Stepping towards causation in studies of neighborhood and environmental effects: How twin research can overcome problems of selection and reverse causation

Glen E. Duncan[a,b], **Brianna Mills**[a], **Eric Strachan**[c], **Philip Hurvitz**[d], **Ruizhu Huang**[e], **Anne Vernez Moudon**[a,d], and **Eric Turkheimer**[f]

[a]Department of Epidemiology, University of Washington, Seattle, WA 98195, U.S.A

[b]Interdisciplinary Graduate Program in Nutritional Sciences, University of Washington, Seattle, WA 98195, U.S.A

[c]Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, WA 98195, U.S.A

[d]Department of Urban Design and Planning, Urban Form Lab, University of Washington, Seattle, WA 98195, U.S.A

[e]Interdisciplinary PhD Program in Urban Design and Planning, Urban Form Lab, University of Washington, Seattle, WA 98195, U.S.A

[f]Department of Psychology, University of Virginia, Charlottesville, VA 22904, U.S.A

Abstract

No causal evidence is available to translate associations between neighborhood characteristics and health outcomes into beneficial changes to built environments. Observed associations may be causal or result from uncontrolled confounds related to family upbringing. Twin designs can help neighborhood effects studies overcome selection and reverse causation problems in specifying causal mechanisms. Beyond quantifying genetic effects (i.e., heritability coefficients), we provide examples of innovative measures and analytic methods that use twins as quasi-experimental controls for confounding by environmental effects. We conclude that collaboration among investigators from multiple fields can move the field forward by designing studies that step toward causation.

Correspondence: Glen E. Duncan, University of Washington, Box 353410, Seattle, WA 98195-3410, telephone: 206-616-2680, fax: 206-685-1696, duncag@u.washington.edu.

The authors have no competing interests to declare in relation to this manuscript.

MeSH Keywords

Causality; Environment Design; Lifestyle Risk Reduction; Social and Built Environments; Twin Studies

INTRODUCTION

The role of built and social environments in supporting healthy lifestyles has garnered increasing attention as research paradigms shift from a focus on individuals to an assessment of macro-level influences operating within ecological models of health. Researchers have long recognized that both built and social environments affect individual health. However, studies that seek to specify the causal mechanisms linking "neighborhood effects" to health outcomes suffer from two intractable problems: selection and reverse causation.

To illustrate the complexity of these problems, we present Figure 1 as a simplified example of the multiple levels of influence that interact to produce the current obesity epidemic in the U.S. (National Heart Lung and Blood Institute, 2004). It is important to appreciate that these influences have an effect not only on behavior, but also on each other through interaction, reinforcement, and even antagonism. Past approaches to slow the obesity epidemic have been less than effective (Flegal et al., 2010; Ogden et al., 2007), primarily targeting individual-level influences (Fig. 1, upper two boxes). Experts increasingly recognize that outcomes in a population are a function of macro-level influences (Fig. 1, bottom two boxes) that affect the lives of all people. *To have the greatest population-level effect, multi-level interventions should start with environmental changes and end with individual-level programming, rather than the reverse.* New methods make it possible to measure energy balance behaviors within the social and physical environments that support or hinder them. Objective macro-environmental data help formulate policies that regulate large-scale environmental influences on healthy lifestyles, such as funding for urban design and infrastructure that support active modes of transportation. A focus on the built and social environment and its implications for policy and legislation is thus a necessary step if we are to manage the epidemics of obesity and related chronic diseases.

In the following sections we outline the problems of selection and reverse causation and argue for the utility of twin study designs in overcoming them. We present examples from our research with the University of Washington Twin Registry (UWTR), including novel analytic approaches, and conclude with suggestions for future research collaborations.

THE SELECTION PROBLEM

Imagine a study in which participants are randomly assigned to residential locations and followed longitudinally to investigate the effects of various social and built environments on their health. For ethical, legal, and practical reasons such a study is impossible. In the real world choices of residential neighborhood are nonrandom, and studies have yet to adequately control for this (Diez Roux, 2002; Oakes, 2004, 2006; Subramanian, 2004). Most existing statistical models simply adjust for individual characteristics associated with residential location, such as age, ethnicity/race, and socio-economic status. More recently,

reasons for choosing a neighborhood have been measured in surveys and subsequently controlled for in analyses (Frank et al., 2007). Even studies that adjust for such characteristics inevitably include unrecognized or difficult to measure factors. Such factors are left uncontrolled and the resulting bias reduces inferential validity (Oakes, 2006). This, in a nutshell, is the selection problem.

Selection problem can also refer to the possibility that individuals select residential environments based on genetics or family upbringing (Duncan et al., 2012; Whitfield et al., 2005; Willemsen et al., 2005). The selection variables, rather than any putative environmental effects, may be responsible for findings that link environmental characteristics to health outcomes. Twins are ideal study subjects for overcoming this problem. Twins are always the same age and ethnicity/race, effectively controlling for those key demographic factors. Monozygotic (MZ) or identical twins are matched genetically and always of the same sex, so within-pair variation in health and environmental risk cannot be attributed to differences in genetics or sex. Among dizygotic (DZ) or fraternal twins, confounding by sex can be eliminated by studying same-sex pairs. Further, because MZ and DZ twins share a common family of origin (twins are almost invariably reared together) differences within pairs cannot be attributed to family background or childhood exposures. In cross-sectional studies, twins can be used as quasi-experimental controls for confounding by genetic, family, and early life environmental effects on outcomes that cannot be held constant by random assignment (Turkheimer, 2008). Because twins often become discordant in neighborhood residence and lifestyle later in life, we can detect environmental effects on health while reducing the confounding inherent in studies of unrelated individuals in nonrandom environments (Moffitt, 2005).

THE REVERSE CAUSATION PROBLEM

The problem of reverse causation (or direction of causation) refers to the possibility that behaviors and health play causal roles in the choice of residential location, rather than the converse. Once again, it is neither ethical nor practical to randomize people to different residential environments and follow them prospectively to determine environmental effects on outcomes of interest. The cross-lagged panel design is a method commonly used in psychology and behavioral research for testing spuriousness by comparing cross-lagged correlations in cases in which the independent variable cannot be experimentally manipulated (Kenny, 1975), such as arises with respect to residential environments. Below we describe how this model has been adopted by twin researchers to a twin cohort allows for supports a genetically informed, cross-lagged panel design that obtains environmental and health data at two or more time points to conduct analyses that support directional causal inferences while controlling for covariation that originates in either genetic or shared environmental variation.

In a cross-lagged design (Fig. 2) two traits, a purported cause (e.g. walkability) and outcome (e.g walking level), are measured at two time points. The outcome at Time 2 is regressed on the outcome at Time 1 (stability), and on the purported cause at Time 1 (cross-lagged regression). The cross-lagged regression represents the effect of the purported cause at Time 1 on the change in outcome between Time 1 and 2. This coefficient can be compared to its

converse, the regression of the purported cause at Time 2 on the purported effect at Time 1, conditional on the stability of the purported cause between Times 1 and 2. If the relationship between Time 1 and 2 is causal, the cross-lagged coefficient is significant, and significantly larger than the regression of the cause on the effect. Using twins, cross-lagged regression coefficients are estimated while controlling for covariation between the purported cause and outcome that originates in either genetic (A) or shared environmental variation (C) that is shared between cause and effect, known as a genetically informed cross-lag design. The residual variation in the purported effect is also decomposed into genetic and environmental components. The partitioning of variance into genetic, shared, and non-shared environmental variation (E) is accomplished with the usual classical twin methods (i.e., based on the correlation between the genetic backgrounds of a pair of twins being 1.0 in MZs and 0.5 in DZs). To the extent there is a causal relationship between the predictor and outcome, the cross-lagged regression will remain significant even after the association arising from familial background variables has been accounted for.

Although the twin design controls for the possibility that individuals select residential environments based on genetics or family upbringing (i.e., common genetic and environmental background in the selection of environments), there remains the possibility of within-pair confounds that the twin design does not control. For example, if one twin is sent away to a private residential school while the other remains home in an urban public school, these experiences may have an influence on later residential selection, or attitudes about walking or public transportation infrastructure, that will not be controlled using the twin design. In this instance, additional data collection would be appropriate, such as reasons for moving to the present neighborhood, desirability of potential public investments in the community, and neighborhood preferences to name a few examples.

OTHER USES OF TWINS TO ADDRESS CAUSATION ISSUES

Innovative Measurements

Linking high-resolution objective measures of built and social environmental exposures with survey and biometric data obtained from a community-based sample of adult twins (see Table 1), the UWTR serves as our example of the possibilities inherent in studying twins longitudinally. The cross-lag study described above and shown in Fig. 2 uses walkability measures for time 1 (baseline) derived from geocoded residential addresses supplied by twins during a recruitment survey; physical activity and other health outcomes are also obtained from this comprehensive survey. Longitudinal addresses (used to generate built environment measures) and health outcomes come from follow-up surveys administered every other year.

One current study of neighborhood effects includes objectively measured physical activity levels and eating episodes in a real space-time continuum. Participants are outfitted with an accelerometer, GPS data logger, and mobile phone for continuous tracking in time and space over two weeks. The data from these three independent data collection tools are joined into a "lifelog" (Kang et al., 2013) indexed by common time stamps across device. An example for one twin pair is shown in Fig. 3. The data from these lifelogs are being used to investigate multiple issues: using MZ twins who live apart to determine the association between the

home built environment (i.e., the home "neighborhood") and objectively measured levels of walking and total physical activity; and to determine whether twins are more active (or often eat) in their home neighborhood than in distal environments outside commonly pre-defined neighborhood boundaries (e.g. ½ mile of home). Beyond overcoming measurement bias inherent in self-report, this study addresses issues of neighborhood definitions and spatial-temporal measurement of behaviors with ecological exposures. Discussion of these issues was begun years ago (Kawachi and Subramanian, 2007) and remains salient today.

Gene by Environment Interactions

A common misconception regarding twin designs is that their chief utility lies in quantifying genetic effects. This view overlooks the potential of twins as quasi-experimental controls for confounding by genetic and environmental effects (Turkheimer, 2008). Advances in twin research have provided powerful new analytical techniques to detect gene by environment (G x E) interactions (Purcell, 2002; van der Sluis et al., 2012), where the proportions of variance attributable to genetic and environmental sources are themselves moderated by environmental variables.

Ongoing projects in this field include analyses of how levels of physical activity and walkability moderate the contribution of genetics to BMI, and how area-level social deprivation moderates the contribution of genetics to depression. In a work in preparation we demonstrate that higher activity levels are associated with lower heritable variance in BMI, even when controlling for the possibility that people with low BMI select environments or activities that require more physical exertion.

"Omics" Studies

Studies have started using omics approaches from systems biology to characterize differences in the epigenetic molecular profiles of co-twins (Gordon et al., 2012), in part to expose differences in the genotype-to-phenotype paths traveled by genomic information and discover where environmental influences impinge upon these paths. Molecular profiles can now be routinely performed at the level of the epigenome (Gordon et al., 2012; Kaminsky et al., 2009), transcriptome (Correa and Cheung, 2004; Tan et al., 2010), and proteome (Naidoo, 2011). With twins, such profiling can assist our search for the molecular source of observable phenotypic differences between MZ twins discordant for chronic conditions. For example, intra-pair differences in mean methylation level of the serotonin transporter gene, which has a role in regulating food intake, body weight, and energy balance, were significantly correlated with intra-pair differences in body mass index, body weight, and waist circumference in a recent study of MZ twins from the Vietnam Era Twin Registry (Zhao et al., 2013).

CONCLUSION

The causation issues faced by neighborhood effects studies are not minor, but nor are they intractable. Correctly accounting for individuals within their physical and social context requires both innovation and collaboration, borrowing from other disciplines to address challenges that researchers face as a collective. Twin studies, while known chiefly in the

realm of genetics, can and should play an important role in epidemiologic and population studies more broadly.

The UWTR is a unique resource for investigators across multiple disciplines who wish to apply twin methods to control for genetic and common familial factors in studies of neighborhood or other environmental effects, and is the only one of few community-based registries in the U.S. that actively recruits twins on an ongoing basis. We are unaware of any other twin cohort that obtains spatial data to support longitudinal investigations of the effects of environmental exposures on health behaviors and outcomes. For these reasons, we intend to make full use of the combination of genetic, geographic, and behavioral data offered by the UWTR.

Even though we firmly believe in the potential of twin cohort studies to overcome many limitations of neighborhood effects research, this approach has its own weaknesses. It is important to recognize that while twins can be genetically identical, expression of those genes can be very different and that twins are in fact individuals that may respond to their shared (or unshared) environment in very different ways. Even a population-based cohort such as the UWTR is subject to certain biases, such as any differences that may exist between twins who agree to join the cohort and those who do not. We are currently undertaking a systematic evaluation of this possibility. Critics have pointed out that while twin studies can address individual-level confounding, little has been done so far to address psychological or cultural confounding, as shown in the middle two boxes of Fig. 1. We encourage researchers interested in behavior change to consider the potential of longitudinal and pooled cohort twin studies. Social and cultural confounding is an ongoing issue, as most twin cohorts are lacking in racial, ethnic, and national diversity. Without this diversity it is difficult to disentangle the physical and social aspects of environment; for example, we cannot readily compare walkability in Tokyo to that in London, or the effect of living in the same neighborhood on African-American twins to that of Hispanic twins. As twins and multiple births become more common, twin studies should prioritize population-based sampling strategies to ensure representativeness. Pooling data from twin registries and cohorts around the world may also help address issues around lack of diversity and representativeness; during the past 10 years the number of twin registries has increased greatly with representation from over 28 countries and five continents (Hur and Craig, 2013). This pooled approach would have its own issues, especially related to population stratification and accounting for phenotypic variation. Population studies that recruit individuals but have been able to identify twin pairs, such as the National Longitudinal Study of Adolescent Health (Add Health), may also be important additions to pooled data efforts. This article serves as an open call to any investigator interested in discussing the potential of twin studies or seeking to use data or samples from the UWTR cohort in their research.

Future research into the effect of neighborhood on health should incorporate novel measurements and twin designs to overcome these and other limitations, with the goal of establishing causal links between built and social environments and health outcomes.

Acknowledgments

References

Correa CR, Cheung VG. Genetic variation in radiation-induced expression phenotypes. Am J Hum Genet. 2004; 75:885–890. [PubMed: 15359380]

Diez Roux AV. Invited commentary: places, people, and health. Am J Epidemiol. 2002; 155:516–519. [PubMed: 11882525]

Duncan GE, Dansie EJ, Strachan E, Munsell M, Huang R, Vernez Moudon A, Goldberg J, Buchwald D. Genetic and environmental influences on residential location in the US. Health Place. 2012; 18:515–519. [PubMed: 22377617]

Flegal KM, Carroll MD, Ogden CL, Curtin LR. Prevalence and trends in obesity among US adults, 1999–2008. JAMA. 2010; 303:235–241. [PubMed: 20071471]

Frank LD, Saelens BE, Powell KE, Chapman JE. Stepping towards causation: do built environments or neighborhood and travel preferences explain physical activity, driving, and obesity? Soc Sci Med. 2007; 65:1898–1914. [PubMed: 17644231]

Gordon L, Joo JE, Powell JE, Ollikainen M, Novakovic B, Li X, Andronikos R, Cruickshank MN, Conneely KN, Smith AK, Alisch RS, Morley R, Visscher PM, Craig JM, Saffery R. Neonatal DNA methylation profile in human twins is specified by a complex interplay between intrauterine environmental and genetic factors, subject to tissue-specific influence. Genome Res. 2012; 22:1395–1406. [PubMed: 22800725]

Hur YM, Craig JM. Twin registries worldwide: an important resource for scientific research. Twin Res Hum Genet. 2013; 16:1–12. [PubMed: 23324558]

Kaminsky ZA, Tang T, Wang SC, Ptak C, Oh GH, Wong AH, Feldcamp LA, Virtanen C, Halfvarson J, Tysk C, McRae AF, Visscher PM, Montgomery GW, Gottesman II, Martin NG, Petronis A. DNA methylation profiles in monozygotic and dizygotic twins. Nat Genet. 2009; 41:240–245. [PubMed: 19151718]

Kang B, Moudon AV, Hurvitz PM, Reichley L, Saelens BE. Walking Objectively Measured: Classifying Accelerometer Data with GPS and Travel Diaries. Med Sci Sports Exerc. 2013; 45:1419–1428. [PubMed: 23439414]

Kawachi I, Subramanian SV. Neighbourhood influences on health. J Epidemiol Community Health. 2007; 61:3–4. [PubMed: 17183006]

Kenny DA. Cross-lagged panel correlation - test for spuriousness. Psychological Bulletin. 1975; 82:887–903.

Moffitt TE. The new look of behavioral genetics in developmental psychopathology: Gene-environment interplay in antisocial behaviors. Psychological Bulletin. 2005; 131:533–554. [PubMed: 16060801]

Naidoo N. Potential of proteomics as a bioanalytic technique for quantifying sleepiness. J Clin Sleep Med. 2011; 7:S28–30. [PubMed: 22003327]

National Heart Lung and Blood Institute. Predictors of obesity, weight gain, diet, and physical activity workshop. National Institutes of Health; 2004. Available at: http://www.nhlbi.nih.gov/meetings/workshops/predictors/summary.htm [Accessed Jan 03, 2013]

Oakes JM. The (mis)estimation of neighborhood effects: causal inference for a practicable social epidemiology. Soc Sci Med. 2004; 58:1929–1952. [PubMed: 15020009]

Oakes JM. Commentary: advancing neighbourhood-effects research--selection, inferential support, and structural confounding. Int J Epidemiol. 2006; 35:643–647. [PubMed: 16556642]

Ogden CL, Carroll MD, McDowell MA, Flegal KM. Obesity among adults in the United States--no statistically significant change since 2003–2004. NCHS Data Brief. 2007:1–8. [PubMed: 19389313]

Purcell S. Variance components models for gene-environment interaction in twin analysis. Twin Res Hum Genet. 2002; 5:554–571.

Subramanian S. The relevance of multilevel statistical methods for identifying causal neighborhood effects. Soc Sci Med. 2004; 58:1961–1967. [PubMed: 15020011]

Tan Q, Ohm Kyvik K, Kruse TA, Christensen K. Dissecting complex phenotypes using the genomics of twins. Funct Integr Genomics. 2010; 10:321–327. [PubMed: 20145969]

Turkheimer, E. A better way to use twins for developmental research. In: Kruse, I., editor. LIFE Newsletter. Max Planck Institute for Human Development; Berlin, Germany: 2008. p. 2-5.

van der Sluis S, Posthuma D, Dolan CV. A Note on False Positives and Power in G x E Modelling of Twin Data. Behav Genet. 2012; 42:170–186. [PubMed: 21748401]

Whitfield JB, Zhu G, Heath AC, Martin NG. Choice of residential location: Chance, family influences, or genes? Twin Res Hum Genet. 2005; 8:22–26. [PubMed: 15836806]

Willemsen G, Posthuma D, Boomsma DI. Environmental factors determine where the Dutch live: Results from the Netherlands Twin Register. Twin Res Hum Genet. 2005; 8:312–317. [PubMed: 16176714]

Zhao J, Goldberg J, Vaccarino V. Promoter methylation of serotonin transporter gene is associated with obesity measures: a monozygotic twin study. Int J Obes. 2013; 37:140–145.
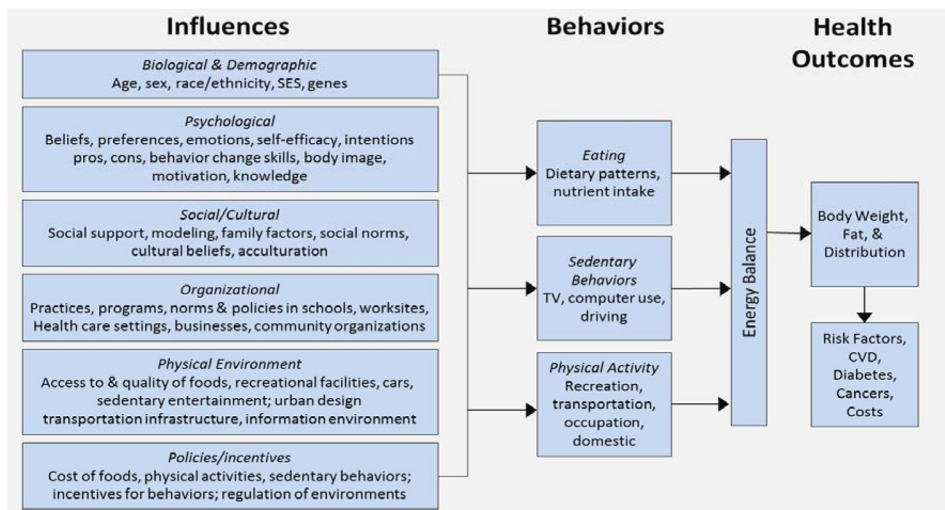
Figure 1.
An ecological model of diet, physical activity, and obesity.

Abbreviation: CVD = cardiovascular disease.

Developed for the NHLBI workshop on predictors of obesity, weight gain, diet, and physical activity; August 4–5, 2004, Bethesda, MD.
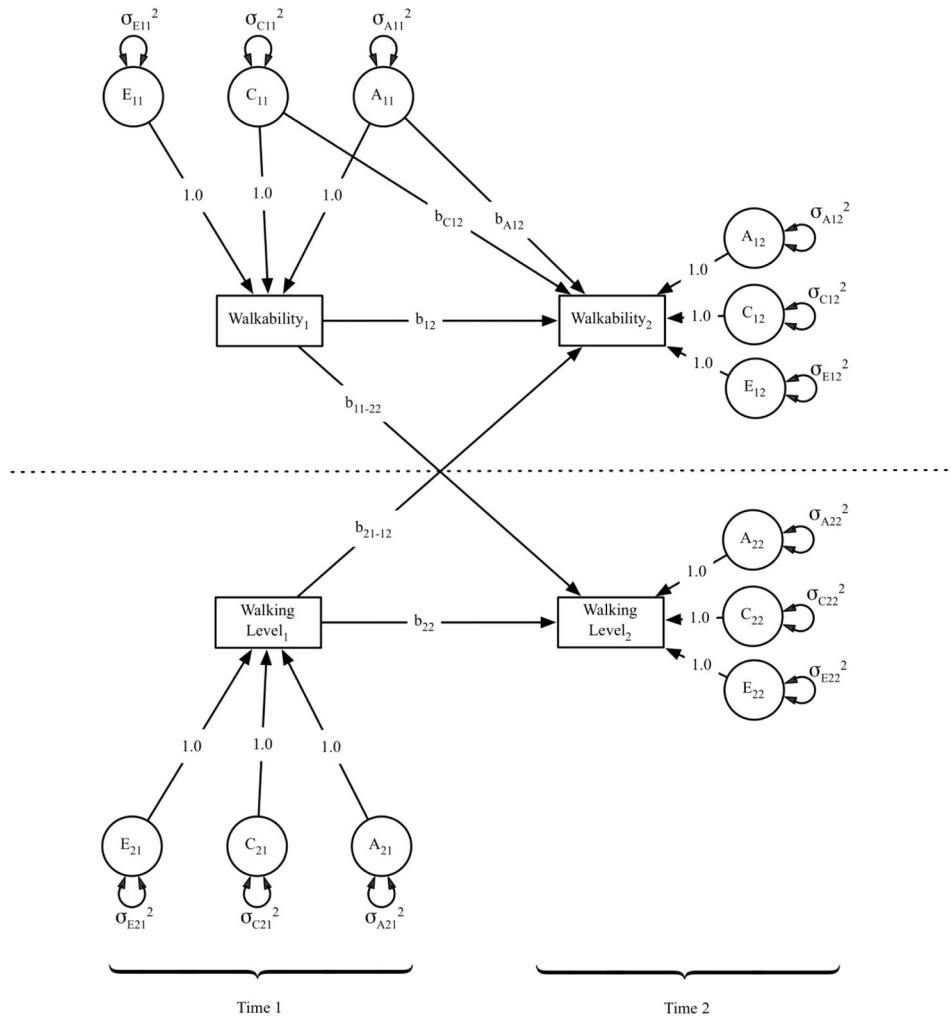
Figure 2.
A genetically informative cross-lagged panel analysis.

The latent variable path diagram for one twin shows that walking level at Time 2 (lower right) is regressed on walking level at Time 1 (lower left, the stability coefficient path b_{22}) and walkability at Time 1 (upper left, the cross-lagged causal path b_{11-22}). The cross-lagged regression coefficients are estimated while controlling for covariation between the purported cause and outcome that originates in either genetic (A) or common environmental variation (C) that is shared between cause and effect, while also accounting for nonshared environmental variation (E), which also includes an error term. The residual variation in the purported effect is also decomposed into genetic and environmental components (noted by σ for A, C, E).
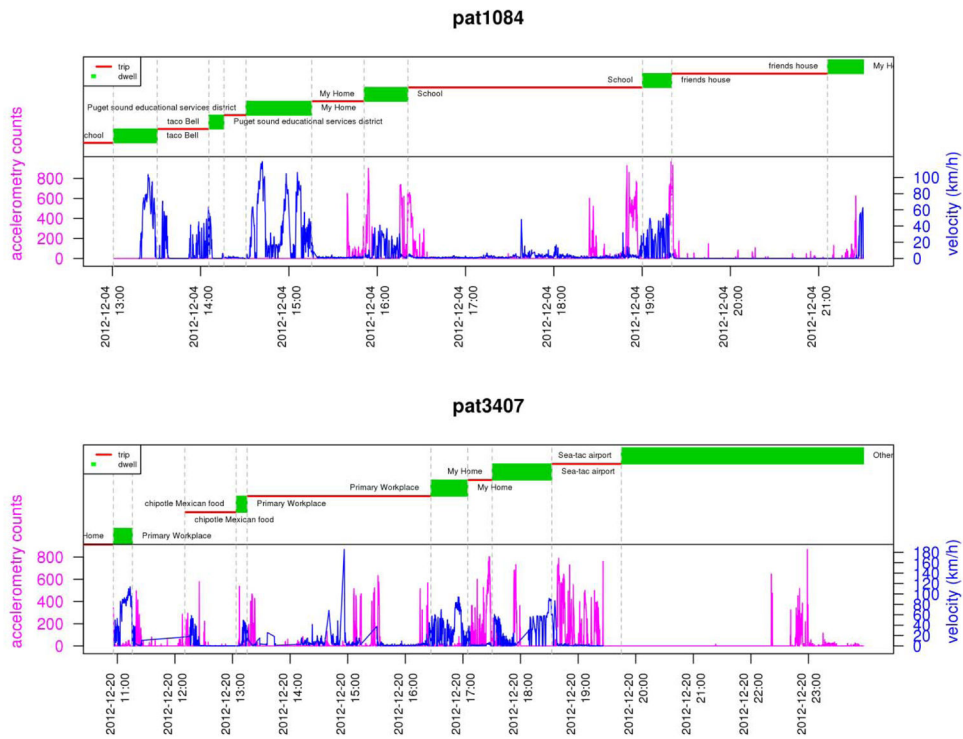
Figure 3.
One-day lifelogs from a twin pair.

The upper panel shows self-reported data on place (red) and trip (green) from the travel diary along with objective accelerometry counts (magenta) and GPS velocity (blue). The lower panel maps travel patterns for the same individuals over the same days, with markers shown at hourly intervals.

Table 1

Current measures addressing neighborhood and environmental effects available for participants in the University of Washington Twin Registry.

Assessment area	Twin Pair N Baseline data collection 2008-present	Twin Pair N Follow-up data collection 2010–2012[†]
Self-report Survey Data		
Zygosity[*]	4811	-
Sociodemographics	4811	2299
Height /weight	4811	2299
Medical history	4811	2299
Allergies	4811	2299
Sleep/physical activity	4811	2299
Eating habits	4811	2299
Health functioning	4811	2299
Alcohol/tobacco use	4811	2299
Pain conditions/fatigue	4811	2299
Head injury	4811	2299
Depression/Anxiety	4811	2299
Posttraumatic stress	4811	2299
Resilient coping	4811	2299
Perceived stress[**]	3734	-
Cognitive functioning[**]	3734	-
Five factor personality traits[**]	3734	-
Environmental Exposures		
Walkability	3530	1132
Urban sprawl	3530	1132
Normalized Difference Vegetation Index	3530	1132
Social deprivation	3530	1132
Property values	3530	1132
Crime	3530	1132
DNA		
Saliva	3422[**]	-

[*] Self-reported zygosity collected in the enrollment survey.

[**] Data collection began in 2010, so no follow-up data are available.

[†] Second follow-up data collection to begin in 2014.