Water Sensitive Cities proposal: Unified statistical analysis plan

19/9/2016

All environmental, health and well-being outcome measures in Objectives 2, 3 and 4 of the proposal can be categorised as being of continuous (approximately normally distributed, possibly after logarithmic transformation), binary or count scales. Therefore we present a unified approach to the statistical analysis of these outcomes.

We note that space constraints in the proposal document resulted in only brief detail of methodology for physical and living environment outcome measurements of Objective 2 being provided. We therefore provide greater detail of these outcome measurements in Table 1 at the conclusion of this document.

Baseline (pre-intervention-period) comparisons

Comparison of the intervention and control arms prior to the implementation of the intervention with respect to demographics, health and environment characteristics will be performed using summary statistics and graphical displays, both at site (i.e. settlement) and individual person levels, in accordance with CONSORT guidelines [1]. A flow diagram will be prepared describing intervention allocation, receipt of intervention, numbers of individuals and environmental samples, and losses to followup at both site and individual person levels per arm.

Effects of the WSC intervention on outcomes

Estimation of the effect of the WSC intervention for each outcome measure will employ generalised estimating equations using the repeated post-construction-period measurements as the dependent variables. For human health and well-being outcomes these dependent variables will be the repeated measurements of individuals, and for environmental outcomes these will be the measurements of environmental samples within sites, or of site-level outcomes (e.g. vertebrate communities). Models will incorporate intervention arm as the covariate of interest; with city of location, tidal/non-tidal location and pre-construction-period values of each outcome measure (aggregated at site level) used as baseline covariates[1]; and with temporal adjustment

[1] By noting that analysis of covariance is equivalent to a repeated measurements model with the baseline as part of the outcome variable and covariates containing time and the interaction of intervention arm with time, but without the main effect of intervention arm (i.e. setting the difference in the intervention and control groups in the pre-construction-period to

using indicator variables for seasonal patterns (e.g. rainy vs dry season), local rainfall data for each site, and an indicator term for any major flood event at a particular site in a particular period. An exchangeable working correlation structure will be utilised with site as the clustering unit, and with robust standard errors scaled with a small-sample degrees-of-freedom adjustment [3]. For binary and count outcome measures, we will use a logarithmic link to estimate ratios of prevalence and ratios of mean counts, respectively; for continuous outcomes, the identity link will be used to estimate differences in means. Primary analyses will use the post-construction-period outcomes from a priori specified lag periods for the effect of the intervention to be displayed (e.g. measurements ≥ 12 months for environmental enteric dysfunction markers), with secondary and more exploratory analyses using a variety of lag periods, plus the construction period for assessment of intervention effects during construction.

Loss to followup is not an issue for environmental sampling. For human health outcomes, the major concern is dropout of entire dwellings, in which case we will employ multiple imputation using imputation models that preserve the hierarchical data structure.

Effects of environmental changes on health outcomes

We will explore relationships between environmental changes and health outcomes by modelling health outcomes at individual level at 12 and 24 months post-construction, using changes in aggregated site-level environmental measures at 6, 12, or 18 months post-construction as principal covariates. Specifically, for health at 12 months post-construction we will assess association with changes in the environment from the pre-construction period to 6 months post-construction; and for health at 24 months we will assess environmental changes from pre-construction to each of 6, 12 and 18 months post-construction. We will estimate these effects using generalised estimating equations with pre-construction health and environmental measurements as baseline covariates aggregated at site level. We will assess potential effect modification of the intervention effect by including interactions of environmental changes with intervention arm, together with adjustment for potential confounders.

Assessment of pathways of action of WSC intervention on health and well-being outcomes

We will capitalise on the longitudinal data structure in this trial to explore the pathways of action of the implementation of the WSC intervention on human health and well-being outcomes, with particular focus on mediation by environmental

be zero) [2], there is alignment between the analysis model specified here (which uses the former) and the model in the sample size calculations (which uses the latter).

changes. Specifically, the broad aim of these analyses will be to estimate the proportion of the effect of the WSC intervention on human health and well-being that occurs via the pathway

WSC intervention → environmental changes → changes in health and well-being, as compared to the proportion of the effect on health and well-being due to other pathways not involving the environment. In the ecology discipline these analyses are typically performed using 'Bayesian belief networks' [4] , whereas in epidemiology the principal quantities are entitled 'direct' and 'indirect' effects and estimated using 'causal mediation analyses' [5]. Both approaches are based on directed acyclic graphs (DAGs), but with differing amounts of data-driven determination of the existence and magnitudes of relationships between variables. During the first year of this trial we will review the commonality and differences in these approaches and develop an analysis approach that appropriately combines them. During this time we will also refine the specific health outcomes to be considered and the specific environmental measurements that would be their plausible mediators.

Interim monitoring

Interim reports for safety considerations will be supplied to the Data Monitoring Committee (DMC). No formal interim analyses of the effect of the WSC intervention are planned. A DMC Charter will be developed in conjunction with DMC members describing the reporting frequency and contents.

REFERENCES:

1. Campbell MK, Piaggio G, Elbourne DR, et al. CONSORT 2010 statement: extension to cluster randomised trials. 2012
2. van Breukelen GJ. ANCOVA versus CHANGE from baseline in nonrandomized studies: The difference. Multivariate Behavioral Research 2013;48(6):895-922
3. Mancl LA, DeRouen TA. A covariance estimator for GEE with improved small-sample properties. Biometrics 2001;57(1):126-34
4. McDonald K, Ryder D, Tighe M. Developing best-practice Bayesian Belief Networks in ecological risk assessments for freshwater and estuarine ecosystems: a quantitative review. Journal of environmental management 2015;154:190-200
5. VanderWeele T. Explanation in causal inference: methods for mediation and interaction: Oxford University Press, 2015.

TABLE 1: Measurements and statistical approaches for physical and living environment outcomes in Objective 2

AREA/QUESTION	**APPROACH**	**REFERENCES**
Physical environment		
Air	Generalized Estimating Equations, detrended data, including spatial autocorrelation effects. All implementation in the R statistical environment.	Denny & Gaines 2002. Chance in Biology. Using Probability to Explore Nature, Princeton Univ. Press. Bini et al. 2009 Ecography. Zuur et al. 2009. Mixed Effects Models and Extensions in Ecology with R. Springer, Berlin.
Water	Generalized Estimating Equations, detrended data, including spatial autocorrelation effects. Multivariate approaches: permutational multivariate analyses of variance and of distance matrices.	As above. R packages VEGAN for multivariate analyses, see Oksanen et al. 2015 vegan: Community Ecology Package. R package version 2.3-0. Available at: http://CRAN.R-project.org/package=vegan
Remote sensing	Spatial processing using ARCGIS and R packages (such as raster). Some processed information obtained from European Space Agency EO4SD. Mixed effects models depending on data form and requirements for detrending and spatial autocorrelation.	Zuur et al. 2009. Mixed Effects Models and Extensions in Ecology with R. Springer, Berlin.. Hijmans & van Etten 2015 raster: Geographic analysis and modeling with raster data. http://CRAN.R-project.org/package=raster. For recent UHI example, Duffy & Chown 2016 J. Ecol.

Living environment		
Microbial diversity (post 16s & 18s pipelines)	Zeta diversity. Shimadzu's approach for temporal community turnover and significant species. Generalized Estimating Equations or, depending on questions, Generalized Linear Models assuming Poisson or Negative Binomial Distributions.	Crawley 2012 The R Book. Wiley. Hui & McGeoch 2015 Am. Nat. Latombe et al. R package Zetadiv https://cran.r-project.org/web/packages/zetadiv/index.html. Shimadzu et al. 2015. Methods. Ecol. Evol. Zuur et al. 2009. Mixed Effects Models and Extensions in Ecology with R. Springer, Berlin.
Mosquito populations	Generalized Estimating Equations, or, depending on the question, Generalized Linear Models assuming Poisson or Negative Binomial Distributions. If zeros, Zero-Inflated Hurdle Models. GEEs if spatial autocorrelation problematic.	Bini et al. 2009 Ecography. Zuur et al. 2009. Mixed Effects Models and Extensions in Ecology with R. Springer, Berlin. Crawley 2012 The R Book. Wiley.
Rodent populations	Spatially explicit capture-recapture methods. Generalized Estimating Equations or, depending on the question, Generalized Linear Models assuming Poisson or Negative Binomial Distributions. If zeros, Zero-Inflated Hurdle Models. GEEs if spatial autocorrelation problematic.	Stevenson et al. 2015 Methods Ecol. Evol. Bini et al. 2009 Ecography. Zuur et al. 2009. Mixed Effects Models and Extensions in Ecology with R. Springer, Berlin. Crawley 2012 The R Book. Wiley.

Bird and bat specific species	Spatially explicit capture-recapture methods, followed by GEE or GLMs as above.	Stevenson et al. 2015 Methods Ecol. Evol. And methods as above.
Acoustic environment or soundscape	Acoustic Complexity Index. Principal Components Analyses and Generalized Linear Models.	Pieretti et al. 2013 J. Acous. Soc. Am. Tobias et al. 2014 PNAS.
Human thermal burden	Universal Thermal Climate Index, related indices, thermodynamic niche models for initial load estimation. Comparison using GEEs.	Blazejczyk et al. 2012 Int. J. Biometeorol. Kearney et al. 2013 Funct. Ecol. Chown & Duffy 2015 Funct. Ecol.
Integration		
Objectives 2-4	Causal mediation methods and Bayesian belief networks, which enable integration of qualitative and quantitative data from different sources.	Van der Weele 2015 Explanation in Causal Inference: Methods for Mediation and Interaction. Oxford University Press. McDonald et al. 2015. J. Env. Manag. Pullin et al. 2016 Biodivers. Conserv.