VNP19

VALUING NATURE PROGRAMME

Demystifying Health Metrics

Valuing Nature Paper | October 2019

Demystifying Health Metrics

Valuing Nature Paper | October 2019

Lead Authors: Deborah Cracknell, Rebecca Lovell, Benedict Wheeler and Mathew White *European Centre for Environment & Human Heath, University of Exeter*

1. Introduction

In a previous Valuing Nature Paper 'Demystifying Health'[1], it was argued that health protection and promotion are among the most important roles of the state, and that natural capital assets are a key determinant of these outcomes. Different approaches to how 'health' is conceptualised were presented (e.g. the bio-medical vs. socio-ecological models), and different domains of health introduced (including physiological, psychological and social processes). As one of the state's largest fiscal commitments, the economic costs of health-care were discussed in relation to primary, secondary and tertiary care provision, and the issue of health-inequalities highlighted. The paper concluded by identifying some of the key ways in which health and wellbeing are measured, i.e. 'health metrics', but stopped short of examining these in any depth. **The aim of the current paper is to begin this process of unpacking quantitative health metrics for the natural capital community.**

[1] https://valuing-nature.net/sites/default/files/documents/Reports/VNN-DemystifyingHealth-Web.pdf

Who is the target audience?

The current paper is written, in particular, for that part of natural capital community most involved in commissioning, conducting and interpreting other's investigations into the links between natural environments and health, but who do not have extensive formal training in research approaches in this area. Third-sector environmental organisations, national park managers, local authority teams and some central government agency staff (e.g. Natural England) are all examples of our target audience. Although the issues discussed are global, to keep the report targeted and manageable, we focus primarily on the UK setting. For the sake of brevity, we use the term 'health' to include both 'health and wellbeing' and refer the reader back to the earlier report for a more extensive discussion on these concepts.

What does the report contain?

The report is structured around **five key issues:**

- Rapid review of relevant quantitative metrics;
- Motivations for measuring health;
- Identifying/developing a theory of change to inform metric choices;
- Factors to consider when choosing quantitative health metrics; and
- Factors to consider when collecting, analysing and communicating quantitative health metrics/outcomes.

The report provides a glossary of key terms at the end and suggestions for next steps the community could take to improve the integration of health metrics in natural capital research and evaluation. The report does not focus on qualitative approaches to understanding health but does discuss them briefly.

Supplementary resources[2] include:

- A list of key health metric reviews across the natural capital literature; and
- A taxonomy of commonly used health metrics, both within and beyond the natural capital literature.

How was the report developed?

The report draws on ideas and feedback from participants at two expert stakeholder workshops (London/Leeds, June 2019) and iterative consultation with the wider VNP community, including responses to earlier drafts of this document (a list of contributors can be found at the end).

[2] https://valuing-nature.net/demystifying-health-metrics-1

BOX 1: **What are metrics?**

Health metrics are measures of health determinants, states, or outcomes. They may relate to general health status, (healthy-) life expectancy, disease (communicable or non-communicable), fitness, function and/or capacity (including mental/cognitive capacity and physical disability), injury, or death. They can cover both acute (short-term) states such as negative mood or temporary back pain, and chronic (long-term or recurring) conditions such as depression or chronic back pain. Health-related metrics usually refer not to the health states themselves but to health determinants or risk factors, such as diet, physical activity, smoking, environmental pollution or unsafe work environments.

Health metrics typically relate to either incidence, the rate of new cases of the health outcome, or prevalence, the proportion of 'cases' in the population during a specific period of time (period prevalence) or on a given date (point prevalence). Health metrics can be used at an individual, community or population level.

Health metrics are used for many different purposes including:

- Monitoring population health and inequalities in health outcomes
- Tracking extent or progression of disease
- Assessing the efficacy of health interventions
- Valuing different health promotion or care options
- Targeting health investment and activity

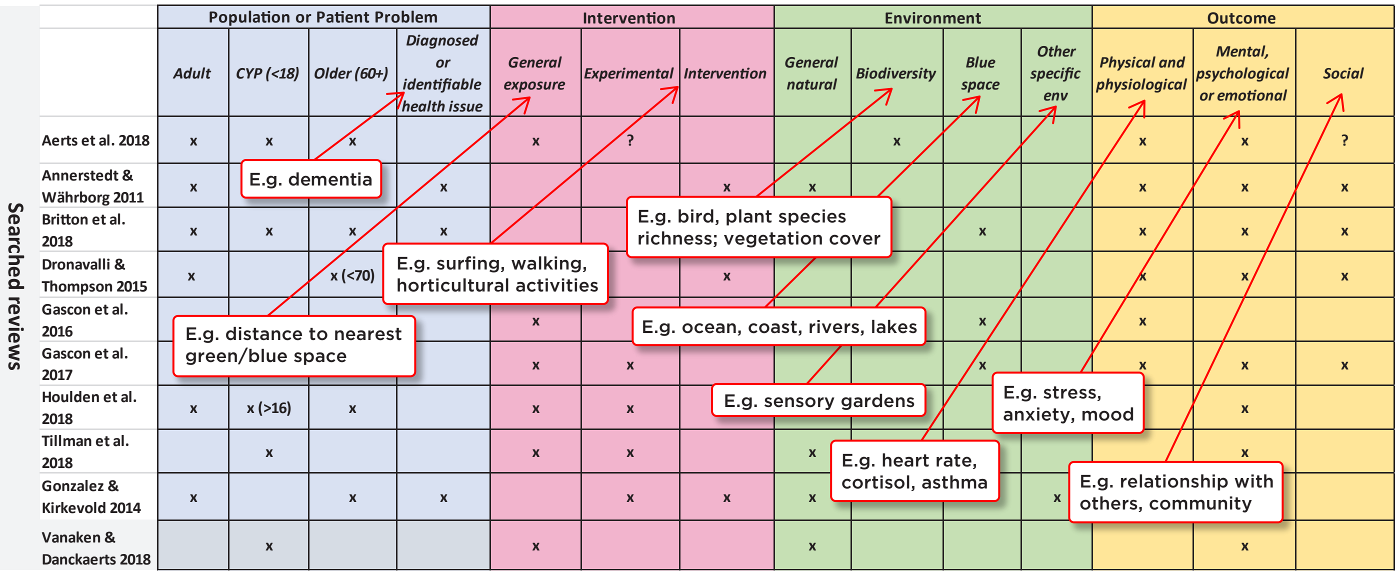
Some health metrics enable comparisons across different health states (e.g. Disability Adjusted Life Years [DALYs] and Quality Adjusted Life Years [QALYs]) and, with caution, can be translated into economic values to support decision making. Individual health metrics can also be brought together to create tools such as the Global Burden of Disease which is used to *'quantify health loss from hundreds of diseases, injuries, and risk factors, so that health systems can be improved and disparities can be eliminated'* (**see Section 6.2**)

2. A rapid review of relevant health metrics

There are thousands of different health metrics and it is not possible to review them all here or to guide nature-health researchers about the most appropriate measures for their specific study without a significant programme of work. Instead, we conducted a rapid review of the key nature-health reviews in the academic literature, and consulted with experts in the field from across disciplines and sectors. This process was used to identify key metrics that have been applied to explore nature-health relationships, and those that were thought to have potential but have not yet been widely implemented.

To prioritise the reviews to search for metrics we used a structured tabulation (using the PICO/PECO format, a framework commonly used to structure systematic review question) to identify key sources (**Figure 1**).

Figure 1: A snapshot of the structured list of key nature-health reviews available at **https://valuing-nature.net/demystifying-health-metrics-1**



Searched reviews	Population or Patient Problem				Intervention			Environment				Outcome		
	Adult	*CYP (<18)*	*Older (60+)*	*Diagnosed or identifiable health issue*	*General exposure*	*Experimental*	*Intervention*	*General natural*	*Biodiversity*	*Blue space*	*Other specific env*	*Physical and physiological*	*Mental, psychological or emotional*	*Social*
Aerts et al. 2018	x	x	x		x	?			x			x	x	?
Annerstedt & Währborg 2011	x			x			x	x				x	x	x
Britton et al. 2018	x	x	x	x						x		x	x	x
Dronavalli & Thompson 2015	x		x (<70)				x					x	x	x
Gascon et al. 2016					x					x		x		
Gascon et al. 2017					x	x				x		x	x	x
Houlden et al. 2018	x	x (>16)	x		x	x							x	
Tillman et al. 2018		x			x	x		x					x	
Gonzalez & Kirkevold 2014	x		x	x		x	x	x			x			
Vanaken & Danckaerts 2018		x			x			x					x	

E.g. dementia
E.g. distance to nearest green/blue space
E.g. surfing, walking, horticultural activities
E.g. bird, plant species richness; vegetation cover
E.g. ocean, coast, rivers, lakes
E.g. sensory gardens
E.g. heart rate, cortisol, asthma
E.g. stress, anxiety, mood
E.g. relationship with others, community

We produced a tabulation of health metrics extracted from the reviews and the community consultation process (as per **Figure 2**). Due to time constraints, the reviews were prioritised (as above) and not all metrics were extracted from all reviews. Summary details of over 270 metrics that have been used in the nature-health field were tabulated, and are presented at https://valuing-nature.net/demystifying-health-metrics-1. The metrics are highly diverse, from the 12 item General Health Questionnaire (GHQ-12), to Head Circumference at Birth, to asthma hospital admission rates. The list is by no means exhaustive but does cover many of the most popular metrics to-date.

To support the Valuing Nature community, the two key resources produced from this process can be found on-line[3].

Broadly speaking there are 5 types of health metrics covered in the list:

- **Routine data** e.g.
  - Hospitalisation incidence
  - Notifiable disease incidence
- **Objective direct** e.g.
  - Lung function tests
  - Accelerometry
- **Objective health-related pathway** e.g.
  - Exposure to PM 2.5
- **Self-report/subjective direct** e.g.
  - Health status scales
  - Quality of life scales
- **Self-report/subjective health-related pathway** e.g.
  - Physical activity participation
  - Smoking

3. Motivations for measuring health

People use, analyse or interpret and apply health metrics for a variety of reasons and in a variety of different ways. Researchers are a key group and use metrics in a wide range of study designs, from primary population surveys, to large scale secondary data analyses, to experimental studies. In applied contexts of policy and practice, metrics are often used for applications such as Environmental and Health Impact Analyses, and to inform local/national health strategies.

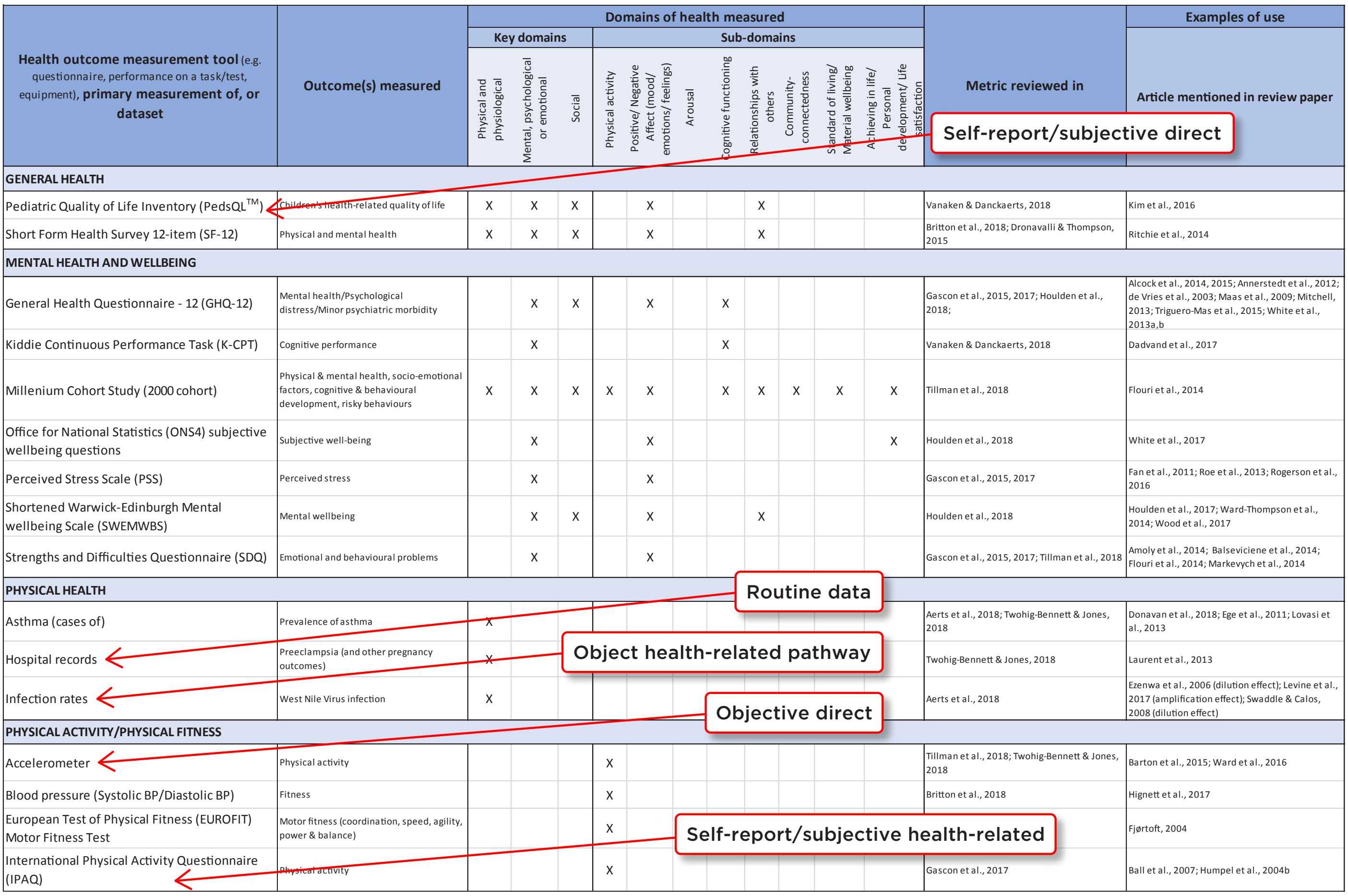
[3] https://valuing-nature.net/demystifying-health-metrics-1

However, health metrics are most often used for two core reasons, monitoring and evaluation, which have unique and overlapping motivations.

3.1 Monitoring

Monitoring, in the health domain, refers to the longitudinal use of a health metric with the aim of understanding both the state of an issue at a particular point in time as well as temporal trends. For instance, monitoring health-related outcomes associated with natural capital assets might include recording the number of tick-borne disease cases recorded over a 10-year period (a health risk), or the number of people regularly visiting urban parks for physical activity (a health benefit). Metrics collected for these purposes are often referred to as *'indicators'*.

Figure 2: Snapshot of the structured list of key health metrics used in health and nature-health studies available at https://valuing-nature.net/demystifying-health-metrics-1



Health outcome measurement tool (e.g. questionnaire, performance on a task/test, equipment), primary measurement of, or dataset	Outcome(s) measured	Domains of health measured											Metric reviewed in	Examples of use
		Key domains			Sub-domains									
		Physical and physiological	Mental, psychological or emotional	Social	Physical activity	Positive/ Negative Affect (mood/ emotions/ feelings)	Arousal	Cognitive functioning	Relationships with others	Community-connectedness	Standard of living/ Material wellbeing	Achieving in life/ Personal development/ Life satisfaction		Article mentioned in review paper
GENERAL HEALTH														
Pediatric Quality of Life Inventory (PedsQL™)	Children's health-related quality of life	X	X	X		X			X				Vanaken & Danckaerts, 2018	Kim et al., 2016
Short Form Health Survey 12-item (SF-12)	Physical and mental health	X	X	X		X			X				Britton et al., 2018; Dronavalli & Thompson, 2015	Ritchie et al., 2014
MENTAL HEALTH AND WELLBEING														
General Health Questionnaire - 12 (GHQ-12)	Mental health/Psychological distress/Minor psychiatric morbidity		X	X		X		X					Gascon et al., 2015, 2017; Houlden et al., 2018;	Alcock et al., 2014, 2015; Annerstedt et al., 2012; de Vries et al., 2003; Maas et al., 2009; Mitchell, 2013; Triguero-Mas et al., 2015; White et al., 2013a,b
Kiddie Continuous Performance Task (K-CPT)	Cognitive performance		X					X					Vanaken & Danckaerts, 2018	Dadvand et al., 2017
Millenium Cohort Study (2000 cohort)	Physical & mental health, socio-emotional factors, cognitive & behavioural development, risky behaviours	X	X	X	X	X		X	X	X	X	X	Tillman et al., 2018	Flouri et al., 2014
Office for National Statistics (ONS4) subjective wellbeing questions	Subjective well-being		X			X						X	Houlden et al., 2018	White et al., 2017
Perceived Stress Scale (PSS)	Perceived stress		X			X							Gascon et al., 2015, 2017	Fan et al., 2011; Roe et al., 2013; Rogerson et al., 2016
Shortened Warwick-Edinburgh Mental wellbeing Scale (SWEMWBS)	Mental wellbeing		X	X		X			X				Houlden et al., 2018	Houlden et al., 2017; Ward-Thompson et al., 2014; Wood et al., 2017
Strengths and Difficulties Questionnaire (SDQ)	Emotional and behavioural problems		X			X							Gascon et al., 2015, 2017; Tillman et al., 2018	Amoly et al., 2014; Balseviciene et al., 2014; Flouri et al., 2014; Markevych et al., 2014
PHYSICAL HEALTH														
Asthma (cases of)	Prevalence of asthma	X											Aerts et al., 2018; Twohig-Bennett & Jones, 2018	Donavan et al., 2018; Ege et al., 2011; Lovasi et al., 2013
Hospital records	Preeclampsia (and other pregnancy outcomes)	X											Twohig-Bennett & Jones, 2018	Laurent et al., 2013
Infection rates	West Nile Virus infection	X											Aerts et al., 2018	Ezenwa et al., 2006 (dilution effect); Levine et al., 2017 (amplification effect); Swaddle & Calos, 2008 (dilution effect)
PHYSICAL ACTIVITY/PHYSICAL FITNESS														
Accelerometer	Physical activity				X								Tillman et al., 2018; Twohig-Bennett & Jones, 2018	Barton et al., 2015; Ward et al., 2016
Blood pressure (Systolic BP/Diastolic BP)	Fitness				X								Britton et al., 2018	Hignett et al., 2017
European Test of Physical Fitness (EUROFIT) Motor Fitness Test	Motor fitness (coordination, speed, agility, power & balance)				X									Fjørtoft, 2004
International Physical Activity Questionnaire (IPAQ)	Physical activity				X								Gascon et al., 2017	Ball et al., 2007; Humpel et al., 2004b

Self-report/subjective direct
Routine data
Object health-related pathway
Objective direct
Self-report/subjective health-related

Monitoring may be directly related to specific targets, *'key performance indicators'* (KPIs), based on informed suggestions ('guidelines') or legal obligations and statutory requirements. For instance, the World Health Organisation (WHO) recommends monitoring the percentage of a given population who have access to outdoor recreational green space within 300 meters of their home, as it believes this is a key determinant of health. However, it has no jurisdiction for setting this as a regulatory target. By contrast, the Marine and Coastal Access Act (2009), places a legal responsibility on Natural England and the Secretary of State for the Environment to improve public access to the English coast by developing an accessible coastal footpath and recreational 'margin', with various clauses stating that these developments still need to protect both the environment and human health.

Where regulations or clear guidelines are lacking, monitoring is closely linked to *benchmarking*. For instance, Public Health England's, Public Health Outcome Framework (PHOF) has a large range of health-related indicators that rely upon regular monitoring at the Local Authority level. Many of these indicators may be indirectly linked to natural capital; an example was indicator 1.16 *'Utilisation of outdoor space for exercise/health reasons'*. The interactive PHOF website[4] uses a traffic light system to highlight trends (green = improvements over time, amber = no change, red = worse situation) and benchmarks by geographical region with particularly high or low scores highlighted. In situations where it is hard to know a priori what a 'good' or 'bad' state of affairs looks like (e.g. is it a good or bad thing if 42% of people in Leicester use outdoor spaces for health purposes?), benchmarking at least provides temporal and geographical comparators. Other useful online resources for access to health metric profiles at different spatial scales include the Local Authority Health Profiles[5] and the Strategic Health Asset Planning and Evaluation (SHAPE) Tool[6].

3.2 Evaluation

Although closely related to monitoring, and sometimes using the same metrics, health evaluation usually refers to an assessment of the success (or otherwise) of an action or intervention aimed at changing one or more of the determinants of a health outcome by comparing health outcomes before and after an intervention. For instance, one might evaluate the effectiveness of a new anti-cancer drug, triaging procedures in an emergency department, or improvement of walking/cycling access to an urban park. Good evaluation is not simply about concluding whether or not an intervention has 'worked', it is also about understanding the pathways between intervention and outcomes and identifying both barriers as well as routes to success. Moreover, success of an intervention is considered in terms of: a) *efficacy* (usually compared to the next best alternative), b) *feasibility* of replicating the intervention at scale or in different settings, and c) *cost-effectiveness* (does the health impact warrant the investment?).

[4] https://fingertips.phe.org.uk/profile/public-health-outcomes-framework

[5] https://fingertips.phe.org.uk/profile/health-profiles

[6] https://shapeatlas.net/

In health evaluation contexts the health metrics selected will often be very specific with clearly defined pathways, e.g. a diabetes management intervention may use the metric of blood glucose levels as the acute outcome because it has well understood pathways to vascular disease.

4. Identifying the most appropriate metric(s) for your study/evaluation

Above, we noted that health-related metrics may be used for monitoring purposes in relation to a legal framework, international guidance or benchmarking activities. We also noted that some funders require/recommend the collection of certain metrics as part of their overall evaluation objectives. In both cases there is often little, or no, choice in the metrics to be used. In other situations, precise health metrics to be explored may not be outlined or there may be opportunities to explore other metrics alongside those required/recommended. The aim of this section is to support members of the Valuing Nature community interested in using health metrics, but unsure of which ones to focus on, in their decision-making processes. At this stage we are unable to make firm recommendations about precisely which metrics are 'best' or to identify a 'gold standard' for certain applications. Rather, we introduce some of the factors health researchers consider in the selection of metrics for their purposes, many of which have direct parallels for metrics used in the natural sciences.

In our consultation with the Valuing Nature community, many individuals stated that they were often more involved in evaluation than monitoring and that a key driver was a requirement of funding bodies to demonstrate the 'success' of their interventions (e.g. biodiversity enrichment of a local park, a forest school for pre-schoolers, or greening of a health-care facility). Moreover, some funders offered guidance on the metrics that might be used, with Heritage Lottery Funded programmes such as EcoMinds, recommending the use of standard metrics such as the Warwick-Edinburgh Mental Wellbeing Scale (WEMWBS). Attempts by funders to encourage applicants to use the same health and wellbeing metrics is sensible because it enables more direct comparison between different interventions than if different interventions used different health metrics.

A very useful document produced by Health Scotland (2007), outlines the steps to consider when choosing mental wellbeing metrics for research and evaluating purposes: 'Mental Health Improvement: Evidence and Practice Guide 5: Selecting scales to assess mental wellbeing in adults. Evaluation guides'[7].

[7] http://www.healthscotland.scot/media/2245/5-selecting-scales-to-assess-mental-wellbeing-in-adults-mental-health-improvement-evidence-and-practice.pdf

4.1 Step 1: **Developing conceptual and theory of change models**

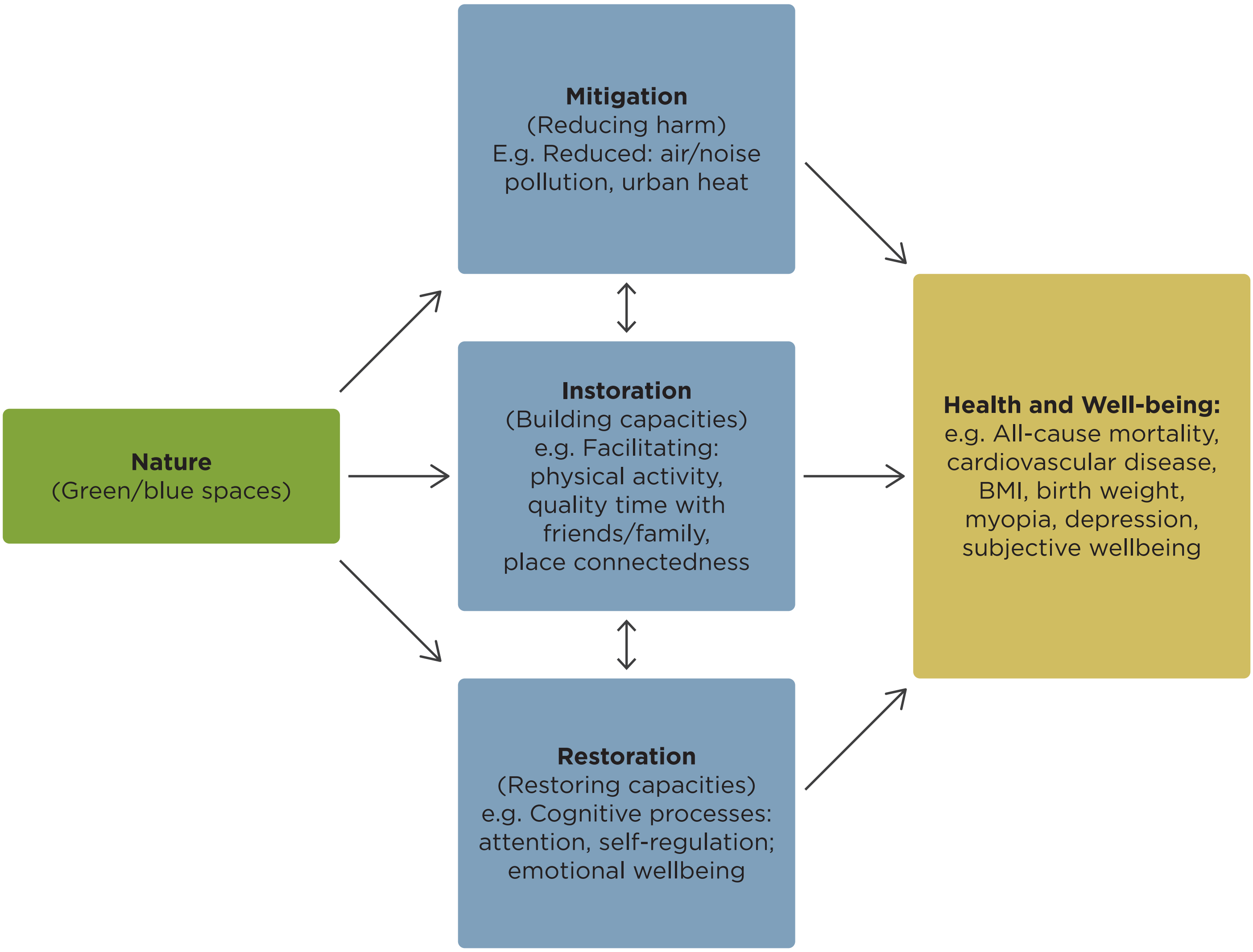
Identification of a suitable health metric relies on a clear understanding of what it is that one is trying to measure. In epidemiology and related sciences conceptual models are used to illustrate pathways between exposure and health outcomes; in intervention research *programme theory* and theory of change (ToC) models are developed to visualise how the action may bring about change.

A well-known conceptual model in the nature-health field was developed by Markevych et al. (2017; **Figure 3**). This model postulates that the links between the natural environment and health are mediated through mitigation processes (e.g. reduced air pollution), instoration processes (e.g. increased physical activity), and restoration processes (e.g. reduced stress). Although still relatively simple the model nonetheless begins to map out the kind of processes that could be measured in addition to the end health outcomes, in order to account for any effects (e.g. did the intervention work by reducing stress, or increasing physical activity?).

It is ***highly recommended*** that any nature-health research or evaluation project is guided by its own specific conceptual or ToC model. In the case of interventions, thinking about and articulating the programme theory is useful because it can help identify how the activity or programme is intended to work, what processes and resources may be needed, the factors which might prevent success, and the timeframe of impacts and outcomes. When used to inform evaluations, programme theory and ToC models are useful in helping clarify the most appropriate metrics and when they should be used. Although they can be developed at any stage of the intervention design, implementation or evaluation, it is perhaps most useful to develop them at the very earliest stages. The Better Evaluation initiative has useful guidance on developing programme theory and ToCs[8].

[8] https://www.betterevaluation.org/en/rainbow_framework/define/develop_programme_theory

Figure 3: A simple conceptual model linking nature & health (Adapted from Markevych *et al.*, 2017)


Mitigation
(Reducing harm)
E.g. Reduced: air/noise pollution, urban heat
Nature
(Green/blue spaces)
Instoration
(Building capacities)
e.g. Facilitating: physical activity, quality time with friends/family, place connectedness
Restoration
(Restoring capacities)
e.g. Cognitive processes: attention, self-regulation; emotional wellbeing
Health and Well-being:
e.g. All-cause mortality, cardiovascular disease, BMI, birth weight, myopia, depression, subjective wellbeing

It is strongly recommended that relevant stakeholders are involved even at this early stage as they will often have detailed knowledge of local health needs, assessments, priorities, mechanisms and possibly earlier relevant reports or studies. This engagement process would ideally be maintained throughout the whole process and stakeholders involved in dissemination activities.

In some cases, developing a ToC may reveal that the processes are so complex and difficult to measure that it may just not be feasible to try and conduct a formal evaluation within a given time frame or with limited material resources. Guidance on deciding whether or not it even makes sense to be trying to measure health metrics at all can found in the literature on 'evaluability assessment' (e.g. Ogilvie et al., 2011), which recognises the challenges of real-world research and argues against conducting an evaluation for its own sake if there is little hope of 'success' in terms of it revealing anything of interest.

4.2 Step 2: Study design

Once a conceptual model has been developed, intervention studies need to develop a clear design for the research in which health metrics are to be used. A classic challenge for nature-health interventions is finding a ***suitable control group,*** i.e. a group of people who did not receive the intervention but who are, in all other respects, the same as those who did. People who sign up to a specific intervention (e.g. nature volunteering or a forest school) will have certain characteristics and identifying an identical group not involved, but willing to have their health metrics collected over multiple time periods, is likely to be difficult. For this reason, health-researchers often employ 'waiting control groups' i.e. people who have also signed up but who will only receive the intervention at a later date, because they are likely to be more similar to the main 'experimental' group, and more willing to be measured over time, in anticipation of receiving the intervention at a later date.

Identifying suitable control sites for physical interventions (e.g. biodiversity improvements in a particular habitat) may be even harder, because by definition no two locations can be identical in all respects. An excellent example of an attempt to identify control sites for a series of physical improvements to woodlands near urban areas was conducted as part of the Woods In and Around Towns (WIAT) project[9]. Control groups are so important because even if the health metrics selected do show an improvement in an intervention group over time, we cannot be sure this was down to the intervention, as opposed to some outside influence (such as the weather or political situation etc.) unless we have also measured similar people at similar times not involved in the intervention. In cases where being able to include a control is not possible, e.g. due to cost constraints, an approximation of a control group might still be achievable by referring back to the ToC. For instance, if it is believed that an environmental improvement influences health via spending time in it, then one could at least compare the health outcomes of people who visited the site more or less often pre- and post- development, while trying to make sure everything else about these people was as similar as possible.

In an ideal world, a study might have multiple control groups to tease apart the mechanisms and pathways elucidated in the ToC model. For instance, a nature-based volunteering effort may have three groups: a 'waiting control group', an 'activity only group' and an 'activity plus engaged leader' group. It is possible that a large part of why some nature-based interventions 'work' is because they are led by enthusiastic, knowledgeable, and engaged individuals or groups and that spending time with such people is good for health and wellbeing over and above the actual (natural) setting in which the activity occurred. Evidence of health benefits of singing groups, knitting groups and other non-nature 'social interventions' supports this possibility. Only by disentangling the group/leader effects from the physical environment setting can we be sure it is nature in and of itself that is key.

[9] http://www.openspace.eca.ed.ac.uk/research-projects/woods-in-and-around-towns/

In nature-health research that uses pre-collected/secondary health metrics, examining issues of causality is likely to be especially hard and 'associations' and 'relationships' are the best we might expect (e.g. people who live nearer the coast tend to report better mental health). Analysts try to reduce the problem of reverse causality (e.g. people with better mental health tend to move to the coast) by statistically controlling for a range of other factors known to influence both the exposure variable (e.g. coastal proximity) and the health metric (e.g. mental health). In some secondary data cases a 'natural experiment' may have occurred where only a sub-set of people for whom health data was already being collected experienced a change (e.g. a new park in their neighbourhood). **See Section 6.1**.

4.3 Step 3: Considering effect size, sample size and statistical power

Assuming a ToC has been developed, and an evaluability assessment suggests it may well be worthwhile to explore health metrics, the precise nature of the relationships and metrics needs to be considered. Most health outcomes are influenced by multiple factors (e.g. genetics, socio-economic circumstances, lifestyle behaviours, and psychological mechanisms) and identifying the potential impact of, for example, a 12-week nature-volunteering programme may be difficult compared to these other influences. That is, detecting nature's 'signal' against a background of 'noise' (the role of other factors), can be hard because the *'effect size'* of nature on health outcomes is generally small, compared to determinants such as poverty or smoking. If a 'signal' is to be picked up, researchers need to use a health metric sensitive enough to detect it, and to collect health metric data from enough people. Generally speaking, the more people that are sampled (the *'sample size'*), the more likely a researcher is to find a 'signal' if one exists. Using health metrics that are unable to detect differences with small samples was something many in the VNP community identified as a possible reason why so many studies report null results, i.e. fail to find evidence that their intervention 'worked'.

To address this problem, researchers in health sciences estimate how many people they will need in their study to detect an effect of the intervention (if one exists) *before* they start the study. The exact health metric selected will be key in determining the required sample size, and statistical *'power analysis'* is used to help estimate the number of people to include based on information from previous studies that used the same (or a very similar) metric. If such an analysis suggests a study would need far more people than are readily available in order to find an effect, there is little point in using that metric, since there is a very high probability a researcher will find no effect. To take a simple/obvious example, although there are now several studies showing reduced mortality rates among people who live in greener areas, there is little point in using reduced mortality rates as a health metric for *evaluating* the success of a local park volunteer scheme – since we are unlikely to see any difference in mortality rates for such a small sample. Rather, a metric that previous studies has demonstrated can detect differences in something less extreme than mortality among relatively small samples, e.g. a metric of psychological wellbeing, would be far more useful.

On the other hand, researchers also need to be aware that using very large sample sizes can detect even very small effects, which may be relatively unimportant in the bigger scheme of things. This issue is more likely to occur with respect to monitoring than evaluation where hundreds of thousands or even millions of people are being sampled (e.g. Census data). To try and reduce the risk of overplaying negligible effects, epidemiologists use various statistical techniques to ensure findings are robust, i.e. are not merely due to confounding or the particular statistical models being used. Nevertheless, due to the ability of these approaches to detect even small effects, it is important for nature-based researchers not to exaggerate the potential impact of nature on health (**see section 6.2 below**).

In sum, before selecting a health-metric for an evaluation project nature-health researchers need to make sure they have enough people to find an effect, if one exists, and if analysing large datasets collected for monitoring purposes, be mindful that the effects detected may be very small and their meaning should not be over-played.

4.4 Step 4: Understanding metric qualities and making trade-offs

Objective data is based on independent observations (e.g. by a measurement technology or health professional), whereas *subjective*, or self-report, data is based on personal perception, opinion or experience. For example, an objective measure of physical activity in the last seven days may be the number of steps registered on an individual's pedometer, whereas a subjective measure would be the self-reported physical activity levels in the last week. Note that an objective metric's quality is only as good as the techniques for measuring it. A pedometer might pick up certain types of sedentary activities and fail to pick up cycling and swimming, and thus a so-called objective measure of physical activity also needs to be treated with caution and its limitations understood. Objective metrics are not always 'better' than subjective measures, and some outcomes are most appropriately (or only) measurable with subjective metrics. Occasionally, self-report measures, such as weekly physical activity, are misinterpreted as being *'qualitative'* health metrics, by virtue of the fact that they are self-reported. This is due to a misunderstanding of the difference between quantitative and qualitative approaches to research (**see Box 2**).

BOX 2: **Quantitative & Qualitative data**

Quantitative approaches are based on a positivistic paradigm, that there is an objective external reality that can be documented using metrics that can be converted into numerical values, and thus submitted to statistical analyses for summary and interpretation. Examples relevant to the nature-healthy field include physiological metrics (e.g. blood pressure, cortisol), weight, scores on standardised attention tests, and self-reported wellbeing measured through survey items. Self-report data is still classed as quantitative if it can be summarised numerically.

Qualitative approaches are typically underpinned by a different ontology (understanding of the nature of reality) and epistemology (way in which we go about trying to understand that reality) to quantitative approaches. Often researchers using qualitative approaches are interpretivist and seek to understand the subjective, socially constructed nature of reality. They aim to *'study things in their natural settings, attempting to make sense of, or to interpret, phenomena in terms of the meanings people bring to them'*. There are a number of different approaches including phenomenological, ethnographic and historical inquiries, and methods including observation and immersion, interviews, open-ended surveys, focus groups and content analysis (**https://esrc.ukri.org/about-us/what-is-social-science/qualitative-research/**). Qualitative research does not depend on large samples, instead some seek to reach 'saturation'. Claims of generalisation can be made from qualitative findings, for instance in relation to transferability, making statements about similar groups or situations.

Due to their different strengths, mixed methods approaches, a combination of both quantitative and qualitative elements, are often used, either at different stages of a study (e.g. interviews/focus groups during the development of a survey) or simultaneously (open-ended questions in a survey).

Metric qualities

When considering which health metric(s) to use, or how to interpret previous findings using specific metrics, there are many factors to consider (**see Box 3**). The level of importance attached to these factors may depend on the type of study/intervention used, the outcomes under consideration, and any time, researcher experience, or budgetary constraints. In practice, researchers will often have to make a series of trade-offs in their choice of metrics.

The *validity* of the health metric is key. A measurement tool's validity is the degree to which the tool (e.g. questionnaire, procedure or assessment) effectively measures what it is supposed to (see Dronavallli & Thompson, 2015, for a list for definitions of different validity types). It is therefore, wherever possible, advisable to use a validated measurement tool.

The metric should also be *reliable*, i.e. it should consistently give the same value if the thing being measured has not changed. Most standard measurement tools should have been tested for reliability as well as validity. It is important to also consider how *applicable* the metric is to the target audience. For instance, the tool should be written in the appropriate language and checked for any cross-cultural factors that may deem the metric less appropriate. It is extremely useful if the measurement tool has *global relevance* (in terms of being comprehensive rather than international). Tools that incorporate a global measure of health or wellbeing, such as "How is your health in general?" enable the metric's values to sit within an overarching context.

How appropriate a particular measurement tool is for your *timeframe* (acute/chronic) should also be considered. Metrics for longitudinal interventions may be very different to those that are appropriate for a cross-sectional study. Ideally, the conceptual or ToC model should consider not just the processes involved, but also the timescales over which one would expect any detectible changes to occur. Changes in many health states may take many months (e.g. Body Mass Index) or years (e.g. diabetes status). Changes to intermediate outcomes indicative of improvement in health outcomes in the future may occur relatively quickly, even within the timeframe of a short intervention (e.g. an increase in regular physical activity). For this reason, it may make more sense for some evaluations to focus on metrics of intermediate outcomes (such as physical activity), rather than health outcome metrics *per se* (such as BMI).

Box 3: **Factors to consider when choosing a health metric**

Appropriateness

- Validity
- Reliability
- Applicability
- Responsive and sensitive to change (if applicable)
- Global relevance
- Timeframe
- Cross-cultural validity
- Ability to compare with previous and future research

Practicalities/feasibility

- Accessibility & Cost
- Acceptability
- Linguistically appropriate (age appropriate/English as a second language)
- Ease of use
- Clarity of tool
- Time requirements
- Interpretation of results & analytical capacity
- Response, and other, biases

There are also several practical issues that need to be considered in choosing a health metric [**Box 3**]. The tool should be accessible and of an appropriate cost. Many measurement tools are available on-line and are free of charge for research purposes. However, some metrics (including some commonly used in nature-health studies) involve an application process for permission and/or require a fee to be paid (sometimes per 'use' or per survey response). Hidden costs should also be considered, for instance, are there additional processing and/or analysis costs for diagnostic tools?

Ideally, a health metric should be easy to *use* – both for the person administering the measurement tool, as well as the person completing it, including considerations of the language abilities of the respondents and whether tools might need to be available in multiple languages among certain communities. Consider whether a certain level of expertise or training is required to administer or complete the tool, which is related also to a tool's *clarity*. Ideally, in order to collect good quality data, the tool should be understandable for a non-specialist and it should be free of ambiguities. A tool that is too complex, either for the participant to understand or the researcher to administer, may result in poor quality data.

It is important to consider the amount of *time* available to collect data. Some measurement tools may consist of a small number of items (or even just one) or involve quick and simple tasks. In contrast, other tools may require the completion of a long questionnaire, or involve lengthy tasks or complex procedures. Although more detailed questionnaires may provide a greater depth of understanding, they can be very time consuming; there is a trade-off with how much data is 'enough'. It is important that the data can be accurately *interpreted*. Some metrics may be relatively easy to interpret (e.g. simple scores on a scale), whereas others may require a specific level of statistical expertise or specialist software.

When conducting a survey or structured interview, it is important to be aware of the potential for *response (or survey) bias*. Response bias is a general term that refers to the various conditions and factors (often unintentional) that can influence participants' responses. Response bias can result in participants providing false or misleading answers: for instance, they may feel they need to give answers that are socially acceptable/desirable or answers that they think will 'help' the researcher with their study (demand characteristics). Response bias can ultimately affect the accuracy and validity of the data, and may be a particular issue with some metrics (for example self-reported alcohol intake or physical activity).

5. Data collection

Once the most appropriate health-metrics have been identified for any given study a researcher needs to be mindful of best-practice data-collection. The best metric available is only as robust as its implementation.

5.1 Who collected or is collecting the data?

In some cases, health data will have already been collected and made available to users, this is called *secondary* or *routine* data. Often this type of data is originally collected for monitoring or research purposes but has application for other uses. Examples of this type of data are government health surveys, or annual hospital admissions from tick bite infections. In other cases, such as intervention evaluations, data will be collected specifically for the purpose of understanding if and how the intervention works. This is called *primary* data. A key concern is ensuring that the data is robust and that potential sources of bias are understood and minimised. Sources of bias differ between study types and data collection methods and some sources of bias relate to who collects the data.

Typically, considerable effort is put into selecting metrics and ensuring data collection for national and local monitoring exercises are as robust as possible. For instance, Natural England's 'Monitor of Engagement with the Natural Environment (MENE) survey uses highly sophisticated sampling techniques and trained interviewers to conduct in-home interviews with very large samples (approx. 40,000 per year) to best capture the nation's interactions with the natural world. Further, it employs an independent third-party organisation to collect and analyse the data, to reduce the potential of inadvertent 'confirmation bias' (the desire to see one's expectations supported). Similar processes occur for much official health metric data collection. The Office for National Statistics (ONS), for instance, requires certain standards of data collection to be met before data can be awarded 'National Statistics' status.

'Official' health metric data should still be treated with caution. For example, rates of hospitalisation are affected by the availability and accessibility of appropriate services for the population that needs them, and also on correct diagnoses being made and recorded. In most intervention settings there is neither the time nor resources to conduct such thorough data collection protocols. Nevertheless, for even for the smallest of projects, if funds permit, third-party data collection to reduce the potential for confirmation bias is recommended.

5.2 Ethics and data protection

For primary data collection, where people are sampled and their data stored for analysis, many studies will require some form of ethical approval from a recognised ethics board. Ethics boards are sensitive to issues such as informed consent, ensuring study protocols maintain participants' dignity, and highlighting participants' right to withdraw at any point. They can also help provide guidance on the new data protection regulations (GDPR)[10] regarding how data should be collected and stored. Many boards will require clear evidence that the evaluation may provide new information of importance, e.g. through detailed power analyses, to reassure them that people's time is not being 'wasted' on an evaluation that does not have a sufficient sample size to detect an effect.

[10] https://eugdpr.org/

Ethical approval is usually required before data collection can begin although piloting is sometimes allowed if it is used to inform the study design and the data is not used for later analytical purposes. Ethics applications often go through several iterations and can take several months (especially if the NHS is involved), so researchers new to collecting human data need to be mindful of this when developing their research timeframe. Ethical approval may not be needed for service evaluation, although the GDPR on data protection rules will still apply. Useful guidance from the Medical Research Council on which approaches to ethics and data protection you will need for your work can be found here.[11]

5.3 When is the data being collected?

Choosing the right time at which to collect data on health outcomes is critical for utility. For example, it's not advisable to measure either resting heart rate or heart rate variability (HRV), even using the most sophisticated equipment available, shortly after strenuous exercise or a large cup of coffee, because these can significantly affect the readings. But timing issues can be subtle [**Box 4**].

Box 4: **Exemplar on the importance of timing**

One member of the VNP community told us about a multi-session nature-based intervention evaluation, along the following lines, that measured momentary mood (e.g. happiness now) *before any intervention* (T1) had occurred and directly *after the final session* (ten weeks later, T10). As he pointed out, demonstrating an increase in momentary mood from T1 to T10 under this situation does not really help us understand whether the 10-week intervention improved mood, because it was confounding engaging with a single session with taking part in all 10 sessions. To tease this out, an alternative would have been to measure mood *immediately after* the first session and compare that with the post-T10 measurement. That way both measurements occurred directly after an intervention session, the only difference was whether there was a longitudinal improvement over the 10 weeks.

[11] http://www.hra-decisiontools.org.uk/research/

Other issues to consider with regards to timing are participant burden, fatigue and drop outs. Collecting data at multiple time points (i.e. longitudinally) is often recommended in the final paragraph of academic papers, but can be hard to achieve in practice. A major issue is participants' ability and willingness to be measured at multiple time points and many longitudinal studies experience considerable 'attrition', i.e. people dropping out over time. In many cases such 'attrition' is non-random, i.e. certain people are more likely to drop out of a program than others, especially those who believe they are not experiencing benefits. This can result in a biased final sample, with those people who benefit from the intervention more likely to remain in the study; this might lead to an incorrect conclusion that the intervention 'worked' for everyone. Health research has developed techniques to try and mitigate such interpretation errors, e.g. Intention-to-treat (ITT) analysis, and it is recommended that these issues are considered carefully before deciding when and how often participants are expected to be measured.

6. Analysing, interpreting and communicating findings

6.1. Cleaning, analysing and interpreting health metric data

Health metric data can be very messy and require considerable 'cleaning' before they can be used for analytical purposes. Even relatively simple metrics such as blood pressure and heart rate are sensitive to all sorts of extraneous factors and responses, especially in small samples, and the data may violate assumptions needed to run certain types of analyses. This can also be true of self-report data. For example, subjective wellbeing data tends to demonstrate negative skews (e.g. most people score 7-8, rather than 5-6, on 0-10 scales of life satisfaction).

A full discussion of analytical approaches and interpretation of health metric data is beyond the scope of this short report, and depend on the exact study design and the measures being used. However, some general issues should be considered, many of which are not specific to this context, but do have particular pertinence to nature-health research and evaluation:

a. Causality: Establishing causality is hard, if not impossible, in most situations when trying to identify the links between nature and human health. Typical study designs (such as non-randomised interventions and observational studies) do not permit substantive inference of cause and effect.

b. Reverse causality: is often possible (or probable), and it should be considered as a possible explanation for results (for example, people with depression may be less likely to visit nature, which may be why depression rates are lower among regular park visitors, rather than visiting leading to less depression).

c. Generalisability: Often the exposure or intervention being evaluated has unique characteristics or is within a specific population or place; these may limit the extent to which we could be sure that the same exposure or intervention in a different place or population group would have the same effect.

d. Confounding: There are often very powerful determinants of health at play (such as socio-economic status) that can also be associated strongly with exposure to natural environments, or participation in a nature-based intervention. Being aware of, and mitigating, potential confounding factors is important as residual (unmeasured) confounding could explain observed relationships.

The basic advice here is, however simple the health metric appears to be, it is a good idea to ensure that the project includes, or has access to, people who are used to dealing with these metrics. Generally speaking, the more complex the metric, the more expertise, time and money will be needed to clean, analyse and interpret it, and these factors should inform the metric-selection process.

6.2 Communicating the findings

Reporting a change in most health-metrics is likely to mean little to all but a small group of experts. Is a 2-point improvement in EQ5D scores something to shout about or not? Clearly, thinking about the audience will be key and, in the experience of many of those we consulted, three key issues were raised: relative effectiveness, cost-effectiveness, and objectivity *vs.* advocacy.

Relative effectiveness

In many people's experience stakeholders and audiences wanted to know not just that something 'worked', but how well it worked compared to other interventions etc., or how important it was compared to other things that we know are important for people's health and wellbeing. Simply demonstrating 'statistically significant' improvements in a given health metric from a nature-based intervention does not in itself make it important. As noted above, the bigger the sample, the more likely one is to find statistically significant effects and it may be that similar improvements can be achieved more cost-effectively through non-nature based intervention means.

Along similar lines, some researchers are starting to communicate the size of 'nature's effects' alongside those of other factors which also influence health and which society is already familiar with. For instance, we know that richer people tend to be healthier, and so by comparing the health metric scores of the rich and poor we can see whether differences in health across nature *vs.*

non-nature intervention groups are smaller, similar to, or larger than the differences observed as a function of income. Such comparisons help to contextualise the relative importance of nature-based exposures in ways that are more readily understandable to many people. Considerable effort is put into 'translating' health metrics into outcomes that are readily understandable for communication and policy purposes.

Cost-effectiveness

Since money is a unit people can readily understand a range of monetary valuation possibilities have been developed and are widely used to help contextualise health effects such as Social Return on Investment (SROI), saved medical costs, or social welfare values in terms of how much society is willing to pay for a reduction in disease incidence or severity. This latter approach underpins the calculation of Quality-Adjusted Life Years (QALYs) and Disability Adjusted Life Years (DALYs), both of which are starting to appear in nature-health research.

To simplify, QALYs take into account not just how much life expectancy an individual might gain from an intervention, but also the quality of life they can expect during that time. The potential trade-offs between life expectancy and quality of life are presented to members of the public, e.g. "would you rather live one extra year in perfect health or two extra years in constant pain" and a rank order of societal health-state preferences created. By being able to compare across different health states using a single metric, QALYs are useful to health services with limited budgets because they can inform decisions about which interventions may have the largest effects for any given investment.

They are useful for the nature-health field because the organisation which informs decisions about health funding, the National Institute for Health and Care Excellence (NICE), has made statements about how much it is willing to spend to achieve specific QALY gains. For the VNP community this may be important because if a nature-based intervention were able to demonstrate that it had delivered health benefits equivalent to 5 years lived in perfect health (i.e. 5 QALYs), then society should, in theory, also be willing to pay the same amount for delivering 5 QALYs whether that is through a new drug or greater exposure to nature. In practice, exactly how much NICE is willing to pay per QALY (possibilities under discussion range from £13,000-£100,000), and whether or not a QALY achieved through different intervention types for different target groups will be judged equally in monetary terms, is unclear and we believe currently under review. VNP researchers interested in QALYs are recommended to explore the issues in more detail since some health metrics are more amenable to QALY conversion than others (e.g. EQ5D, SF6D), and some health-related metrics such as physical activity can be converted under further assumptions, but this needs to be done with caution.

Although QALYs are used by the UK health system, researchers interested in more global nature-health relationships may want to consider whether they can express any health metric improvements in terms of DALYs, which are more recognised globally[12]. Like QALYs, one DALY is considered to be one year of perfect health, although lost rather than gained. Like QALYs one part of their calculation is based on life expectancy, i.e. Years of Life Lost (YLL) from a disease or illness. Where they primarily differ is on the calculations concerning the incidence (or more recently prevalence) of diseases and the relative weights given to them, as well as how long the disease/illness usually lasts. Again, more details are beyond the scope of this report, but DALYs are potentially useful because they are used to inform reports on the Global Burden of Disease, and thus can be used for cross-national comparisons.

Feasibility, replicability, fidelity and 'payback' time frame

Even if analysis of a pilot intervention suggests it is relatively cost-effective in terms of health and wellbeing, this is not the end of the story. Before commissioners consider investing more widely, they will also want to know how *feasible* a 'successful' pilot would be at scale. The pilot may have relied on circumstances that are hard to replicate in other contexts, e.g. high volunteer engagement, or may depend on a particular set of skills of those involved in delivering the intervention that are not easy to ensure beyond a specific project. For these reasons research using health metrics usually considers the *fidelity* of an intervention, i.e. how easy it is to replicate according the original formula. In general, the simpler the intervention with fewer component parts and less expertise of those delivering it, the easier it will be to replicate and thus the greater fidelity the roll out would have. In other words, cost-effectiveness assessments need to consider not just the circumstances of the trials and pilots but also the implications for a wider roll out which may not be obvious.

Commissioners may also want to know about how soon they can expect meaningful returns on investment. As noted above, nature-based interventions that aim to increase life-expectancy have a much longer time horizon than those who want to show an improvement in momentary mood, neither of which may have much direct relevance to health care providers. Instead they may be looking at disease rates over set periods of time such as annual rates, five-yearly rates etc., and ideally these issues will already be considered at the time specific health metrics are chosen; can you show meaningful returns in a time-frame under which your key audiences operate (e.g. local electoral cycles)? More detailed insights into these processes can be found here.[13]

[12] https://www.who.int/healthinfo/global_burden_disease/metrics_daly/en/

[13] https://www.cipfa.org/policy-and-guidance/reports/evaluating-preventative-investments

Advocacy through health metrics

Finally, many people we consulted were keen to point out that we should be cautious of using health metrics to try and justify, or 'advocate', a specific theory, circumstance or intervention. Instead we should try to be as 'removed' from the findings as possible and merely report the outcomes as they are, rather than trying to put a positive spin on things. Of course, it is perfectly understandable why individuals involved with any given intervention want to stress the potential benefits, but if these are overstated it can damage the reputation of the field as a whole, and potentially valuable lessons learnt about what could have been done differently or better are missed. Researchers should also be mindful that nature poses many risks/threats to health (e.g. accidents/falls, drownings, bites/stings, vector-borne diseases etc.) and ideally will try to balance the benefits with the risks in their communication of an intervention's overall health effects.

7. Common health metric scenarios for valuing nature

In an attempt to put these relatively abstract discussions in context, we developed a number of hypothetical nature-health project 'archetypes' as a way of helping to demystify health metrics in practice. Earlier versions were discussed at both VNP health metric workshops and developed further based on participant feedback. Although hypothetical, they hopefully serve to stimulate those new to using health metrics to consider some key issues.

Archetype 1. **Small-scale environmental intervention: Evaluation of a new greenway**

A local council wants to understand if and how their investments in a new green route linking a number of neighbourhoods with the train station has benefited local health and wellbeing and whether the investment was value for money. The council planted street trees and flower beds, improved the connectivity and quality of local off-road cycle routes, and installed safety infrastructure such as additional lighting. The council are interested in a wide range of benefits so that they can fully understand value for money.

They decide to use an *evaluative* (rather than a forecast) Social Return on Investment (SROI) approach. The council held stakeholder and community consultation events to ensure the plans were suitable for local users. They also worked with a local consultancy to map out how the greenway might bring about positive change and what the most appropriate health metrics to use might be.

They soon realised that they would not be able to demonstrate direct health changes in the time available, so instead decided to focus on potential increases in physical activity as a key determinant of health in the ToC model.

Focusing on increased physical activity allowed them to use the World Health Organisation's Health Economics Assessment Tool (HEAT)[14] to value the additional active transport and physical activity from the new greenway using quantitative data from long running visitor surveys and local transport assessments. The combined datasets have questions on how people travel about in their neighbourhoods, how many trips are taken using different transportation methods, their duration, frequency and distance. The council were able to use local data on mortality rates and accurate population estimates to help ensure the results were relevant to their context.

The HEAT tool produced an estimate of reduced mortality and a monetary estimation, based on the value of a statistical life (VOSL), of the value of the average amount of walking or cycling per person per day on the new greenway. Outcomes of the results for the HEAT analysis were integrated into the wider SROI analysis[15] to indicate whether the new greenway was cost-effective and value for money.

A key challenge for the council remained in interpreting whether any improvements in physical activity were 'due' to the greening aspect of the intervention (e.g. trees, flowers) and they realised they should also gather people's perceptions about the extent to which the 'greenness' of travel infrastructure influenced their willingness to use it.

Archetype 2. **Small-scale behavioural intervention evaluation: Volunteering with an environmental charity**

An environmental charity wishes to understand whether the volunteering programme they run benefits the health of the adult participants. The participants take part in a variety of hands-on environmental management activities such as scrub clearance, litter picking and hide construction. Sessions last two hours and are held once a week over a three-month period.

The evaluators construct a Theory of Change to show how they think the volunteering activities benefit health. They review the evidence and think about both pathways to health and direct health outcomes. The evaluators work through an evaluability assessment and decide that existing evidence on environmental volunteering suggests that some health outcomes or contributory factors might be: a) modifiable with the intervention; and b) detectable using a robust study design. The charity does not have the resources to undertake a controlled study. Instead they chose an individual level before and after design;

[14] https://www.heatwalkingcycling.org/#how_heat_works

[15] http://www.socialvalueuk.org/resource/a-guide-to-social-return-on-investment-2012/

with a baseline assessment of health before participants start and then again at the completion of the activity, and a final assessment six weeks later. They decide to focus on general health, subjective wellbeing, and physical activity.

The metrics they chose are the EQ5D for general health status, the ONS measure of life satisfaction for subjective wellbeing, and the IPAQ for physical activity. The tools were selected because they are shown to be sensitive to change, appropriate for use on adult populations, and present minimal burden to the participants. Because the EQ5D is a licenced tool the EuroQoL office was contacted to explain the study before the licence to use it was granted. A power calculation was then performed on their primary outcome variable (EQ5D) to identify how many participants they would need.

The data was collected and processed by trained charity volunteers. The analysis sought to find out whether there had been any change in health states between time points. The EQ5D data was used to derive Quality Adjusted Life Year (QALY) estimates. Of note, any QALY gains for a small-scale intervention like this are likely to be substantially less than 1, but this information could still be useful for estimating the cumulative gains over a longer period or across multiple similar projects. The charity worked with a local academic to help them analyse the findings correctly (and cost effectively) by agreeing to make the study part of an MSc student's final thesis work.

Archetype 3. **Modelling population health: Tracking mortality rates in relation to a regional tree planting scheme**

A metro mayor funds the planting of one million trees to promote ecosystem services. She wants to know if the new trees have a positive effect on health. One of the pathways she is interested in is mitigation of poor air quality and associated health impacts. The Mayor asks her team to commission an evaluation.

An evaluability assessment indicates that it might be many years before measurable health impacts are realised and that a direct evaluation might be costly, requiring primary data to be collected on very large numbers of people. Instead the team search for existing data sources which could be used to monitor impacts. They decide to use the new Public Health Outcomes Framework (PHOF) measure 3.01 Fraction of mortality attributable to particulate air pollution at the regional level to track the new trees' contribution to reduced health burden of poor air quality.[16] They use similar regions (matched on topography, baseline vegetation, urban layout and socio-demographics) to benchmark progress and to act as 'quasi-controls'.

[16] https://fingertips.phe.org.uk/search/air%20pollution#page/3/gid/1/pat/6/par/E12000004/ati/102/are/E06000015/iid/30101/age/230/sex/4

The team use published studies to estimate the potential impacts of the new trees on particulate air pollution concentrations, and consequently on mortality rates for the whole city. They also work with the city environment department to collate data from their particulate pollution monitoring programme into the future, to assess whether or not the anticipated reduction in pollution concentrations actually occurs. However, they recognise that these future trends will also be impacted by other city policies such as a low emission zone, which may mean that the direct impact of the tree planting programme may not be precisely measurable.

8. Recommended next steps for the valuing nature community interested in health metrics

Having consulted with a wide range of valuing nature stakeholders we believe there is a strong appetite to take several further steps in this field. More specifically there were three key, inter-related, suggestions which could be taken forward:

a) Developing a traffic-light (quality appraisal) system for existing metrics.

It was suggested that the >270 metrics we identified could be appraised in terms of the metric qualities identified in **Box 3** (e.g. reliability, cost, ease of use etc.). Similar systems have been used elsewhere (Dronavalli, & Thompson, 2015; Health Scotland, 2007) but were beyond the scope of the current project. We strongly support a systematic piece of work to attempt such an appraisal in the future.

b) Identifying the 'golden thread'

One workshop participant proposed a 'golden thread' or a small collection of health metrics to be used across most if not all evaluation projects in this area, alongside more intervention specific metrics. Similar attempts at standardisation have occurred in the field of physical activity, where tools such as the International Physical Activity Questionnaire (IPAQ), were developed to standardise how self-reported physical activity is measured, and subjective wellbeing where the Office for National Statistics (ONS) and the Organisation of Economic Cooperation and Development (OECD) recommend the use of four core questions for all assessments, alongside more domain and respondent specific questions where appropriate. More subtle approaches at harmonisation also exist, for instance with mental health metrics, that do their best to use existing data in as similar ways as possible, but again these are very complex and time-consuming initiatives.[17] At the moment we do not believe such a consensus on health metrics exists in the natural capital field and suggest that such a consensus is developed quickly for the field to progress. The traffic-light exercise may need to occur first to identify the strongest metrics, harmonisation of existing metrics may then be the next logical step before full standardisation is achievable at a later stage.

c) Decision-support tree

Ultimately, we believe a decision-support tree could be constructed to help researchers/practitioners select the most appropriate metric for their study given their theory of change, their study design, and their time and budgetary constraints. The tree would guide researchers through a series of decision points and help reduce the number of options from which to choose, giving ‘traffic-light’ appraisals on the remaining options. Again, this would need to be an extensive, co-ordinated piece of work across the whole community, but we believe could deliver considerable benefits in the long-term by supporting researchers/practitioners select the most appropriate metrics for not just their own use, but ones that can then be synthesised across multiple nature-health studies to give the whole community far greater clarity about the benefits of nature to health cross multiple contexts.

[17] https://www.closer.ac.uk/research-fund-2/data-harmonisation/harmonisation-mental-health-measures-british-birth-cohorts/

Glossary

Term	Definition/Explanation	More information
Confounding factor (variable)	An extraneous (and usually uncontrolled) variable that is allowed to change systematically alongside the variables being studied. In an experiment, an extraneous variable changes systematically along with the independent variable and also has the potential to influence the dependent variable. Confounding variables can threaten internal validity.	Gravetter, F. J. & Forzano, L.-A. B. (2009). Research methods for the Behavioral Sciences. Third Edition. Wadsworth, Cengage Learning.
Control group	The group in a research study that does not receive a treatment or receives a placebo treatment.	Gravetter, F. J. & Forzano, L.-A. B. (2009). Research methods for the Behavioral Sciences. Third Edition. Wadsworth, Cengage Learning.
Cost-Benefit Analysis	A method of reaching economic decisions by comparing the costs of doing something with its benefits.	https://www.nefconsulting.com/our-services/evaluation-impact-assessment/prove-and-improve-toolkits/glossary/ https://valuing-nature.net/demystifying-cost-benefit-analysis
Cross-sectional research design	Analysis of data of variables collected at one given point of time across a sample population (also known as cross-sectional analysis, transverse study or prevalence study).	https://www.questionpro.com/blog/cross-sectional-study/amp/
Disability Adjusted Life Years (DALYs)	"The sum of years of potential life lost due to premature mortality and the years of productive life lost due to disability."	https://www.who.int/mental_health/management/depression/daly/en/
Evaluation	An evaluation is an assessment, conducted as systematically and impartially as possible, of an activity, project, programme, strategy, policy, topic, theme, sector, operational area or institutional performance. Evaluation is a general term for the process of determining what has been achieved during or after a particular activity.	United Nations Evaluation Group (2016). Norms and Standards for Evaluation. New York: UNEG. https://www.nefconsulting.com/our-services/evaluation-impact-assessment/prove-and-improve-toolkits/glossary/

Term	Definition/Explanation	More information
Health	"The ability to adapt and to self-manage, in the face of social, physical and emotional challenges"	Huber, M., Knottnerus, J. A., Green, L., Jadad, A. R., Kromhout, D., Leonard, B., Smid, H. (2011). How should we define health? BMJ, 343:d4163 doi: https://doi.org/10.1136/bmj.d4163
Independent variable	In an experiment, it is the variable manipulated by the researcher.	Gravetter, F. J. & Forzano, L.-A. B. (2009). Research methods for the Behavioral Sciences. Third Edition. Wadsworth, Cengage Learning.
Indicators	Indicators are specific pieces of information, conditions, signs or signals that can be measured to determine whether a given thing has occurred or has been achieved (e.g. an activity, an output, an outcome).	https://www.nefconsulting.com/our-services/evaluation-impact-assessment/prove-and-improve-toolkits/glossary/
Intervention	Process or action that can be repeated over time, or can be a one-off action. Participants are assigned to groups that receive one or more intervention/ treatment (or no intervention) so that researchers can evaluate the effects of the interventions.	Adapted from https://www.nefconsulting.com/our-services/evaluation-impact-assessment/prove-and-improve-toolkits/glossary/ and https://clinicaltrials.gov/ct2/about-studies/glossary
Longitudinal research design	Series of observations or measurement of the same population over a period of time (looks at changes over time).	https://learning.closer.ac.uk/introduction/types-of-longitudinal-research/longitudinal-versus-cross-sectional-studies/
Measures	The items in a research study to which the participant responds (e.g. survey or interview questions, or constructed situations).	http://www.uniteforsight.org/research-methodology/module4
Mental health	Mental health is defined as a state of wellbeing in which every individual realizes his or her own potential, can cope with the normal stresses of life, can work productively and fruitfully, and is able to make a contribution to her or his community.	World Health Organisation https://www.who.int/features/factfiles/mental_health/en/

Glossary

Term	Definition/Explanation	More information
Metric	Using or relating to a system of measurement (for metrics see measures).	https://dictionary.cambridge.org/dictionary/english/metric
Monitoring	To watch and check a situation carefully for a period of time in order to discover something about it. Regularly and systematically collecting and recording information in order to check progress against plans.	https://dictionary.cambridge.org/dictionary/english/monitoring https://www.nefconsulting.com/our-services/evaluation-impact-assessment/prove-and-improve-toolkits/glossary/
Objective Measures	Data based on solid measurements or observations (e.g. physiological data, such as heart rate, or how someone performs on a task, i.e. their score).	
Proxy (indicator)	A proxy indicator is used to replace indicators that are difficult to measure directly.	https://www.nefconsulting.com/our-services/evaluation-impact-assessment/prove-and-improve-toolkits/glossary/
Qualitative Research	Research that is based on observations that are summarised and interpreted as a narrative report (see **Box 2**).	Gravetter, F. J. & Forzano, L.-A. B. (2009). Research methods for the Behavioral Sciences. Third Edition. Wadsworth, Cengage Learning.
Quality-adjusted life year (QALYS)	“QALYs are a measure of the state of health of a person or group in which the benefits, in terms of length of life, are adjusted to reflect the quality of life. One QALY is equal to 1 year of life in perfect health. QALYs are calculated by estimating the years of life remaining for a patient following a particular treatment or intervention and weighting each year with a quality-of-life score (on a 0 to 1 scale). It is often measured in terms of the person’s ability to carry out the activities of daily life, and freedom from pain and mental disturbance.” See **Section 2**.	https://www.nice.org.uk/glossary?letter=q

Term	Definition/Explanation	More information
Quality of Life	An individual's perception of their position in life in the context of the culture and value systems in which they live and in relation to their goals, expectations, standards and concerns. It is a broad ranging concept affected in a complex way by the person's physical health, psychological state, personal beliefs, social relationships and their relationship to salient features of their environment.	World Health Organisation https://www.who.int/healthinfo/survey/whoqol-qualityoflife/en/
Quantitative Research	Research that is based on variables for individual participants or subjects to obtain scores, usually numerical values that are submitted to statistical analyses for summary and interpretation (see also **Box 2**).	Gravetter, F. J. & Forzano, L.-A. B. (2009). Research methods for the Behavioral Sciences. Third Edition. Wadsworth, Cengage Learning.
Reliability	The degree of consistency and stability of measurements. A reliable measurement procedure will produce identical, or near identical measurements, if the same individuals are measured under the same conditions.	Gravetter, F. J. & Forzano, L.-A. B. (2009). Research methods for the Behavioral Sciences. Third Edition. Wadsworth, Cengage Learning.
Sample	A set of individuals selected from a population, usually with the intention that they are representative of the population in the research study.	Gravetter, F. J. & Forzano, L.-A. B. (2009). Research methods for the Behavioral Sciences. Third Edition. Wadsworth, Cengage Learning.
Social Return on Investment (SROI)	Social Return on Investment (SROI) is an outcomes-based measurement tool that helps organisations to understand and quantify the social, environmental and economic value they are creating.	https://www.nefconsulting.com/our-services/evaluation-impact-assessment/prove-and-improve-toolkits/sroi/
Subjective (Self-report) Measures	Personal perceptions, opinions or self-reported experience.	

Glossary

Term	Definition/Explanation	More information
Tool	Measurement tools are instruments (e.g. scales, surveys, interviews, observations) that are used by researchers and practitioners to aid in the assessment or evaluation of study participants, clients or patients.	https://guides.lib.uw.edu/hsl/measure
Validity (of a measurement procedure)	The degree to which a measurement process measures the variable it claims to measure.	Gravetter, F. J. & Forzano, L.-A. B. (2009). Research methods for the Behavioral Sciences. Third Edition. Wadsworth, Cengage Learning.
Validity (of a research study)	The degree to which the study accurately answers the question it was intended to answer.	Gravetter, F. J. & Forzano, L.-A. B. (2009). Research methods for the Behavioral Sciences. Third Edition. Wadsworth, Cengage Learning.
Wellbeing	Wellbeing can be understood as how people feel and how they function, both on a personal and a social level, and how they evaluate their lives as a whole.	New Economics Foundation (2012) Measuring Wellbeing: A guide for practitioners, London: New Economics Foundation.

References

Denzin, N. K. & Lincoln, Y. (2000). The Discipline and Practice of Qualitative Research. In: Denzin NK and Lincoln Y (eds) *Handbook of Qualitative Research* 2ND Edition. Sage, London Page 8

Dronavalli, M. & Thompson, S. C. (2015). A systematic review of measurement tools of health and wellbeing for evaluating community-based interventions. *Journal of Epidemiology and Community Health, 69:* 805-815.

Markevych, I., Schoierer, J., Hartig, T., Chudnovsky, A., Hystad, P., Dzhambov, A. M., ..& Lupp, G. (2017). Exploring pathways linking greenspace to health: Theoretical and methodological guidance. *Environmental Research, 158,* 301-317.

Ogilvie, D., Cummins, S., Petticrew, M., White, M., Jones, A. & Wheeler, K. 2011. Assessing the Evaluability of Complex Public Health Interventions: Five Questions for Researchers, Funders, and Policymakers. *The Milbank Quarterly*, 89, 206-225.
https://onlinelibrary.wiley.com/doi/full/10.1111/j.1468-0009.2011.00626.x

Lead Authors:

Drs. Deborah Cracknell, Rebecca Lovell, Benedict Wheeler and Mathew White at the *European Centre for Environment & Human Heath, University of Exeter.*

Key consultants throughout the process:

Val Woods and Dr. Sunita Sarkar at the *Centre for Ecology and Hydrology,* and Dr. Ruth Garside, Prof. Lora Fleming and Prof. Ian Bateman at the *University of Exeter.*

Contributors who helped set the agenda and commented on earlier drafts included:

London/Leeds workshop attendees:

Jim Burt, *Natural England*
Alison Darlow, *Natural England*
Dr. Nicola Dempsey, *University of Sheffield*
Dr. Eirini Flouri, *University College London*
Prof. Marc Jones, *Manchester Metropolitan University*
Dr. Rosie McEachan, *NHS/Born in Bradford*
Dr. Amy Mizen, *Swansea University*
Dr. Liz O'Brien, *Forest Research*
Rebecca Robertson, *University of Leeds*
Prof. Sarah Rodgers, *University of Liverpool*
Dr. Mike Rogerson, *University of Essex*
Prof. Catherine Ward Thompson, *University of Edinburgh*
Dr. Sue Williams, *Natural Resources Wales*
Dr. Cheryl Willis, *Natural England*

On-line survey contributors:

Lauriane Suyin Chalmin-Pui, *University of Sheffield*
Dr Cláudia Costa, *University of Coimbra*
Dr Mike Gent, *Public Health England*
Angela Hands, *Public Health England*
Mikaël J.A. Maes, *University College London*
Dr Jean McKendree, *Stockholm Environmental Institute*
Jim Stewart-Evans, *Public Health England*
Jim Wharfe, *Independent Consultant*

Suggested citation: Cracknell, D, Lovell, R, Wheeler, B and White, M. 2019. Demystifying Health Metrics, Valuing Nature Paper VNP19.

This work was funded by the Valuing Nature Programme, which is a 5-year year £7M research programme which aims to improve understanding of the value of nature both in economic and non-economic terms, and improve the use of these valuations in decision making. It funds interdisciplinary research and builds links between researchers and people who make decisions that affect nature in business, policy-making and in practice. See www.valuing-nature.net

The Valuing Nature Programme is funded by the Natural Environment Research Council, the Economic and Social Research Council, the Biotechnology and Biological Sciences Research Council, the Arts and Humanities Research Council, and the Department for Environment, Food and Rural Affairs.

Additional funding was also provided through the National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Environmental Change and Health at the London School of Hygiene and Tropical Medicine in partnership with Public Health England (PHE), and in collaboration with the University of Exeter, University College London, and the Met Office. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, the Department of Health, Public Health England, or of any of the Research Councils and Defra.

Additional resources downloadable through link below:

A list of key health metric reviews across the natural capital literature (word document)

A taxonomy of commonly used health metrics, both within and beyond the natural capital literature (tabulated information)

https://valuing-nature.net/demystifying-health-metrics-1

Further information contact the Programme Coordination Team:
info@valuing-nature.net

@ValuingN
valuing-nature.net

Design by Countryscape