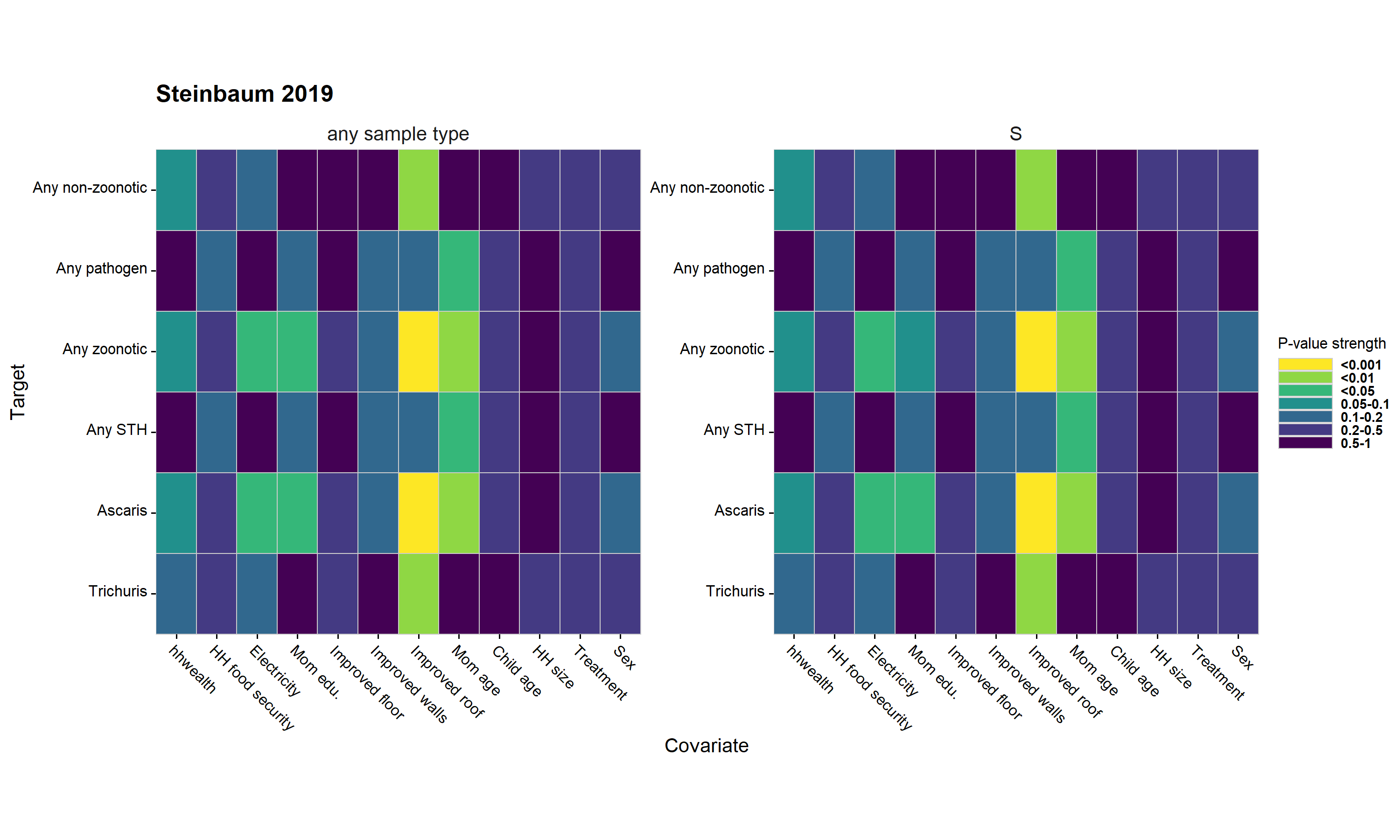
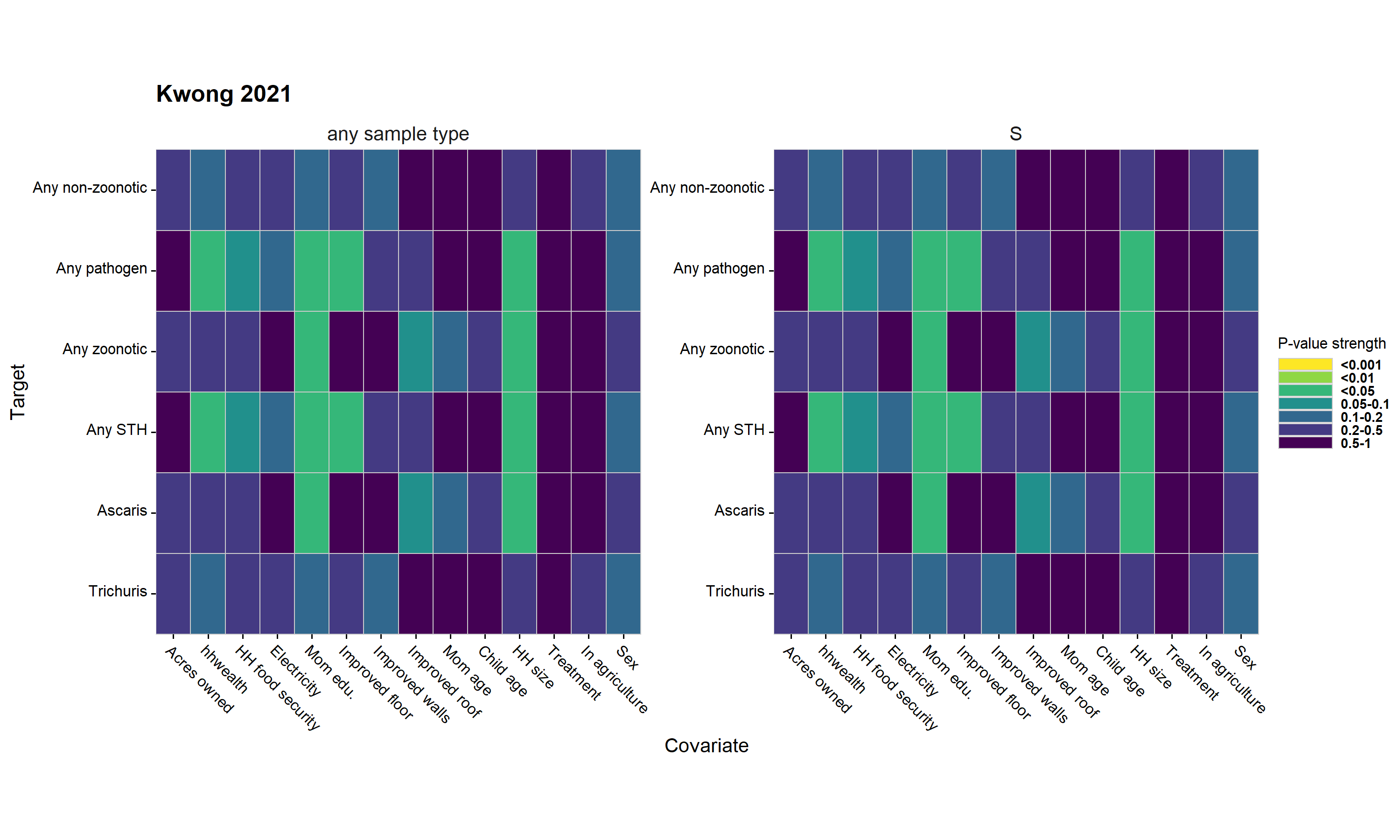
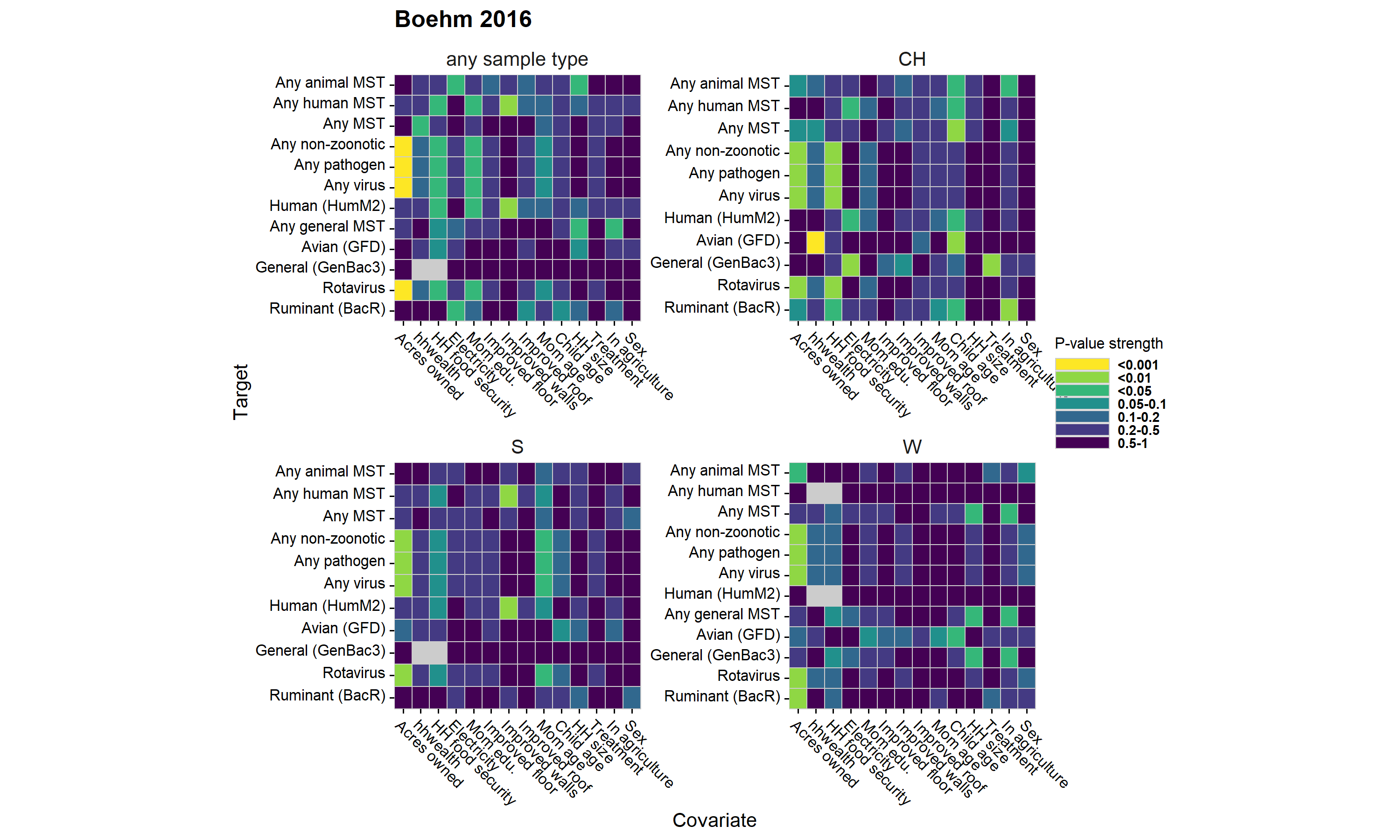
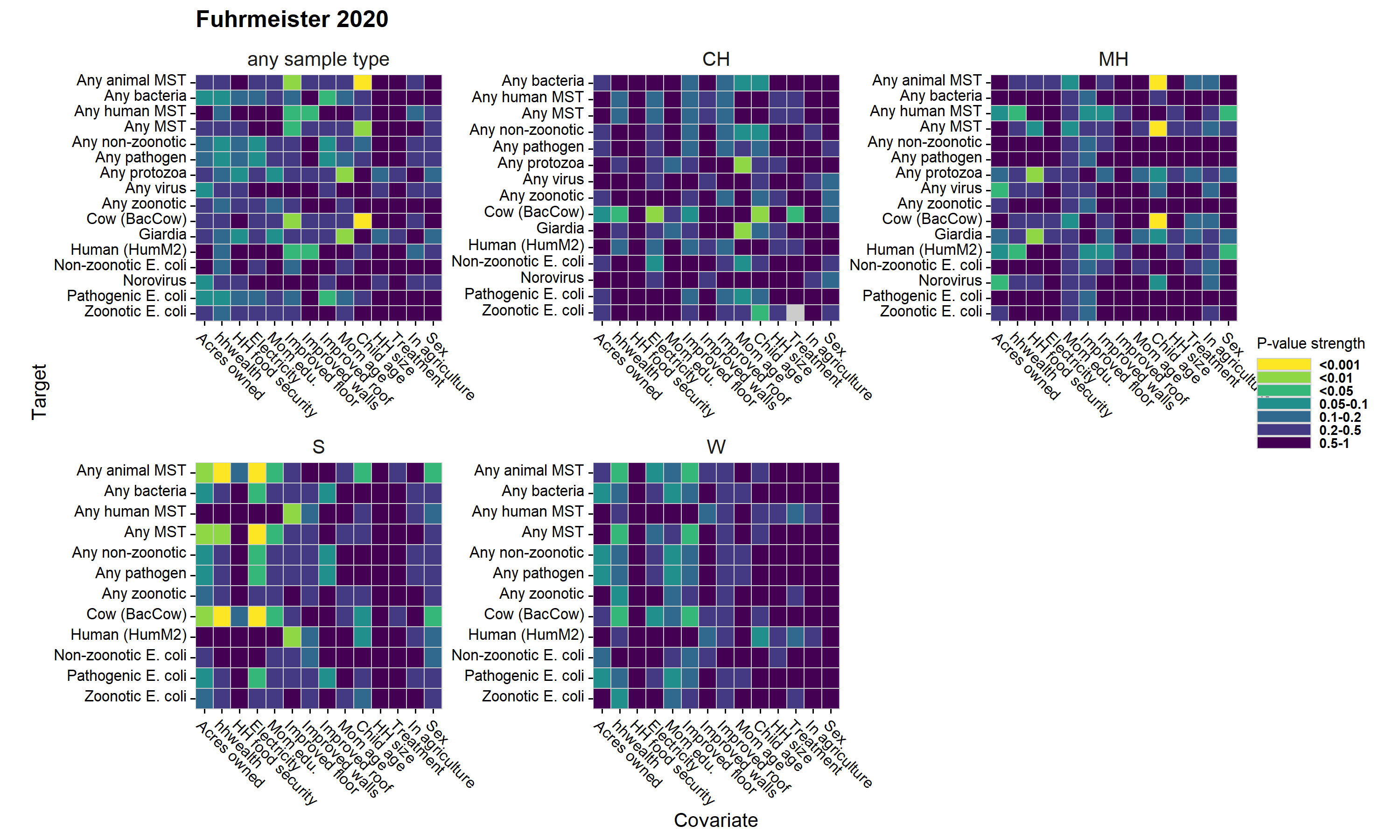
**Interpretation:** The estimation approach chosen did not affect our conclusions about associations between environmental contamination and diarrheal disease.



**Figure 9.** Comparison between associations estimated with generalized linear models (GLM) and machine-learning based targeted likelihood estimation models (TMLE) for the height-for-age Z-score outcome.

**Interpretation:** The estimation approach chosen did not affect our conclusions about associations between environmental contamination and child linear growth.

#### Covariate tables



**Figure 10.** Associations between adjustment covariates and the presence of different pathogen and MST markers in different environmental samples. The columns are different pre-screened confounders, and the rows are specific pathogens and MST markers. Cells of the heatmaps are colored by P-values of bivariate likelihood ratio tests, and heatmaps are stratified by sample type.

**Interpretation:** Most covariates were not strongly associated with pathogen or MST marker presence in the environment, meaning they were not strong confounders of the relationship between environmental contamination with pathogen or MST markers and child infections or poor growth. Measures of household wealth generally had the strongest association with environmental contamination, though the association varied by study, sample, and microbial target. Between the low association between covariates and environmental contamination, and the generally limited differences between unadjusted, adjusted, and TMLE estimates, we believe our modeling approach adequetly adjusted for measured confounding, but unmeasured confounding may bias the results.

# 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 |

**Interpretation:** Due to the smaller sample size of the environmental samples within the WASH trials and quasi-randomized studies, the rarity of diarrheal disease in children, and the rarity of many of the pathogen in environmental samples, data sparsity affected what was possible in this analysis. Many exposure-outcome associations were not estimated due to data sparsity, and others were estimated but could only be adjusted for a subset of potential confounders.

## 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).