The relevance of transport to promote physical activity : addressing challenges related to the measurements and the observational analysis of transport-related physical activity, and the simulation of shifts in transportation mode

Ruben Brondeel

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PIERRE ET MARIE CURIE

Spécialité : Epidémiologie

École doctorale Pierre Louis de santé publique à Paris : épidémiologie et sciences de l'information biomédicale

présentée par

Ruben Brondeel

pour obtenir le grade de :

DOCTEUR DE L'UNIVERSITÉ PIERRE ET MARIE CURIE

Sujet de la thèse :

La pertinence du transport pour promouvoir l'activité physique: une prise en compte des défis liés à la mesure, à l'analyse empirique et à la simulation des changements de modes de transport

soutenue le 16/12/2016

devant le jury composé de :

Basile Chaix	Dir. de recherche INSERM	Directeur de thèse
Jay Kaufman	Prof. McGill University	Rapporteur
David Ogilvie	Prof. Cambridge University	Rapporteur
Jimmy Armoogum	Ch. de recherche IFSTTAR	Examinateur
Jean-Michel Oppert	Prof. Université Paris VI	Examinateur

Thèse réalisée dans l'institut Pierre Louis d'Epidémiologie et de Santé Publique – UMR S 1136 dans le cadre du réseau doctoral en santé publique animé par l'EHESP

Basile, thank you for all your advice. You made my Phd a very interesting learning and working experience, during which you helped me to develop skills crucial for a researcher. Discussing research with you is a true pleasure, and I hope we can continue collaborating in the future. Thanks for your patience and persistence.

Camille, Laura, Jasper, Antoine, Tarik, and Julie, thank you very much for reading the manuscript. Your remarks truly made this a better manuscript.

I am very grateful to all the members of the team, past and present, for your advice and friendship. I enjoyed the coffee breaks, the birthday cakes, the short and not so short after-work drinks, the trips we made together, ... I am convinced that our friendships will continue, no matter where our careers might lead us.

My friends from Ghent, thank you for being there for me. Every time I knock on your doors, you open them with a smile, even though you know I am going to empty the fridge. You were a great support during all the years I studied and worked in Ghent; and despite the distance, you stayed as close to me while I lived in Paris.

My friends from Paris, thanks for the past 5 years. Some helped me out with housing, others with work, others with drinks... You will all be missed. I am sorry I can't stay, but it's just time to go.

Special thanks go to my family. Thank you for your love and constant support.

Abstract

The relevance of transport to promote physical activity: Addressing challenges related to the measurements and the observational analysis of transport-related physical activity, and the simulation of shifts in transportation mode

Introduction Physical activity has an important impact on various health outcomes, and transport accounts for a substantial part of total physical activity. Previously observed social inequalities in mode of transport used are likely to lead to social inequalities in transport-related physical activity. This has, as yet, not been investigated with objective accelerometer data.

The use of accelerometer data to evaluate transport-related physical activity is not without challenges. Many decisions made in regards to how raw accelerometer data are processed can substantially alter the physical activity levels estimated. In order to differentiate between transport and non-transport related physical activity, it is necessary to have very precise data on mobility patterns, i.e. where, when, and how people go between places.

Objectives This PhD work aimed to improve measures of transport-related physical activity and to report empirical findings on the transport-related physical activity of adults aged 35 to 83 years living in Ile-de-France. First, we investigated the impact of the epoch length - the time unit in which accelerometer data is processed - on physical activity indicators. Second, we developed a model to automatically detect the transportation mode used, in order to facilitate the collection of mobility pattern data. Third, we integrated two datasets from the same population; a small sample with precise sensor data and a large representative sample with survey data. This enabled the analysis of accelerometer-based indicators for a large sample. Fourth, we investigated through simulation the impact of the choice of transportation mode on transport-related moderate-to-vigorous physical activity (T-MVPA), with a particular focus on potential social inequalities in T-MVPA.

Methods The RECORD GPS Study collected GPS and accelerometer data for 236 participants over a 7-day period, resulting in the observation of 7425 trips.

The Global Transport Survey (Enquête Globale Transport) collected data over one day, resulting in the observation of 82084 trips for 21332 participants. The methods used include random forest prediction models, geographical information systems, negative binomial regressions, and multiple imputation.

Results and discussion Four articles are incorporated into this PhD thesis. In the first article, we evaluated the accelerometer-based measure of moderate-to-vigorous physical activity (MVPA), and found that shorter epochs resulted in considerably larger estimates of the daily accumulated MVPA. Moreover, there was a larger impact of the epoch length when using tri-axial data compared to uni-axial data. This is an important finding, as most accelerometer studies will be based on tri-axial data in the future. This finding supports calls from the literature for further harmonisation of accelerometer based indicators of physical activity.

An algorithm based on a random forest prediction model was developed that correctly predicted 90% of the transportation modes in the RECORD GPS dataset. This algorithm could facilitate data collection by informing prompted recall systems; however, future work should further optimise integrated prediction methods based on machine learning to make completely passive data collection possible.

By integrating the two datasets (RECORD GPS and Global Transport Survey), we observed an average 18.9 minutes of daily T-MVPA (95% confidence interval: 18.6; 19.2 minutes) in this representative sample from Ile-de-France. Participants with a higher level of education did more T-MVPA than their less educated counterparts. In contrast, people with a higher household income did less T-MVPA per day.

The simulated scenarios of transportation mode shift with the highest impact were those promoting walking or discouraging car use, compared to the scenarios promoting biking or public transport. The lower impact of the latter two strategies may be attributable to the low prevalence of biking and to a reverse effect of promoting public transport due to a decreased number of walking trips (even if public transport promotes physical activity compared to private motorised transport). The simulations also showed that interventions may increase inequalities by education level in transport-related physical activity.

Conclusion This PhD work was the first study to combine a very detailed dataset - including GPS, accelerometer, and mobility behaviour data - and a large-scale transport survey; both originating from the same population of 35 to 83 year old people in Ile-de-France. The innovative methods developed for this work enabled the analysis of T-MVPA for a large representative sample. The results were the first based on accelerometer data for a large population sample that gave insight into transport-related physical activity. This study provided insight into social inequalities in transport-related physical activity, and simulated the impact of transportation mode shift. Future research will need to further improve and standardise data collection of accelerometer-based physical activity. Only such improvements in data collection and analytical techniques will make it possible to understand the complex associations between motivational, behavioural, and environmental determinants of physical activity behaviour.

Résumé

Introduction L'activité physique a un impact important sur la santé populationnelle, et les comportements de transport constituent une partie substantielle de l'activité physique totale. De plus, les inégalités sociales observées dans les choix des modes de transport utilisés sont susceptibles de conduire à des inégalités sociales dans l'activité physique liée au transport. Or, à ce jour, ce problème majeur de santé publique n'a été étudié que via l'utilisation de données auto-rapportée d'activité physique (questionnaire, journal, etc.), et n'a pas fait l'objet d'études utilisant la précision des données objective d'accélérométrie.

L'utilisation de données de l'accéléromètre pour évaluer l'activité physique liée au transport n'est pas sans défis. De nombreuses décisions prises en ce qui concerne la façon dont les données de l'accéléromètre brutes sont traitées peuvent modifier substantiellement les niveaux d'activité physique estimés. Afin de différencier entre l'activité physique liée au transport et l'activité physique hors transport, il est nécessaire d'avoir des données très précises sur la mobilité quotidienne ; c.à.d. où, quand, et comment les gens se déplacent.

Objectifs Ce travail de thèse a pour objectif d'améliorer les mesures de l'activité physique liées au transport et d'utiliser ces nouvelles mesures dans des études de cas empirique sur l'activité physique liée au transport des adultes âgés de 35 à 83 ans résidant en Ile-de-France. Premièrement (i), nous avons étudié l'impact de l'unité de temps dans laquelle les données de l'accéléromètre sont traité sur les indicateurs de l'activité physique. Dans un deuxième temps, nous avons développé un modèle pour détecter automatiquement le mode de transport utilisé, afin de faciliter la collecte des données de mobilité. Troisièmement (iii), nous avons intégré deux ensembles de données de la même population ; un échantillon avec des données de capteurs précises (n = 236) et un grand échantillon représentatif avec des données base sur des questionnaires (n = 21332). Cela a notamment permis d'analyser des indicateurs d'activité physique liée au transport basés sur l'accéléromètre pour un large échantillon. Quatrièmement, nous avons étudié par simulation l'impact du choix du mode de transport sur l'activité physique modérée à vigoureuse liée au transport (T-APMV), avec un regard particulier sur les

inégalités sociales potentielles dans T-APMV.

Méthodes Des données GPS et d'accéléromètre ont été collectée dans le cadre de « RECORD étude GPS » pour 236 participants sur une période de 7 jours, ce qui permis d'observer 7425 déplacements. L'Enquête Globale Transport a recueilli des données sur une population de 21332 participants sur une période d'un jour, comptabilisant l'observation de 82084 voyages pour 21332 participants. Les méthodes statistiques utilisées pour l'analyse des modes de transports ainsi que de leurs déterminants incluent des modèles forêt aléatoires (Random Forests), des régressions binomiales négatives utilisant la méthode de l'imputation multiple ; ainsi que l'utilisation de systèmes d'information géographique.

Résultats et discussion Cette thèse présente quatre articles empiriques faisant état des travaux effectués. Dans le premier article, nous avons évalué l'impact de l'unité de temps des données de l'accéléromètre sur la mesure de l'activité physique modérée à vigoureuse (APMV). Nous avons mis en évidence que les unités de temps plus courtes ont donné lieu à des estimations d'APMV beaucoup plus importantes. De plus, nous avons observé un impact plus grand de l'unité du temps lors de l'utilisation des données tri-axiaux par rapport aux données uni-axial. Cette constatation est importante, car la plupart des études utilisant des accéléromètres seront basés sur les données tri-axiaux dans l'avenir. Cette constatation renforce les recommandations de la littérature d'une harmonisation plus poussée des indicateurs de l'activité physique basés sur l'accéléromètre.

Dans le deuxième article de cette thèse, nous avons créé un algorithme basé sur un modèle de prédiction Random Forests et correctement prédit 90% des modes de transport utilisés par la population RECORD GPS. Cet algorithme pourrait faciliter la collecte de données en informant les systèmes de rappel guidé. Cependant, d'autres études devraient permettre d'optimiser les méthodes de prédiction intégrées basées sur machine learning afin de pouvoir développer une méthode de collecte de données complètement passif.

En intégrant deux ensembles de données complémentaires (RECORD GPS et Enquête Globale Transport), nous avons observé 18,9 minutes T-APMV par jour en moyenne (95% intervalle de confiance : 18,6; 19,2 minutes) dans cet échantillon représentatif de l'Ile-de -France. Les participants ayant un niveau d'éducation plus

élevé ont plus de T-APMV que les participants moins instruits. En revanche, les personnes ayant un revenu du ménage plus élevé ont moins T-APMV par jour.

Lors de la simulation du changement de mode de transport, les pratiques qui avaient le plus d'impact étaient les personnes privilégiant la marche ou étant découragé de l'utilisation de la voiture, comparé à ceux promouvant l'utilisation du vélo ou des transports en commun. L'impact plus faible des deux dernières stratégies peut être attribuable à la faible prévalence des déplacements en vélo en Île-de-France et à un effet inverse de la promotion des transports publics résultant en une diminution du nombre de voyages à pied (même si les transports en commun favorisent l'activité physique par rapport au transport motorisé privé). Les simulations ont également montré que les interventions peuvent accroître les inégalités d'activité physique liée au transport, et ce, selon le niveau d'éducation.

Conclusion Ce travail de thèse a été la première étude à combiner une base de données très détaillées - comprennent GPS, accéléromètre, et des données de comportement de mobilité - et une base de donnees d'une enquête de transport à grande échelle provenant de la même population d'individus que RECORD étude GPS (individus âgés de 35-83 ans et résidant en Ile-de-France). Les méthodes innovantes développées pour ce travail de thèse ont permis l'analyse de T-APMV pour un large échantillon représentatif. Nos résultats, basés sur des données d'accéléromètrie au sein d'un large échantillon représentatif, sont également les premiers à donner un aperçu de l'activité physique liée au transport. Cette étude a fourni un aperçu des inégalités sociales d'activité physique liée au transport, et a simulé l'impact du changement de mode de transport. Les recherches futures devront améliorer et normaliser la collecte de données de l'activité physique mesurées par accéléromètres. Seules de telles améliorations dans la collecte de données et dans les techniques d'analyse permettront de comprendre les associations complexes existant entre les déterminants motivationnels, comportementaux et environnementaux d'activité physique.

Table of contents

Scientific publications and communications during the PhD

Published articles

Brondeel, R., Pannier, B., and Chaix, B. Associations of socioeconomic status with transport-related physical activity: combining a household travel survey and accelerometer data using random forests. Journal of Transport & Health, (In press)

Brondeel, R., Pannier, B., and Chaix, B. (2015). Using GPS, GIS, and Accelerometer Data to Predict Transportation Modes. Medicine & Science in Sports & Exercise, 47(12):2269–2675

Chaix, B., Dustin, D., **Brondeel, R.**, Méline, J., El Aarbaoui, T., Pannier, B., and Merlo, J. (2016). A GPS-based methodology to analyze environment health associations at the trip level: case-crossover analyses of built environments and walking. American Journal of Epidemiology, (In press)

Perchoux, C., Chaix, B., **Brondeel, R.**, and Kestens, Y. (2016). Residential buffer, perceived neighborhood, and individual activity space: new refinements in the definition of exposure areas - The RECORD Cohort Study. Health & Place, (In press)

Perchoux, C., Kestens, Y., **Brondeel, R.**, and Chaix, B. (2015). Accounting for the daily locations visited in the study of the built environment correlates of recreational walking (the RECORD Cohort Study). Preventive Medicine, 81:142–149

Chaix, B., Kestens, Y., Duncan, S., Merrien, C., Thierry, B., Pannier, B., **Brondeel, R.**, Lewin, A., Karusisi, N., Perchoux, C., et al. (2014). Active transportation and public transportation use to achieve physical activity recommendations? A combined GPS, accelerometer, and mobility survey study. International Journal of behavioral Nutrition and Physical Activity, 11(1):124

Brondeel, R., Weill, A., Thomas, F., and Chaix, B. (2014). Use of healthcare services in the residence and workplace neighbourhood: The effect of spatial accessibility to healthcare services. Health & Place, 30:127–133

Karusisi, N., Thomas, F., Méline, J., **Brondeel, R.**, and Chaix, B. (2014). Environmental conditions around itineraries to destinations as correlates of walking for transportation among adults: The RECORD Cohort Study. PLoS ONE, 9(5):e88929

Chaix, B., Méline, J., Duncan, S., Jardinier, L., Perchoux, C., Vallée, J., Merrien, C., Karusisi, N., Lewin, A., **Brondeel, R.**, and Kestens, Y. (2013). Neighborhood environments, mobility, and health: Towards a new generation of studies in environmental health research. Revue d'epidemiologie et de Sante Publique, 61:S139–S145

Articles under review

Brondeel, R., Kestens, Y., and Chaix, B. Simulating the impact of transport mode shifts on transport-related physical activity.

Brondeel, R., Schipperijn, J., Kestens, Y., and Chaix, B. The impact of epoch lengths on moderate-to-vigorous physical activity, light physical activity and sedentary behavior estimates for adults.

El Aarbaoui, T., Méline, J., **Brondeel, R.**, Chaix, B. Short-term association between personal exposure to noise and heart rate variability: a sensor-based study.

Oral communications

Invited communications

Brondeel, R., and Chaix, B. Use of Healthcare Services in the Residence and Workplace Neighbourhood: The Effect of Spatial Accessibility to Healthcare Services. Conférence Francophone SIG 2014, Versailles, France

Brondeel, R., and Chaix, B. L'influence du niveau d'éducation et du revenu sur la pratique d'activité physique liée au transport. ARS-Ile-de-France. Invited talk during the seminary: "Séminaire profession banlieue - Santé, pratique de l'activité physique et renouvellement urbain" 2016, Paris, France

Brondeel, R., and Chaix, B. Social inequalities in transport-related physical activity: Measurement issues and analysis. Department of Movement and Sport Sciences 2016, Ghent, Belgium

Other communications

Brondeel, R., and Chaix, B. Using GPS, GIS, and accelerometer data to predict transportation modes. ISBNPA 2015, Edinburg, UK

Brondeel, R., and Chaix, B. Social disadvantages in transport-related physical activity: Combining a household travel survey and accelerometer data using Random Forest. ALR 2016, Clearwater Beach, Florida, USA

Brondeel, R., and Chaix, B. Associations of the socioeconomic status with transport-related physical activity and the impact of transport interventions: a simulation study based on random forests. ISBNPA 2016, Cape Town, South-Africa

Brondeel, R., and Chaix, B. The impact of epoch lengths on MVPA and sedentary

time estimations. ISBNPA 2016, Cape Town, South-Africa

Brondeel, R., and Chaix, B. Applying sensor-based knowledge to large population surveys to simulate the impact of interventions on urban and transport systems (during the symposium 'Causal inference methods for estimating health impacts of environmental policies'). ISEE 2016, Rome, Italy

Press Release

'Les transports en commun, un générateur d'activité physique quotidienne lié à la mobilité.' A press release in collaboration with the Observatory of Mobility in Ile-de-France (L'Observatoire de la mobilité en Île-de-France). http://www.omnil.fr/IMG/pdf/reperes5_tc_sante.pdf

The document used for the press release was based on the article 'Associations of socioeconomic status with transport-related physical activity: combining a household travel survey and accelerometer data using random forests' (Brondeel et al. 2016. Journal of Transport & Health).

The study was discussed 33 times in the following media: printed articles (e.g. Le Parisien, Premiere heure), internet articles (e.g. TF1.fr, LeFigaro.fr, 20Minutes.fr, Ladepeche.fr, France3.fr), radio programs (France inter, VoltageFM and NRJ), one television program (iTele), and two news agencies (AFP and Relaxnews).

List of Abbreviations

CI	Confidence Interval
CNAMTS	Caisse Nationale de l'Assurance Maladie des Travailleurs Salariés
CPM	Counts Per Minute
DRIEA	Direction Régionale et Interdépartementale de l'Equipement et de l'Aménagement
EGT	Enquête Globale Transport
GIS	Geographical Information Systems
GPS	Global Positioning System
HDOP	Horizontal Dilution Of Precision
IAU	Institut d'Aménagement et d'Urbanisme
IGN	Institut national de l'information géographique et forestière
INSEE	Institut National de la Statistiques et des Etudes Economiques
IPC	Centre d'Investigations Préventives et Cliniques
LPA	Light Physical Activity
MET	Metabolic Equivalent Task
MVPA	Moderate-to-Vigorous Physical Activity
PDOP	Positional Dilution Of Precision
RECORD	Residential Environment and CORonary heart Disease
SB	Sedentary Behaviour
STIF	Syndicat des transports d'Île-de-France
T-MVPA	Transport-related Moderate-to-Vigorous Physical Activity
VDOP	Vertical Dilution Of Precision
WHO	World Health Organisation

List of figures and tables

List of Figures

List of Tables

Résumé en Français

Introduction

L'activité physique est connue pour avoir un effet protecteur contre plusieurs maladies non transmissibles telles que les maladies coronariennes,[1,2] l'obésité,[3] le diabète de type 2,[2] la maladie d'Alzheimer,[2] la démence,[2,4] la dépression[5] et certains cancers tels que le cancer du rein,[6] du côlon,[7,8] de la prostate,[8] des testicules,[7,8] du sein,[7–9] de l'ovaire.[8] Un manque d'activité physique est reconnu par l'Organisation Mondiale de la Santé (OMS) comme la quatrième cause de mortalité.[10] L'OMS recommande, par conséquent, au moins 150 minutes d'activité physique d'intensité modérée à vigoureuse (APMV) par semaine pour les adultes.

Les comportements de transport constituent une partie substantielle de l'activité physique totale. Les interventions concernant les modes de transport pour promouvoir l'activité physique sont traditionnellement concentrées sur la marche[11] et le vélo.[12] Récemment, des études ont montré le potentiel des interventions touchant aux transport public.[13]

Les résultats de ces interventions de promotion de l'activité physique dans l'utilisation des transports demeurent incohérents.[11,14] Les différences dans ces résultats sont dues, entre autres, à la grande diversité entre les mesures de l'activité physique ainsi qu'une efficacité et un progrès relativement lent dans le développement des méthodes d'intervention

pour la promotion de l'activité physique et du transport actif.[15] La plupart des études d'intervention de transport ont été évaluées par des mesures de niveaux d'activité physique auto-rapportées. Ce type d'indicateur est utile pour évaluer les changements dans le comportement de mode de transport liés à l'activité physique, mais donne peu d'informations sur l'impact exact d'une intervention sur le niveau d'activité physique. à ce jour, peu d'études ont étudié l'impact des interventions de transport à l'aide de données issues de l'accéléromètrie pour mesurer l'activité physique.[16]

L'utilisation de données de l'accéléromètre pour évaluer l'activité physique liée au transport n'est pas sans défis. De nombreuses décisions prises en ce qui concerne la façon dont les données de l'accéléromètre brutes sont traitées peuvent modifier substantiellement les niveaux d'activité physique estimés. Un aspect important dans le traitement des données souvent négligé est la longueur de l'unité de temps dans laquelle les données de l'accéléromètre sont traitées. Les indicateurs de l'intensité de l'activité couramment utilisés ont été étalonnés avec une unité de temps de 60 s. Néanmoins, les accéléromètres actuels permettent une collecte et un traitement de données dans des unités de temps plus courtes. Des indicateurs adaptés aux unités de temps plus courtes peuvent cependant fournir des estimations de l'APMV et du comportement sédentaire (SB) très différentes.[17–19]

Afin de différencier l'activité physique liée au transport et l'activité physique hors transport, il est nécessaire d'avoir des données très précises sur la mobilité quotidienne (c.à.d. où, quand, et comment les gens se déplacent). La collecte de ces données est une tâche très exigeante pour les chercheurs et les répondants.[20,21] Globalement, deux méthodes peuvent être envisagées pour faciliter la collection de données de l'accéléromètre dans la mesure de l'activité physique liée au transport. La première solution 1) est de faciliter la mesure de la mobilité. La deuxième solution 2) consiste à éviter la collecte de données de l'accéléromètre en prédisant l'activité physique.

Ce travail de thèse a pour objectif d'améliorer les mesures de l'activité physique liées

au transport et d'utiliser ces nouvelles mesures dans des études de cas empiriques sur l'activité physique liée au transport d'adultes âgés de 35 à 83 ans résidant en Ile-de-France. Premièrement (i), nous avons étudié l'impact de l'unité de temps dans laquelle les données de l'accéléromètre sont traitées sur les indicateurs de l'activité physique. Dans un deuxième temps, nous avons développé un modèle pour prédire le mode de transport utilisé, afin de faciliter la collecte des données de mobilité. Troisièmement (iii), nous avons intégré deux ensembles de données de la même population ; un échantillon avec des données de capteurs précises (n = 236) et un grand échantillon représentatif avec des données basées sur des questionnaires (n = 21,332). Cela nous a permis, notamment, d'analyser des indicateurs d'activité physique liés au transport basés sur l'accéléromètre sur un large échantillon. Quatrièmement (iv), nous avons étudié, par simulation, l'impact du choix du mode de transport sur l'activité physique modérée à vigoureuse liée au transport (T-APMV), avec un regard particulier sur les inégalités sociales potentielles dans T-APMV.

Méthodes

Nous avons utilisé les données de deux échantillons d'études indépendantes : l'échantillon de l'étude RECORD GPS et l'échantillon de l'Enquête Globale Transport (EGT). Ces deux bases de données ont été considérées comme complémentaires pour étudier l'activité physique liée au transport d'une population adulte (entre 35 et 83 ans) dans la région d'Ile-de-France.

Dans la deuxième vague de l'étude de cohorte RECORD,[21] 410 participants ont été invités à participer à l'étude GPS RECORD. Les participants portaient un GPS BTQ1000XT (QStarz) et un accéléromètre GT3X + (Actigraph) sur la hanche droite avec une ceinture élastique dédiée, pour la journée de recrutement et 7 jours supplémentaires, pour toute la journée : du réveil jusqu'à l'heure du coucher. Les participants

devaient remplir un journal de déplacement en signalant, durant les 7-8 jours, chaque arrivée et départ sur les différents lieux d'activité. Les données GPS ont été recueillies toutes les 5 secondes. Sur la base des données GPS, l'application 'Mobility Web Mapping' a été utilisée pour visualiser les modèles d'activité et de transport sur une carte par participant et par jour. L'application a été utilisée pour sonder les participants sur l'activité réalisée à chaque lieu d'activité visité et sur les modes de transport privilégiés pour chaque déplacement. Cette procédure a abouti à l'identification de 7,138 déplacements pour 229 participants.

L'Enquête Globale Transport (EGT) est une enquête de mobilité effectuée tous les 10 ans en Ile-de-France. Au cours des entretiens face-à-face, les données ont été recueillies pour tous les déplacements effectués la veille de l'entrevue. Nous avons sélectionné les participants âgés entre 35 et 83 ans, ce qui donne 82,084 déplacements effectués par 21332 personnes.

La mesure d'APMV de Sasaki et al.[22] a été adaptée pour notre étude. Concernant cette mesure, une minute de APMV est définie comme une minute au cours de laquelle 2,690 CPM (counts par minute) ou plus ont été enregistrés sur le vecteur magnitude. Le vecteur magnitude résume le mouvement enregistré sur les trois axes orthogonaux du GPS, en prenant la racine carrée de la somme de la CPM de chaque axe au carré.

Au cours des enquêtes EGT et RECORD, les participants ont rapporté une séquence chronologique des modes de transport pour chaque déplacement. Aux fins de la modélisation, cette information a été codée dans une variable de mode de transport composée de quatre catégories : «marche », «vélo », «véhicule motorisé privé »et «transports en commun ». Les autres variables utilisées incluent des indicateurs d'inégalités sociales (par exemple, niveau d'éducation), des variables sociodémographiques (âge), des caractéristiques de déclenchement (par exemple de la durée d'un déplacement), et des caractéristiques environnementales du quartier résidentiel et des quartiers de départ / arrivée des déplacements (par exemple, la densité de population).

Deux méthodes ont été développées au cours de cette thèse : l'intégration de deux bases de données, et une méthodologie de simulation pour étudier les changements de mode de transport. La méthode de l'intégration des données présentées dans ce travail est basée sur deux bases de données de la même population (Ile de France) avec un grand ensemble de variables communes. Ceci nous a permis d'utiliser les meilleurs aspects des deux ensembles de données : les données très détaillées de l'accéléromètre de la base de données RECORD GPS et le large échantillon de la base de données EGT. Dans une première étape, un modèle de prédiction pour APMV au niveau du déplacement a été construit sur les données RECORD GPS. Des prédictions d'APMV ont ensuite été conduites pour chaque déplacement dans la base de données EGT où les mêmes variables prédictives étaient disponibles pour chaque déplacement. Enfin, l'APMV par déplacement a été résumé par jour pour chaque participant d'EGT. Cela a abouti à une estimation pour T-APMV par jour qui a été mise en corrélation avec des variables individuelles, par exemple le niveau d'éducation.

Le but de l'étude de simulation était d'évaluer, par simulation, l'impact des changements dans les modes de transport. Douze scénarios de simulation ont été examinés, trois pour chacun des 4 modes de transport : la marche, le vélo, les transports en commun, et de transport motorisé privé. Dans une première étape, le mode de transport pour un nombre prédéterminé de déplacements a été changé par un mode alternatif. Dans une deuxième étape, la durée du déplacement a été estimée pour les déplacements auxquels un nouveau mode de transport a été attribué à l'étape 1. La prédiction était basée sur un modèle de random forests pour la durée des déplacements dans les données de EGT. Dans une dernière étape, l'APMV était prédite pour les déplacements avec un mode de transport modifié. La simulation de chaque scénario a été répétée 100 fois pour éviter les erreurs d'échantillonnage aléatoire dans les résultats. D'autres méthodes statistiques utilisées dans ce travail de thèse incluent les forêts aléatoires, la régression binomiale négative, et l'imputation multiple.

Résultats et discussion

Objectif 1. Des unités de temps de 1, 15 et 30 secondes produisaient des estimations d'APMV et SB beaucoup plus importantes que des unités de temps de 60 secondes. Notre étude a également montré que l'impact de l'unité de temps était très grand par rapport aux autres décisions de traitement de données.

Objectif 2. Une méthodologie a été développée pour automatiser la détection des modes de transport, qui a correctement prédit 90% des modes de transport dans l'ensemble de données RECORD GPS. Ce résultat est comparable ou supérieur à celui d'études antérieures.[23–27]

Objectif 3. Une méthodologie d'intégration de données a été développée, basée sur les bases de données RECORD GPS et EGT. Le modèle de prédiction pour T-APMV basé sur la base de données RECORD GPS expliquait 67% de la variance de cette variable. Le modèle a ensuite été appliqué à l'ensemble de données EGT au niveau du déplacement ; et ensuite, l'APMV liée au transport a été calculée par jour.

Objectif 4. Une approche de simulation a permis l'estimation des variations de T-APMV associées à des changements dans les modes de transport observés. Pour chacun des quatre modes de transport, trois scénarios de promotion du transport actif ont été construits. Des modèles de prédiction random forests ont été utilisés pour sélectionner les trajets pour lesquels le mode de transport a dû être modifié, pour prédire la nouvelle durée du déplacement, et enfin pour prédire la T-APMV par déplacement. La T-APMV accumulée par jour a été calculée dans chaque scénario et comparée à l'estimation initiale.

Objectif 5. En intégrant deux ensembles de données complémentaires (RECORD GPS et Enquête Globale Transport), nous avons observé 18,9 minutes T-APMV par jour en moyenne (95% intervalle de confiance : 18,6 ; 19,2 minutes) dans cet échantillon représentatif de l'Ile-de-France. En comparaison avec les recommandations de l'OMS (150 minutes d'APMV par semaine, ou 30 minutes par jour pour la plupart des jours au cours de la semaine), une moyenne de 19 minutes d'APMV liée au transport indique une contribution considérable à l'activité physique totale du transport. Toutefois, les recommandations de l'OMS sont elles-mêmes basées sur des études utilisant des enquêtes ; ils sont toujours en attente d'une mise à jour basée sur des études basées sur l'accéléromètre.[10] Il n'est donc pas clair dans quelle mesure les mesures de APMV par accéléromètre peuvent être comparées aux recommandations de l'OMS.

Les participants ayant un niveau d'éducation plus élevé ont plus de T-APMV que les participants moins instruits. En revanche, les personnes ayant un revenu du ménage plus élevé ont moins de T-APMV par jour. Les personnes ayant un revenu du ménage supérieur ont moins de T-APMV par jour. Dans des études précédentes, un niveau d'éducation plus élevé a été associé à plusieurs minutes de marche pour le transport,[28] plus de déplacements avec les modes de transport actifs,[28,29] et plus de déplacements à vélo ;[30] tandis que le revenu plus élevé a été associé à moins de minutes de marche et de moins fréquents déplacements avec les modes actifs.[28]

Objectif 6. Les simulations ont porté dans chaque scénario sur l'un des quatre modes de transport. Avant les changements de mode de transport (à savoir les données observées), la marche représentait une moyenne de 6,8 minutes de T-APMV par jour, le vélo 1,2 minutes, les transports en commun 6,8 minutes, et le transport privé motorisé 4,2 minutes par jour. Les pratiques qui avaient le plus d'impact étaient celles privilégiant la marche ou décourageant l'utilisation de la voiture, comparé à ceux promouvant l'utilisation du vélo ou des transports en commun. L'impact plus faible des deux dernières stratégies peut être attribuable à la faible prévalence des déplacements

en vélo en Ile-de-France et à un effet inverse de la promotion des transports publics résultant en une diminution du nombre de voyages à pied (même si les transports en commun favorisent l'activité physique par rapport au transport motorisé privé). Les simulations ont également montré que les interventions peuvent accroître les inégalités d'activité physique liées au transport, et ce, selon le niveau d'éducation.

Conclusion

Ce travail de thèse a contribué à l'évaluation critique des indicateurs les plus couramment utilisés d'activité physique : activité physique modérée à vigoureuse (APMV), le comportement sédentaire (SB), et l'activité physique légère (LPA). La conclusion selon laquelle la longueur d'unité de temps a un impact considérable sur ces indicateurs d'intensité de l'activité s'ajoute à une série d'études qui ont évalué les décisions prises lors de la collecte et du traitement des données de l'accéléromètre. D'une part, cette constatation renforce les recommandations de la littérature d'une nécessité d'harmonisation plus poussée entre les indicateurs de l'activité physique basés sur l'accéléromètre ;[31] d'autre part, ces résultats soutiennent le développement de nouveaux indicateurs qui pourraient éviter certaines, sinon la plupart, des décisions en utilisant des données brutes et des algorithmes de machine learning.

Le deuxième article est basé sur les avancées méthodoloques pour l'identification automatique de la mobilité quotidienne. Basé sur un algorithme développé précédemment pour détecter les lieux de départ et d'arrivée et l'heure des déplacements,[32] cette étude développe une méthode pour détecter automatiquement le mode de transport. Avec un nombre de source de données croissant (par exemple les smartphones), il reste un travail important pour la recherche future afin d'optimiser l'utilisation de ce type de données.

Dans un troisième et un quatrième article, nous avons appliqué une approche d'intégration

de données à des indicateurs de l'activité physique dans un grand échantillon représentatif de la population en Ile-de-France. Cette approche a permis de nous éloigner de la catégorisation habituelle entre les déplacements actifs tels que la marche et le vélo, et les déplacements non-actifs tels que voiture, moto, et les déplacements en transport public. Pour comprendre l'activité physique liée aux transports, il est nécessaire de tenir compte des épisodes actifs qui se produisent lors d'un déplacement non actif, comme la marche jusqu'à un arrêt de bus ; et les épisodes non-actifs ou moins actifs lors de déplacements actifs, tels que la marche lente et un épisode d'attente à un feu rouge.

Dans le dernier article, nous avons utilisé une approche de simulation pour étudier l'impact des changements de mode sur l'activité physique liée au transport. Cette approche utilise une méthodologie innovante basée sur le machine learning pour étudier l'impact approximatif des interventions. Cette méthodologie de simulation nécessite encore un certain degré d'essai et de réglage fin pour améliorer la précision des inférences faites. Cependant, elle a déjà fait ses preuves en tant que méthode très puissante et flexible pour modéliser des scénarios d'interventions.

Les résultats empiriques, basés sur des données d'accéléromètre au sein d'un large échantillon représentatif, sont les premiers à donner un aperçu de l'activité physique liée au transport. Cette étude a fourni un aperçu des inégalités sociales d'activité physique liée au transport, et a simulé l'impact du changement de mode de transport. Les recherches futures devront améliorer et normaliser la collecte de données de l'activité physique mesurées par accéléromètres. Seules de telles améliorations dans la collecte de données et dans les techniques d'analyse permettront de comprendre les associations complexes existant entre les déterminants motivationnels, comportementaux et environnementaux de l'activité physique.

1. Introduction

1.1 Physical activity and health

Physical activity is known to be protective against several noncommunicable diseases including coronary heart disease,[1,2] obesity,[3] type 2 diabetes,[2] Alzheimer's disease,[2] dementia,[2,4] depression,[5] and certain cancers such as renal,[6] colon / colorectal,[7,8] prostate,[8] testicular,[7,8] breast,[7–9] ovarian,[8] and endometrial cancer.[8,33] A lack of sufficient physical activity is recognised by the World Health Organisation (WHO) as the fourth leading cause of mortality.[10] Physical activity is a protective factor for health through its associations with biological factors such as energy expenditure, sex hormone levels, the immune function, insulin levels, blood pressure, triglyceride concentrations, high-density lipoprotein cholesterol concentrations and others.[1,7,34]

The WHO and many governments have adopted health promotion strategies focused on encouraging higher levels of participation in regular physical activity.[4,35] Central in these promotion strategies are the recommendations offered by the WHO.[10] In summary, the WHO recommends at least 150 minutes of moderate-to-vigorous physical activity (MVPA) per week for adults and 60 minutes per day for children older than 5 years. For adults, the recommended levels can also be obtained by 75 minutes of vigorous physical activity per week, or by an equivalent combination of moderate and vigorous physical activity. Furthermore, higher levels of physical activity than those

recommended are likely to give extra health benefits; and inactive people benefit from increasing their physical activity even without reaching the recommendations.[10]

Despite a long tradition in physical activity research and despite the health plans to promote regular physical activity which have been adopted around the world, continuous surveillance systems are still very rare.[4,19] Therefore, there is little conclusive evidence about the recent trends in physical activity worldwide or even on the current physical activity levels.[36,37]

The information which is available is predominantly based on self-reported measures and the results from these studies are very mixed.[4] For example, two renowned sources reported very different results on one of the most commonly used indicators of physical activity at the population level: the percentage of inactive people, i.e. people that do not reach the recommendations by the WHO. In a review in the Lancet in 2012, Hallal et al.[36] reported that an estimated 34% of adults in Europe were inactive; while the WHO[38] reported 71% of adults in Europe to be inactive, a result cited from a study conducted in 15 European states.[39]

Reported temporal trends in physical activity show equally inconsistent results, with studies reporting both positive, stable and negative trends over time.[4,40,41] These mixed results in both prevalence and temporal trends of physical activity are partly due to measurement problems.[41] For example, Knuth et al.[40] showed in a review that the inconsistent results in the temporal trends were partially due to the domain of physical activity that was measured. They reported that in high income countries leisure-time physical activity has been increasing (slightly), while occupational physical activity is decreasing. Therefore, depending on the focus of the questionnaire, studies reported very different trends. Some authors[41,42] have even argued that self-reported indicators are not appropriate for measuring (temporal trends in) physical activity due to its inherent inaccuracies. They strongly advocate the use of sensors, such as accelerometers, to measure physical activity.

1.2 Domains of physical activity

Roughly four domains of physical activity are recognised in research: leisure-time, occupational and transport-related physical activity, and household chores.[43] However, the definitions of these domains are not standardised.[34,44] Leisure-time, for example, could be defined as exercise activities only,[45] or all non-occupational activities,[34] or something in between such as all non-occupational activities excluding household chores.[46]

Considering these domains separately can provide important information for health promotion. There are also indications that the different domains of physical activity have different impacts on health. Holtermann et al.[46] found a protective effect of leisure-time physical activity on the long-term absence of sickness, while occupational physical activity had the inverse effect. Similarly, Hinrichs et al.[45] found a protective effect of vigorous leisure time physical activity on mobility limitation in old age and a detrimental effect from vigorous occupational physical activity. Hu et al.[47] found that leisure-time physical activity was protective against mortality and cardiovascular events, while occupational physical activity was a risk factor for both health outcomes. In a review study of Samitz et al.,[34] this latter finding on all-cause mortality was not confirmed. They did, however, find a lessened protective effect from occupational physical activity compared to leisure-time physical activity.

A clear explanation of why leisure-time and occupational physical activity impact health differently is still missing. It is possible that important confounding factors were omitted or badly measured. Also, all of these studies are based on self-reported physical activity, with some studies using different questionnaires;[34] and currently there is little or no accelerometer-based studies that were able to differentiate between the domains. However, it is clear that studying the different domains of physical activity may important in order to gain a better understanding of the impact of physical

activity on health.

Transport and physical activity

Traditionally, leisure-time and occupational physical activity have received the most attention in physical activity research. More recently, transport-related physical activity was recognised as an important physical activity domain in research [13,21,48] and an important target for health prevention authorities. Relative to the other domains, few studies have examined the impact of active transport on health. [49] However, there is emerging evidence that active transport is associated with improved cardiovascular health, [50,51] physical wellbeing, [52] and reduced overweight and obesity, [50,53] independently of the level of physical activity undertaken in other domains.

Transport is not only an important physical activity domain; it may be a key opportunity for effective physical activity interventions. Transport-related physical activity is a regular, incidental type of physical activity. [54,55] Focusing on incidental physical activity makes it possible for people to integrate physical activity into their day-to-day life. More people participate in incidental physical activity than organised physical activity. [54,56] Therefore, promoting incidental physical activity such as transport-related physical activity might have a larger effect at the population level than promoting organised physical activities such as sports or exercise, especially for populations who are less likely to participate in organised physical activity.

1.3 The impact of transport interventions on physical activity

Transport interventions to improve physical activity levels have traditionally focused on walking [11] and biking; [12] and recently there has been some attention to the potential of public transport interventions. [13] Interventions have had individual approaches

tailored at motivated persons or groups, or community-wide approaches targeting the whole community, for example built environment interventions.[15] Physical activity interventions have been implemented in work and school settings, neighbourhoods (for example changing transport infrastructure), or whole regions (for example media campaigns). The results have been mixed and the impacts have often been modest from this wide variety of interventions.[11,14] A recent review[16] on scaled-up physical activity interventions – i.e. interventions first found successful in research and later introduced on a large scale – was more optimistic about the long term impacts; they found several successful examples, including some transport interventions, while concluding that there is still much work to do.

Measurement inconsistency could be one reason, among others, for the mixed results in effectiveness and the relatively slow advancements in transport intervention methods to improve physical activity.[15] Most transport intervention studies were evaluated by self-report measures of physical activity levels. Therefore, interventions were often evaluated on indicators that are only approximate measures for physical activity such as the 'usual transport mode to go to work / school' or the self-reported 'number of active trips per week'. This type of indicators is useful to approximately evaluate the changes in transport behaviour related to physical activity, but gives limited information on the exact impact on physical activity levels. To date, there is limited accelerometer-based evidence on the impact of transport interventions on physical activity levels.[16]

1.4 Social inequalities in physical activity

Social inequality in health is a good indicator for social inequity and social injustice.[57] Social inequity in health has long been recognised as a fundamental problem,[58] but it is still often overlooked in research and policy. Therefore, the WHO has put equity centrally within their global agenda, the 2030 Agenda for Sustainable Development,[59]

urging countries to 'leave no one behind'.

Social inequalities have been consistently found in many areas of health,[60] including physical activity.[44] The evidence base specifically for transport-related physical activity is scarce. Socioeconomic status was found to lead to social inequalities in transport behaviour related to physical activity in some studies.[44] For example, a higher personal level of education has been associated with more minutes of walking for transport,[28] more trips with active transport modes,[28,29] and more cycling trips.[30] In contrast to the finding that higher levels of education are positively associated with active transport, higher income has been associated with fewer minutes of walking and less frequent trips with active modes.[28]

These studies were based on survey measures of physical activity and often used very distant approximations of transport-related physical activity such as number of trips with an active transport mode. To our knowledge, there are no studies investigating social inequalities in physical activity based on accelerometer data, and none investigating social inequalities in transport-related physical activity. This is due to the measurement issues which make large-scale data collections in this domain very difficult.

1.5 Challenges in measuring physical activity

1.5.1 Physical activity conceptualised

A commonly used definition of physical activity is '*any bodily movement produced by skeletal muscles that results in energy expenditure*'.[61] The four main components of physical activity are the type of activity, the intensity, the frequency and the duration.[62,63] The most common activity types in previous physical activity research in free-living settings were lying, sitting, standing and ambulating.[43] Given the context

of the study, other activities including household work, running, cycling, jumping or others were sometimes considered.[23,64] The activity intensity is often seen as a continuum from total inactivity to high intensity vigorous activity, with sedentary behaviour (SB), i.e. no or very low intensity; and moderate-to-vigorous physical activity (MVPA) as the most commonly studied activity intensity levels.[43] SB is regarded as a measure for sitting time; while MVPA is considered to be the intensity level of physical activity equivalent to brisk walking or more intense physical activity. Frequency and duration refer to the episodes in which an activity type or intensity level takes place. These two dimensions of physical activity are most often reported by calculating the accumulation per day (or per week) of the activity types or the activity intensity.[43]

1.5.2 Survey and accelerometer measures

Indicators based on surveys were for a long time the only instruments to measure physical activity, and they are still often used. Survey data are relatively easy and cheap to collect, but they are limited in accuracy and precision.[42,65,66] Accelerometers and other sensors were therefore very promising when they were introduced in physical activity research. In theory, accelerometers can measure accelerations of the body with a high precision and a high accuracy. More recent technical advances in accelerometers made it also possible to measure physical activity with great precision over long periods of time, at many different positions of the body (e.g. hip, wrist, ankle) and in all three spatial dimensions; as compared to uni-dimensional accelerometers.

Fairly low correlations have been found between surveys and accelerometer measures of activity intensity.[42,67] The estimated activity intensity such as the minutes of MVPA per day, is very dependent on the measurement method used. These finding have been used as arguments that survey-based indicators are not very reliable measures. However, the comparison between survey and accelerometer indicators is not a completely correct method to investigate the validity of survey indicators.

First, accelerometer indicators cannot be considered as criterion measures of activity intensity.[68] Accelerometer measures have some serious limitations. As described in section 1.5.3, there are several methodological choices to be made when using accelerometers to measure activity intensity, which can lead to very different estimates. To date, there is no agreement in the literature on the best way to measure activity intensity with accelerometers.

Second, accelerometer indicators measure something slightly different than survey-based indicators.[68] Activity intensity levels correspond to levels of energy expenditure expressed in Metabolic Equivalent Task (METs), with 1 MET being equal to the energy expenditure at rest. In theory, sedentary behaviour (SB) corresponds to a maximum energy expenditure level of 1.5 MET (while sitting), light physical activity (LPA) to a maximum level of 2.99 METs (while not sitting) and moderate-to-vigorous physical activity (MVPA) to 3 METs or more. MVPA is sometimes subdivided into moderate physical activity (3 - 5.99 METS), and vigorous physical activity ($\geq$ 6 METs).[69]

The survey indicator of SB is reported as minutes of sitting[70] while the accelerometer indicator is movement under a certain threshold, thus some standing can be included;[71] both indicators capturing only one part of the definition of SB. There are also different definitions between survey and accelerometer indicators of LPA and MVPA. For example in the short-version IPAQ (International Physical Activity Questionnaire), the participants report how intense the activities are for them, by responding to questions such as: 'During the last 7 days, on how many days did you do moderate physical activities'.[70] The measures therefore reflect the respondents' personal experience of activity intensity. In accelerometer studies, on the other hand, the acceleration data is transformed based on the findings of a previous calibration study (see section 1.5.3 for more details). Therefore, the measures do not reflect the participants' energy expenditure but the mean energy expenditure related to activity intensity of the participants in the calibration study. Accelerometer measures of activity intensity do not include the per-

sonal differences in physical fitness that lead to different levels of energy expenditure needed for a given activity intensity.

Other physical activity questionnaires use questions on the duration and type of activities (for example, 60 min of football) to calculate activity intensity per day. Also this type of questioning lead to very different estimates of activity intensity levels compared to accelerometer data, since it does not take the intermittent nature of activities into account.

In summary, the low correlations between survey and accelerometer measures are not only the result of the measurement errors in the survey indicators, but also of the measurement errors in the accelerometer indicators and the differences in definition between the two types of indicators. The lack of a true criterion measure of activity intensity in free-living conditions[72] makes it currently impossible to formally validate either type of indicators. Nevertheless, accelerometer indicators of activity intensity are widely considered to be far more precise and accurate than survey indicators.[42,72]

1.5.3 How to measure activity intensity with accelerometers

In transforming the acceleration data to activity intensity, several decisions in the data collection protocol and data processing can have a substantial impact on the results. Here, we present the decision-making process when using devices that allow for the collection of raw accelerometer data. Since older types of accelerometers did not enable raw data collection due to a lack of memory, the data were already transformed into counts within the device. Therefore, more of these decisions had to be made in the data collection protocol.

Decisions in the data collection protocol include the placement of the device on the body; the sampling rate i.e. the number of observations per second; the range of accelerations the device captures; and the number of axes (uni-axial or tri-axial) on

which the accelerometer data is to be registered by the device. These decisions are important in the choice of the device.

The data processing to obtain activity intensity estimates includes the following steps, each step subject to data processing decisions. The raw data is acceleration expressed in g-units, i.e. the acceleration caused by the force of gravitation; and needs to be filtered for non-human accelerations. Then, the filtered data is transformed into counts. The counts reflect how many times the acceleration level exceeded a threshold within an epoch – a predefined time unit. When a long period of zero values (e.g. 60 minutes)[73] is detected, this is considered non-wear time. Then, to determine the activity intensity within a time unit, the counts are compared to a cut point. For example, more than 1952 counts per minute (on the vertical axis) indicates MVPA, according to the cut point for adults calibrated by Freedson et al.[74] Finally, the accumulated time spent in a certain activity intensity is calculated per day (or another meaningful time unit).

Recently, tri-axial accelerometers have become the standard in physical activity research, measuring in all three orthogonal planes: vertical, antero-posterior and medio-lateral planes,[22] compared to the vertical plane used in the older uni-axial accelerometers; which may improve measurement of physical activity. The counts on the three axes are summarized into a vector magnitude count: the square root of the sum of the uni-axial counts squared. To obtain activity intensity measures, a vector magnitude cut point is then applied in the same manner as for uni-axial data. Table 4 gives an overview of the decisions discussed, as well as the choices made for this work.

New accelerometers and data processing methodologies are currently under development. The improvements in devices and data handling techniques will make it possible to avoid some of the above decisions in data protocol and data processing. For example, smaller accelerometers will make it easier to use accelerometers at multiple sites on the body for 24/7 wear time; and the use of raw data in combination with machine learning will change completely the use of counts, cut points and epochs by deriving

Table 1.1: Technical details of data collection and processing

Data collection	
Manufacturer	ActiGraph
Type	GT3X+
Wear location	Right hip
Sampling rate	30 Hz
Sample range	± 6 G
Sensitivity	3 mg/LSB
Data processing	
Data type downloaded from device	Raw data (in G-units)
Bandpass filter	Standard filter. Details not released by manufacturer
Epoch length	60 second epochs
Cut point activity intensity	Tri-axial MVPA [1]: > 2690 VM-CPM
Axis	Tri-axial
Data reduction	
Non-wear detection	1-min time intervals with consecutive zero counts for at least 90-min time window, allowing a short time intervals with nonzero counts lasting up to 2 minutes if no counts are detected during both the 30-min of upstream and downstream from that interval; any nonzero counts except the allowed short interval are considered as wearing. [5] Non-wear was detected using 60-s epoch counts; which were calculated with the standard bandwidth filter.
Valid days	Minimum 10 h wear time
Software	ActiLife 6 was used for applying the bandpass filters, calculating counts and non-wear detection. R 3.3.0 was used for the calculation of activity intensity levels per epoch and per day, and for the valid day detection.

Hz: Hertz; G: gravitational unit; mg: milli-G; LSB: least significant bit; LFE: low frequency extension; MVPA: moderate-to-vigorous physical activity; SB: sedentary behavior; LPA: light physical activity; CPM: counts per minute; VM-CPM: vector magnitude CPM; [1] MVPA cut point calibrated by Sasaki et al. (2011) for adults; [2] SB cut point calibrated by Aguilar et al. (2014) for older adults; [3] MVPA cut point calibrated by Freedson et al. (1998) for young adults; [4] SB cut point calibrated by Treuth et al (2004) for adolescent girls and confirmed by Matthews et al. (2008) for adults; [5] algorithm proposed by Choi et al. (2011)

physical activity indicators. To date, hip-worn devices and count based methodologies to process the data are by far the most common practices in the field.[43]

1.5.4 Challenges in accelerometer measures

Accelerometer indicators of activity intensity have limitations. The choices made during data collection can alter the results considerably. The filter for non-human acceleration (e.g. accelerations caused by a car), the counts-threshold, the cut point and the epoch length can all impact the activity intensity indicators. Until recently the memory capacities of the accelerometer devices did not allow the collection of raw data. The calculation of the counts was done within the accelerometer device, and there is therefore a limited amount of research on the impact of the filter[75] and the count-threshold on the physical activity indicators.

More is known about the wear-time detection algorithms and the cut points to transform counts into activity intensity indicators. Wear-time algorithms have been shown to have a modest impact on estimates of MVPA, but a relatively large impact on SB.[43,76] The choice of the 'best' cut points is probably the most debated topic in this field of processing accelerometer data;[77–79] and many cut points have been proposed for both MVPA and SB for children, adults and older adults. The impact of the cut point on these indicators is fairly intuitive. Since the counts accumulated within a 60-s epoch have to exceed a cut point value for that 60-s epoch to be recognised as MVPA time, the estimated time spent doing MVPA will be lower when the cut point is higher; and the inverse is true for SB. The calibration studies that defined the cut points had relatively small sample sizes and were done in laboratory settings; therefore, it is very hard to determine which cut point represents best the activity intensity for data collected in free-living conditions.

One important aspect in data processing that is often overlooked is the epoch length. Due to the limited memory space, older accelerometers did not enable data collections over long observation periods (typically 7 days) with epoch lengths much shorter than 60 seconds.[19] More recent devices provide the possibility of collecting data in shorter

epochs and the collection of raw accelerometer data, so that the epoch length can be as short as 1 second.

To measure the physical activity of children, there is a consensus on the use of shorter epochs to capture the more intermittent nature of the physical activity of children compared to adults.[80] Specific cut points for the physical activity of children have therefore been developed and calibrated on 15-s epoch data and studies have used epochs as small as 2 seconds.[80,81]

Studies on adult populations have also collected and used accelerometer data with shorter epoch lengths.[63] Only recently, however, a cut point for 15-s epochs has been developed, and this cut point is only applicable for adults 60 years and older.[82] Therefore, the cut points based on 60-s epochs have been adapted to shorter epochs by dividing the cut point by the corresponding factor. For example, the 30-s epoch cut point would be the 60-s epoch cut point divided by 2.

Adapted cut points may provide significantly different estimates of daily MVPA and SB. This has been shown mainly in studies on child populations,[17,80,81,83] and also in some studies based on adult populations.[18,19]

In a sample of overweight post-menopausal women, Gabriel et al.[18] found an extra 16 min of MVPA per day when adapting the Troiano[73] cut point to 10-s epochs compared to the original 60-s epoch cut point. Orme et al.[19] found similar results for a slightly younger sample of men and women using the same cut point, reporting an extra 16 min of MVPA per day on average between 5-s and 60-s epochs (equivalent to a decrease of 38.3%). Gabriel et al.[18] also found 1 hour and 39 minutes more SB per day when using the 10-s compared to 60-s epochs. Similar results were found in studies on children, with clearly higher estimates of MVPA and SB when using shorter epoch lengths.[17,80]

On a sample of children, one recent study[17] investigated the impact of epoch lengths on activity intensity estimates when applying the Romanzini cut points[84] for tri-axial

data. Vector magnitude counts in 15-s epochs were used for the validation of the Romanzini cut point.[84] The results showed a similar effect of the epoch length to those for uni-axial data for both MVPA and SB; MVPA and SB time estimates were higher when using shorter epoch lengths. To our knowledge, the impact of the epoch length on physical activity measures based on tri-axial accelerometer data originating from an adult population has not been studied.

1.5.5 Measuring transport-related physical activity

Numerous studies have relied on accelerometers to derive objective measures of physical activity.[31,42,43] However, studies were less successful in linking transport behaviour with physical activity. To measure physical activity that is specific to transport, it is important to have exact information about the respondents' mobility patterns: i.e. when, where and how were trips made. Collecting these patterns is a very demanding task for both researchers and respondents.[20,21] Data collections including both mobility patterns and accelerometer data therefore often result in datasets with very precise measures but with limited sample sizes.

Broadly, two types of solutions to this challenge in the collection of accelerometer-based physical activity in large-scale transport studies can be considered. The first solution is to facilitate the measurement of the mobility patterns, i.e. depart and arrival times of the trips and the transportation modes. The second solution is to avoid the collection of accelerometer data by predicting physical activity. These two solutions are discussed in more detail in the following two sections.

1.5.5.1 Automated detections of mobility patterns

Methods proposed to automatize the measurement of mobility patterns have focussed on determining the departure/arrival time and location,[32] the transportation mode,[23]

or a combination of both.[27,85] The departure/arrival time and location of trips can be determined by manually processing global positioning system (GPS) data in a geographical information system (GIS), i.e. geographical mapping software;[86] or by using an prediction algorithm.[32]

Most recent assessment methods for transportation modes are based on prediction models that use GPS data, accelerometer data, GIS data, or a combination of those three sources. Another method is the use of a body worn camera such as the SenseCam that regularly takes pictures.[87] The pictures are then coded by researchers.

For large data collections, the automatic algorithms for the recognition of both departure/arrival time and location and transportation mode detection diminish the workload for both researcher and respondent dramatically. However, the prediction accuracy of these methods is, as yet, relatively low. Hybrid methodologies have therefore been developed which combine automatic detection of mobility patterns with verification by survey.[20,21] These methodologies first use the algorithms to detect both departure/arrival time and location and transportation mode; and then verify this information during a mobility survey with the respondents, i.e. a so-called GPS based prompted recall survey. The mobility surveys can be held in real time with an automatic trigger system on smartphones; at the end of each observation day, usually an internet survey; or at the end the full observation period, usually a phone survey. The surveys done during the observation period avoid more memory bias but result in higher rates of participant drop out over the week compared to surveys held at the end of the observation period.[88,89]

These hybrid methodologies are still work intensive. For example, in our own RECORD GPS Study, the mobility survey was done at the end of the observation period. The collection of the 7-day mobility data took approximately a full day of work for a research assistant per respondent, including the preparation of the phone survey and data processing after the survey. This clearly affects the number of respondents that

can be included in a study with a given budget. However, the data resulting from these hybrid methods is of higher quality than survey data which results in memory biases; or data resulting from automatic detection methods, which results in prediction errors.

1.5.5.2 Predicting transport-related physical activity

A second complementary way to address the problem of collecting data related to transport-related physical activity, is to rely on a large mobility survey sample and then to estimate the physical activity level based on previously established knowledge. The compendium of Ainsworth et al.[90] enables this by providing an estimated physical activity level in MET per minute for numerous activities. The researcher has to determine which category of the compendium relates best to each trip, given the transportation mode, duration of the trip, and intensity of use of certain active modes.

The measure of transport-related physical activity expressed in METs has similar characteristics to an accelerometer based measure. It is a single indicator that takes the frequency and duration of the activity intensity into account. However, despite the usefulness of the compendium, the accuracy of its predictions can be criticised. The measures in the compendium are based on findings in very restricted settings, mostly laboratories,[90] and are not adaptable to the characteristics of trips in real settings, which likely vary between cities and countries. Therefore, they may not reflect free-living physical activity in a specific study context.

1.6 Objectives

Physical activity is an important protective factor for several health outcomes. Despite a longstanding interest of health researchers in physical activity, the progress of knowledge in the field is limited by, among other reasons, the difficulties in measur-

ing physical activity in free-living conditions. Accelerometers are increasingly used, which means a great improvement for the objective measurement of physical activity compared to surveys. However, there are still important challenges for the use of accelerometers in large-scale studies. While transport seems to be an important source of physical activity and provides unique opportunities to promote physical activity, accelerometer data about transport-related physical activity are particularly difficult to collect. This has restricted the research community's ability to reliably answer important research questions such as: 'what is the exact contribution of transport-related physical activity to the total physical activity?'.

In the light of these findings, this PhD project aimed to improve the estimations of transport-related physical activity for large-scale studies and to analyse the transport-related physical activity of adults living in the French capital region.

The specific methodological research aims were to:

- Investigate the bias introduced in accelerometer-based measurement of physical activity by applying cut points in combination with epoch lengths shorter than the epoch lengths used to calibrate the cut points.
- Develop a prediction method for the transport mode that enables a more cost- and time-efficient data collection method for transport-related physical activity.
- Develop a data integration method combining accelerometer datasets with survey dataset in order to provide accelerometer-based estimates of physical activity for large samples, with a better accuracy and precision than survey data.
- Develop a simulation technique that enables estimating the changes in transport-related physical activity due to transport mode shifts.

The specific empirical research aims were to:

- Analyse the population distribution of transport-related physical activity for adults living in the French capital region, with a specific focus on social inequalities.
- Simulate the impact of transport mode shifts on the population average of transport-related physical activity and the social inequalities in transport-related physical activity.

2. Methods

2.1 Study samples

This study used the data of two independent study samples: the RECORD GPS study sample and the EGT ('Enquête Global Transport') study sample. The two samples were considered complementary to investigate the transport-related physical activity for an adult population (between 35 and 83 years old) in the French capital region, Ile-de-France.

The RECORD GPS Study included highly detailed data on 7425 trips collected by accelerometers, GPS receivers and a mobility survey. However, this dataset included only 236 persons, which was considered too few for analyses at the person-level especially of population-wide social disparities. Moreover, due to the sampling procedure described below (see Section 2.1.1.1), this sample was not considered representative of the population.

For these reasons, we included the EGT dataset in this study. This dataset did not include accelerometer data, but it had a sample size large enough to analyse the data at a person-level (21332 persons and 82084 trips). The EGT sample was also representative of the background population. In sections 2.1.1 and 2.1.2, the data collections of the RECORD Cohort Study, the RECORD GPS Study (a subsample of the RECORD

Cohort Study) and the EGT study are described. In the section 2.1.3, both datasets are compared to the background population.

2.1.1 The RECORD GPS Study

2.1.1.1 The RECORD Cohort Study

The RECORD Cohort Study (Residential Environment and CORonary heart Disease), of which the RECORD GPS Study is a subsample, was established to investigate environmental determinants of territorial disparities in health.[91] The RECORD participants were recruited during preventive health check-ups in 2007-2008 and 2011-2013, were born between 1928 and 1978, and resided at baseline in 112 municipalities of the Ile-de-France Paris region.[92–94]

The recruitment took place during free preventive medical examinations in four centres of the IPC Medical Centre located in the Paris metropolitan area.[92,95,96] The medical examinations are offered every five years by the French National Health Insurance System for Salaried Workers (Caisse nationale de l'assurance maladie des travailleurs salariés, CNAMTS) to all working and retired employees and their families. People not insured by the CNAMTS could not be recruited for the RECORD Study: self-employed occupations (lawyers, architects, etc.), shopkeepers, craftsmen, farmers, and salaried farm workers. However, in the Ile-de-France region, working and retired employees and their families represent almost 95% of the population.[95]

No a priori sampling of individuals was performed in the general population as a basis for inviting potential participants to the healthcare centre. Participants were recruited among people visiting the healthcare centres for a reason independent of the study, i.e., a convenience sample was established. The employed, unemployed, or retired workers or their families visiting the healthcare centres for a preventive check-up either came

on their own, or were sent by their family physician or work physician, or were referred to the centre by various associations. A priori, 10 (out of 20) administrative districts of Paris and 111 other municipalities of the Paris Ile-de-France region were selected for the study. The selection favoured districts and municipalities of which it was expected that relatively many inhabitants would visit one of the four sites of the IPC medical centre during the recruitment period. The selection also ensured the inclusion of areas with different socio-economic backgrounds and from urban and peri-urban areas.

2.1.1.2 The RECORD GPS Study

In the second wave of the study (2011-2012),[91,93] after undergoing a medical check-up and filling in computerised questionnaires at the IPC Medical Centre, 410 participants were invited to enter the RECORD GPS Study of which 247 subjects agreed to participate. The participants wore a GT3X+ accelerometer (ActiGraph) and a BT-Q1000XT GPS (QStarz) on the right hip with a dedicated elastic belt, at the recruitment day and for 7 additional days, all day long from wake up to bedtime. The participants had to fill out a travel diary by reporting their activity places over the 7 to 8 days, each time with arrival and departure times. Written informed consent was obtained from all participants. The RECORD GPS Study was approved by the French Data Protection Authority.

The mobility data were collected with a prompted recall mobility survey, i.e. a survey using GPS data uploaded in the Mobility Web Mapping application and a travel diary to prompt the memory of the participant. The GPS data were collected with the BT-Q1000XT GPS device every 5 seconds. In certain circumstances (e.g. the participant was indoor) the GPS data would be missing or invalid. After removal of invalid data and linear interpolation of the missing data, the GPS data were analysed with an algorithm (ArcGIS Python script) that identified all of the activity locations of the participants (any activity at a stationary location) with their start and end time from

the accumulation of GPS points over 7 days.[32] Based on these outputs of the algorithm, the Mobility Web Mapping application was then used to visualise the activity locations on a map per participant per day. The application was used to survey the participants on the activity performed at each visited location and on the modes used in each trip. The survey operator could report activity locations and trips undetected by the algorithm and could modify / remove detected visits to locations that were inaccurate or incorrect. This procedure resulted in the observation of 7425 trips for 236 participants over 1584 observation days.

The physical activity data was collected by the ActiGraph accelerometer worn at the hip, which is the most commonly used device and the most commonly used position on the body.[97] The GT3X+ version of the ActiGraph was chosen for this study because it enables collecting raw tri-axial accelerometer data over relatively long periods of time. Also, due to the common use of the Actigraph, the procedures to transform the data into physical activity indicators have been relatively well-documented and formalised compared to other devices.

2.1.2 The EGT study

The Global Transport Survey (Enquête Global Transport, EGT) is a household travel survey conducted every 10 years in Ile-de-France, the French capital region. The latest EGT survey was conducted in 2010 by two French transport institutions: the Ile-de-France Transport Authority (Syndicat des Transports d'Île-de-France, STIF) and the Regional and Interdepartmental Direction for Equipment and Planning (Direction Régionale et Interdépartementale de l'Equipement et de l'aménagement, DRIEA). The main purpose of the survey is to inform local authorities and transport planners on the mobility and transport use in Ile-de-France. During face-to-face interviews with members of randomly selected households, data were collected for all the trips made during the day before the interview. For this study, we selected participants between

35 and 83 years old to match the population targeted by the RECORD Study. This resulted in a dataset including 82084 trips made by 21332 people.

2.1.3 Comparison of the RECORD GPS sample, the EGT sample, and the background population

In the studies presented here, we do not claim the RECORD data to be representative of the background population at the participant-level or the trip-level. In fact, this lack of representativeness is the main motivation to integrate the RECORD GPS data with the EGT data. However, the trips were observed in real-life settings and can be assumed to include a wide variety of the trips performed by the background population (people between 35 and 83 years old residing in Ile-de-France). This property made it possible to use these data to illustrate the methodologies presented in Article 3.1 and Article 3.2 and to inform the data integration used in Article 3.3 and Article 3.4.

Table 2.1 compares the RECORD GPS sample, the EGT sample and census data obtained from the French National Institute of Statistics and Economic Studies (INSEE, http://www.insee.fr/) on the background population, i.e. people between 35 and 83 years old living in Ile-de-France. This comparison supports the hypothesis that the EGT sample represents the background population better than the RECORD GPS sample. The EGT sample included more women, more young people and less people from the inner city compared to the RECORD GPS sample.

2.1.4 Inclusion criteria

In the article on the impact of the epoch length (Article 3.1), the accelerometer data was analysed at the day-level. The data collection failed for four participants and five participants wore an older version of the ActiGraph (ActiGraph GT3X vs ActiGraph GT3X+), which did not collect raw accelerometer data. Furthermore, an observation

Table 2.1: Overview of the demographic characteristics of the background population (people between 35 and 83 years old in Ile-de-France), the EGT sample and the RECORD GPS sample

	I-d-F [a] (%)	EGT [b] (%)	RECORD (%)
Gender			
Female	52	53	37
Male	48	47	63
Age [c]			
35-44 years	30	32	16
45-59 years	39	37	36
60-74 years	24	25	41
75-83 years	7	6	7
Location of residence			
Inner city (Paris)	19	14	26
First crown of counties around Paris	37	36	41
Second crown of counties around Paris	44	51	30
Population / sample size	5,887,647	21,332	236

[a] I-d-F: 2012 Census data from Ile-de-France, the French capital region; [b] EGT: Enquete globale transport; [c] The data for the age groups 35-44 and 75-83 were not available in the population statistics. The percentages for these categories are based on the assumption that the distribution within the broader category is uniform; [d] The categorization of urbanicity is based on an official administrative subdivision of the Ile-de-France region.

day had to include a minimum of 10 h to be considered valid. By retaining only valid observation days, another three participants were deleted from the dataset. The final dataset for analysis included 1389 observation days for 224 participants. Thus overall, the data from 12 participants could not be used.

In the three other articles, we used the accelerometer data at the trip-level. For the article on the prediction of transportation mode (Article 3.2), we had to exclude 96 trips performed by a combination of non-walking transportation modes due to a low prevalence of trips with these combinations (e.g. car and bike). This resulted in a dataset with 7329 trips for 236 participants. For the article on data integration and social inequalities (Article 3.3) and the article on the simulations of transportation modes (Article 3.4), the RECORD GPS dataset was used for the prediction model,

not the analysis model. Therefore, it was of no use to include the trips with missing accelerometer data, since these trips would not have added any information to the prediction model. Also the previously mentioned 96 multimode trips were deleted. This resulted in a dataset with 7138 trips for 229 participants.

In the EGT study, only the participants aged between 35 and 83 years old were included, to correspond to the RECORD GPS participants. No further selection was done, not at the participant level nor at the trip level. This means that the multimode trips (i.e. trips with two or more non-walking modes) were not excluded from the EGT dataset, in contrast to the inclusion criteria for the RECORD GPS trips. Excluding these trips would have implied a strong bias in the estimated accumulated T-MVPA per day. With an average of 4 trips per person per day, not recording the physical activity for even only one trip would have resulted in a considerable underestimation of the individual's daily T-MVPA. Instead, for these trips, we used the 'main' transportation mode, i.e. the transportation mode with the longest duration (information that was not available in the RECORD dataset). The lack of a prediction model based on multimodal trips has probably introduced bias in the estimation of T-MVPA for these trips; however, this bias was considered to be considerably less than the underestimation resulting from excluding the trips.

2.2 Measurements

In this section, the large set of variables used throughout the four studies are presented. First, the construction of the physical activity indicators is described, with special attention to the indicator used in all four studies: moderate-to-vigorous physical activity (MVPA). Then, two sets of predictors of the transportation mode are described: GPS- and accelerometer-based predictors. In the following subsections, the trip characteristics, person characteristics and environmental characteristics are presented. The GPS

and accelerometer based variables were only available for the RECORD sample. However, in Section 2.3.1, we describe how the MVPA indicator was made available for the EGT sample by data integration.

2.2.1 Physical activity indicators

Among the four dimensions of physical activity (activity type, activity intensity, frequency and duration), this PhD work focussed on activity intensity. Activity intensity is often seen as a continuum from total inactivity to high intensity vigorous activity. IN the article on the impact of the epoch length, the continuum was divided in three categories: moderate-to-vigorous physical activity (MVPA), light physical activity (LPA) and sedentary behaviour (SB). In the other three articles, only the most commonly used indicator of physical activity, MVPA,[43] was used.

Raw accelerometer data represents the acceleration of a participant's body, and not the intensity of physical activity (e.g. MVPA) nor the bodily position (e.g. sitting). Raw accelerometer data is transformed into indicators of physical activity in three steps. Depending on the device and the software used to transform the data, the details of the procedures might be slightly different. In general, the acceleration signal is first filtered for accelerations caused by non-human activity. Then, the remaining signal is transformed into counts, which expresses the number of times the acceleration measured in a sampling rate between 30 Hz and 100 Hz surpasses a certain limit during an epoch (a predefined time unit). Finally, these counts are transformed into indicators of MVPA, LPA and SB by using cut points for the counts per minute (CPM).

Cut points have been calibrated in laboratory research for specific accelerometer devices and specific positions on the body.[22,73,74,98] In these studies, MVPA was defined as a minimum energy expenditure of 3 metabolic equivalent of task (MET) or 3 times the energy expenditure at rest. The criterion for energy expenditure was indirect

calorimetry, a method to measure the oxygen uptake. The criterion for SB was direct observation. LPA was defined as the physical activity with an intensity level lower than MVPA but higher than SB.

For the studies in this PhD work, the MVPA cut point of Sasaki[22] for tri-axial accelerometer data was adapted. Using this cut point, a minute of MVPA is defined as a minute during which 2690 or more CPM were recorded on the vector magnitude. The vector magnitude summarizes the movement registered on all three axes (vertical, antero-posterior and medio-lateral axis) by taking the square root of the sum of the squared CPM of each axis. For Article 3.1, also the cut point of Freedson[74] for MVPA (1952 or more CPM on the vertical axis) was used and compared to the Sasaki cut point. The Freedson cut point is the oldest and probably most used cut point in the literature.

People cannot be expected to wear the accelerometer device at the hip 24 h per day. Therefore, an algorithm determines when the sensor was not worn, so this part of the data can be left out of the analyses. In this study, we used the algorithm proposed by Choi et al.[99], which defines a non-wear time as periods of 90 minutes of consecutive zero's, allowing a short time intervals with nonzero counts lasting up to 2 minutes if no counts are detected during both the 30 min before and 30 min after that interval. In the calculation of the accumulated activity intensity per day, the assumption is made that the non-wear time is mostly due to inactive periods during the day (e.g. sleep time). However, when there is too much non-wear time detected during a day, it is likely that the device malfunctioned or that the respondent did not wear the device during active periods. Following previous research,[100–103] an observation day had to include at least 10 h of wear time to be considered valid for the day-level analyses in the article on the impact of the epoch length (Article 3.1). There was no minimal number of days required to retain a participant in the analysis. For the other three articles, MVPA was analysed at the trip-level. The observation of MVPA for a trip

was considered valid if the full period of the trip was recognised as wear time.

Accelerometers worn at the hip underestimate physical activity during biking trips.[104] For the articles on the epoch length (3.1) and on the prediction of the transportation mode (3.2), it was important to report and use the accelerometer data as observed. However, for the articles on the social inequalities 3.3 and on the simulated transport mode shifts 3.4, it was important to report accurate estimates of T-MVPA. Therefore, we used an estimate of biking physical activity from the compendium of Ainsworth.[90] A drawback of this is that the variability inside the trips was lost. So, all minutes of biking trips were considered to be physically active disregarding the stops over the way. The impact on the results was probably small with around 6.2% of T-MVPA obtained from cycling in this population. A slight overestimation of this small share of T-MVPA probably only led to a minor overestimation of the daily T-MVPA.

2.2.2 Transportation mode

During the EGT and the RECORD surveys, participants reported a chronological sequence of transportation modes for each trip. For modelling purposes, this information was coded into a transportation mode variable consisting of four categories: 'walking' (i.e., only walking), 'bicycle', 'private motorised', and 'public transport'. When both walking and another transportation mode were sequentially used within a trip, the non-walking mode was attributed to the trip. The walking part during these trips was considered to be the consequence of the choice of the non-walking mode.

2.2.3 Individual characteristics

The following self-reported individual characteristics were used in the articles on social inequalities (Article 3.3) and the article on the simulation of transport mode shifts (Article 3.4). The household income was coded as a continuous variable. The educational

level was measured in three categories: 'no diploma of secondary education', 'diploma of secondary education or lower tertiary education', and 'diploma of higher tertiary education'. Working situation was categorised as 'employed', 'unemployed', 'retired', or 'other'. Participants indicated whether a bike, a motorbike, and / or a car was available in their household; and whether they had a public transport pass. Finally, age (continuous) and gender (male / female) were included in the analyses.

2.2.4 Environmental characteristics

For the departure and the arrival points of each trip and for the residence, environment characteristics were generated within a Geographical Information System (GIS). Two types of characteristics were used: the shortest street distance to a point of interest (e.g. closest bus stop) or the density of something (e.g. density of green spaces) or people (e.g. density of highly educated people) within a predefined area. The residential area and the area around the departure / arrival point were defined as 1 km buffers around the location following the street network; corresponding to a 10-to-15 minute walk that reflects the local resources easily accessible within a 'walkable' distance.[93,94,105,106]

Based on the 2010 population census of the French National Institute of Statistics and Economic Studies (INSEE), the educational level of an area was measured as the proportion of persons in that area with a diploma of tertiary education. From the same dataset, the population density was extracted. The density of destinations (supermarkets, other shops, administrations, public/private shops, health services, entertainment facilities) was based on the 2011 permanent database of facilities of INSEE. The number of intersections within an area was extracted from a database of the French National Geographic Institute (Institut National de l'information Géographique et forestière, IGN) collected in 2014. Data from the Institute of Urban Planning of region Ile-de-France (Institut d'Aménagement et d'Urbanisme Ile-de-France, IAU-IDF) on public parks and green spaces in 2008 enabled extracting the proportion of area's surface

covered with green space.

Based on a dataset from the Ile-de-France Transport Authority (Syndicat des transports d'Île-de-France, STIF), the distance to the nearest transport service following the street network was extracted for bus, metro, train and tram stops. A fifth variable was the distance to the nearest transport service, i.e. the minimal value of these four distances. Finally, from a IGN database, the administrative location of the residence or departure/arrival point of a trip was determined. These locations could be in one of the three following areas: the inner city of Paris, the first crown of counties around Paris (i.e. the counties adjacent to the city centre) or the second crown of counties around Paris (i.e. the counties non-adjacent to the city centre). This categorization is based on a recognized and official administrative subdivision of the Ile-de-France region, previously used in studies in the Ile-de-France region.[107–109]

2.2.5 Other predictors of transportation modes

For the article on the prediction of transportation modes (Article 3.2), additional accelerometer-based predictors were used. There is (nearly) no influence of multicollinearity on the predictive value of random forests prediction models; and there was no analytical aim for this model. Therefore, many relevant predictors based on the GPS, accelerometer and geographical information systems (GIS) data could be constructed, even if they had no real clinical value.

In addition to the standard filter for human acceleration, a low-frequency extension filter was used.[110] The optional low-frequency extension filter extends the lower end of the filter, which might be useful for example when processing the data of people who move slowly. For both filtering approaches, we estimated the number of footsteps taken, the energy expenditure in kilocalories calculated from the activity counts and participant weight based on the Sasaki and Freedson equation,[22] and whether the

participant was sedentary[111] at the 5 sec epoch-level. To capture a maximum of relevant information, we derived for each trip standard measures of central tendency such as the mean and median, and measures of dispersion such as standard deviation, minimum, maximum, 10th and 90th percentiles. On the basis of the accelerometer data, the accelerations on each of the three axes separately, the number of steps taken, MVPA, sedentary time, and energy expenditure in kilocalories were aggregated in this way. In addition, we calculated the total number of steps taken, the number of MVPA epochs, the number of sedentary epochs, and total energy expenditure for each trip. We also determined the percentage of epochs that were SB or MVPA.

Next to accelerometer based predictors, also GPS based predictors were used for the article on the prediction of transportation modes (Article 3.2). Every 5 seconds, the GPS device registered the position coordinates (i.e., latitude, longitude, and elevation), speed, and the following three indicators of the quality of the observation: horizontal, vertical, and positional dilution of precision (HDOP, VDOP, and PDOP, respectively). Only the good-quality observations (HDOP < 6, VDOP < 7, PDOP < 8) were retained[21] for the aggregation of time-unit observations at the trip-level. GPS observations were determined to be valid, invalid (high dilution of precision), or missing (less than three satellites in view). On average, 27% of GPS observations were missing and 1.5% of the existing observations were invalid. The distribution of potential GPS data points across these three categories provides information on the circumstances of the trips (e.g., underground public transport, tunnels, high buildings). To capture this trip characteristic, the total number of GPS observations, number of valid GPS observations, percentage of valid GPS observations among recorded observations, and percentage of valid GPS observations relative to the maximum number of observations (including missing ones) were also used as predictors for transportation modes.

On the basis of the GPS data and geographical information on the street network provided by the National Geographic Institute, four distance measures between the

departure and arrival points of a trip were calculated: the straight line distance, the shortest walking distance following the street network, the shortest street network distance by car, and the map-matched distance. The latter distance is determined by projecting the GPS data points onto the street network.[112] Finally, the duration of the trips was obtained from the mobility survey.

2.3 Statistical analysis

First, two methodologies are presented that were developed during this doctoral work: the integration of the EGT and RECORD datasets and the procedure to simulate the impact of transport mode shifts. Then, the random forest prediction method is described. This prediction method is the basis of the data integration (Article 3.3), the simulation procedure (Article 3.4) and the prediction of transportation modes (Article 3.2). Finally, the negative binomial regression method and multiple imputation method are described. These methods were used to analyse the integrated MVPA measure in the EGT dataset.

2.3.1 Data integration

The data integration method presented in this work relied on datasets from the same population that have a large set of variables in common. These two characteristics enabled the use of the best aspects of both datasets: the highly detailed accelerometer data from the RECORD GPS dataset and the large sample size of the EGT dataset.

The goal of the data integration was to obtain an accelerometer-based measure of transport-related MVPA (T-MVPA) that would enable person-level analyses. In the article on the data integration and social inequalities (Article 3.3) and the article on the simulation of transportation modes (Article 3.4), we presented and applied a method

based on a random forest prediction model and a set of variables common to both datasets. In a first step, a prediction model for trip-level MVPA was built on the RECORD GPS data. Then, predictions of MVPA were made for each trip in the EGT dataset where the same predictor variables were available for each trip. Finally, the estimated MVPA per trip was summed up per day for each EGT participant. This resulted in an estimate for T-MVPA per day which was then correlated with individual variables, e.g. educational level.

2.3.2 Simulation procedure

In the article on the simulation of transportation mode shifts (Article 3.4), the aim was to evaluate the impact of shifts in transportation modes by simulation. Twelve simulation scenarios were considered, 3 for each of the 4 transportation modes: walking, biking, public transport, and private motorised transport. All scenarios were designed to promote more active transportation modes. So, in the private motorised scenarios, private motorised trips were changed into walking, biking, or public transport trips. For the other three modes, trips not performed by the respective mode were changed into this mode. For example, in the walking scenarios, non-walking trips were changed into walking trips.

The simulation process for all scenarios consisted of three consecutive steps. In a first step, the transportation mode for a predefined proportion of trips was changed into an alternative mode. The predefined proportion of trips was chosen in function of the prevalence of the mode under consideration. In the 'private motorised' scenarios, the percentages of private motorised trips changed into walking, biking, or public transport trips were of 10%, 20%, and 30%. In the walking scenarios, the number of non-walking trips changed into walking trips was of 10%, 30%, and 50% of the observed walking trips. The same percentages were applied in the public transport scenarios. In the biking scenarios, the percentages applied to the non-biking trips were 100%, 200%,

and 300% of the observed biking trips.

Any trip of interest for the scenario could be selected for a transport mode shift (for example, every non-walking trip in the walking scenarios), but the selection was weighted by the likelihood of a trip to be performed by the alternative mode. For the 'private motorised' scenarios, the alternative mode was the most likely alternative transportation mode for the respective trip. For the other scenarios (a change to walking, biking, or public transport), trips were selected based on their likelihood to be performed by the target mode. Taking as an example the scenarios to promote walking, non-walking trips were selected for change based on the likelihood that these trips were performed by walking. In this example, the likelihood of performing these non-walking trips by walking was extracted from a random forest model predicting the transportation mode. The average predicted probability of walking in these non-walking trips was rescaled to the pre-specified level of change in the scenario of interest (e.g., to 10%, 30%, or 50%). The rescaling relied on a transformation of the probabilities to the logit scale and then back to the probability scale to avoid probabilities outside of the [0–1] range. These transformed probabilities enabled drawing random samples of the trips selected for change, weighted by the likelihood for the trip to be performed by the alternative mode given the predictor variables.

In a second step, the duration of the trip was predicted for the trips to which a new transportation mode was attributed in step 1. The prediction was based on a random forest model for the duration of trips in the EGT data. In a final step, the MVPA was predicted for the trips with a changed transportation mode and duration with the same model than the one used for the data integration, i.e., a random forest model for MVPA based on the RECORD GPS data. The simulation of each scenario was repeated 100 times to avoid random sampling error in the results.

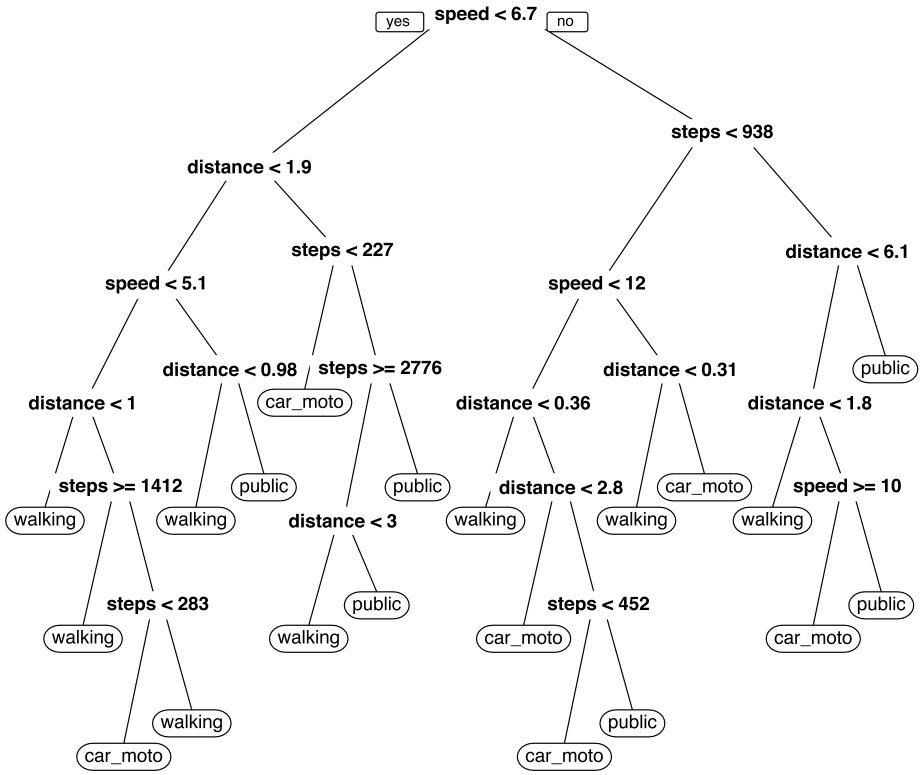
2.3.3 Random forests

The random forests prediction method[113] is based on the decision tree method. Decision trees classify data into groups in subsequent steps, aiming to obtain homogenous groups in terms of the outcome variable.

At each step in the tree (or knot), the algorithm first determines the features, i.e. the dichotomous versions of the variables that best differentiate the sample. For continuous variables, the algorithm searches for the best cut point for that variable; for categorical predictors, the categories are regrouped into two categories. Out of the set of features (one for each variable), the algorithm selects the feature that best differentiates the (sub-)sample. Figure 2.1 provides a simplified example of such a decision tree, with the transportation mode as the outcome variable and the mean speed ($km \cdot h^{-1}$), distance (km), duration (min) and total number of steps as the predictor variables. At the first knot, the sample is divided in two subsamples: the trips with a mean speed below 6.7 $km \cdot h^{-1}$ to the left, the others to the right. The two new subsamples are sequentially subdivided until they are homogeneous or until the algorithm is not able to further differentiate the trips within in the subsample. For illustrative purposes, the algorithm for this particular decision tree also stopped if the subsample contained less than 30 observations.

Simple decision trees, like the one presented above, are very performant in predicting the outcome variable of the dataset at hand, but they often provide predictions of poor quality for new datasets. This is a problem also known as overfitting, where prediction models take too much the particularities of the dataset into account to be performant on new datasets. To obtain better generalisability, the random forests method adds two sources of randomness to the simple decision tree method and repeats the process a large number of times, thereby resulting into a forest of decision trees. The first source of randomness consists of considering only a random subsample of the features in each

Figure 2.1: Example decision tree - predicting transportation mode



Notes: 'walking' = Walking trips; 'Car_moto' = Private motorised trips; 'public' = Public transport trips; 'Speed' = Average speed during the trip in km/h; 'distance' = Distance of the trip in km; 'steps' = Number of steps taken during the trip.

knot. Secondly, for each tree, only a random subsample of the observations is used. The subsample is obtained by selecting as many observations as in the original sample with replacement, i.e. the same observation can be selected multiple times. As a result, the subsample contains approximately 64% of the observations once or multiple times, and does not contain about 36% of the observations (out-of-bag observations).

Like for decision trees, random forest can be used for both the prediction of categorical variables (classification) and continuous variables (regression).[114] The procedure is almost identical for both, only the criterion of diversity (or impurity) and the way predictions are calculated are different. For the classification random forests, the Gini

impurity is calculated to compare the diversity within the subsamples. The Gini impurity expresses the probability of a wrong prediction, if for any of the observations the category was predicted randomly given the frequencies of the categories in the subsample. The Gini impurity is equal to 0 if all observation in the subsample belong to the same category. Regression random forests uses the percentage of variance explained (R^2) as the indicator of diversity. This is calculated by subtracting from 1 the ratio between the mean square error in the predictions after the split and the variance of the outcome variable before the split. So, the algorithm will select for each knot the feature that reduces the most the Gini coefficient or the R^2.

Finally, the random forests are validated on their prediction value. Classification forests are validated by the prediction error rate, or the proportion of wrong predictions. Regressions trees are validated by the R^2 comparing the final subsamples with the original sample. To calculate the prediction value, predictions for each observation in the dataset are calculated. First, tree predictions are obtained for the observations that are not used to grow the respective tree (so-called out-of-bag data). Then, for the classification random forest, a forest prediction is obtained for each observation as the majority of the out-of-bag tree predictions. The regression random forest method uses the averages of the out-of-bag predictions as the final forest predictions.

In most other prediction methods – such as regression models or decision tree models – the models have to be validated by a training and test set procedure (or similar, e.g. cross-validation). In a training-test set procedure, the prediction model is built on a training set and evaluated on the test set. The procedure prevents overfitting models; on the other hand, it uses only a subset of the data to inform the prediction model. The random forest method uses the tree predictions of the out-of-bag data to validate the model, instead of a separate test set. This works as a built-in validation mechanism. Therefore the evaluation criterion (either prediction error rate or R^2) is a good estimation of the prediction value for new data; and the full dataset can be used

to build the prediction model.

In response to a reviewer, we verified the value of the out-of-bag validation method for our prediction model of the MVPA variable used for the data integration in Article 3.3. We made the comparison between the built-in coefficient for accuracy and a coefficient for accuracy based on the training / test set validation. The test was designed as follows. We divided the dataset into a training set (75%) and a test set (25%). A random forest with 1000 trees was grown on the training set. The R^2 reported by the model was 0.60, the R^2 in the test set was 0.63. To see if this was due to bias or random sampling variation, we repeated the process 100 times (i.e. 100 different divisions into training and test sets; for each training set we have grown a forest of 150 trees, and each time we calculated the R^2 for both the training and test set). The mean R^2 in the training sets was 0.64 and the mean R^2 in the test sets was 0.65. So, the built-in validation mechanism resulted in a good approximation of the classical training and test set validation, by slightly underestimating the R^2. It should be noted that the R^2 reported in the article on data integration (Article 3.3) from the built-in validation procedure (0.67) is higher, since the full 100% of the data was used to build the prediction model. We can therefore conclude that the model outperforms a model that would be based on a training and test set validation procedure, as indicated by the higher R^2.

2.3.4 Negative binomial regression

The negative binomial regression was used in the article on data integration and social inequalities (Article 3.3) to analyse the minutes of T-MVPA accumulated per day. The time variable could be considered as continuous and analysed with a regular linear regression. However, given the left-censored distribution of the variable (i.e. 0 as the absolute minimum and many observations equal 0 or close to 0), we preferred the negative binomial regression that is adapted to count variables with overdispersion

(a high variance compared to the mean).

2.3.5 Multiple imputation

In the analyses of the minutes of T-MVPA accumulated per day, there were missing values on 8 independent variables for 24 % of the respondents, of which 6 % had more than 1 missing value. Therefore, multiple imputations were performed.[115,116] This method enabled analysing the data under the hypothesis that the unobserved values are randomly distributed given the observed data.[115] To account for the non-linear and interaction effects in the imputation process, random forests methods were also used for the multiple imputations.[117] Five imputation datasets were constructed though an iterative process using 100 trees for every imputed variable at each iteration. The convergence of the imputations was checked with plots of the means and standard deviations over the iterations.

2.4 Data management and data analyses software

To handle the accelerometer data, we used the program provided by the ActiGraph company: ActiLife (version 5.1). The geographical information system ArcGIS (version 10.3, automated by Python version 2.7) was used to calculate the geographical variables. All other data management and all data analyses were performed in R (several versions between version 2.7.0 and version 3.3.0). The random forests were grown using the R package 'randomForest'.[114] The negative binomial regression analysis was performed using the R package 'MASS',[118] and the multiple imputations were performed with the 'mice' package in R.[119]

3. Results

3.1 Article 1: The impact of epoch lengths on moderate-to-vigorous physical activity, light physical activity and sedentary behavior estimates for adults

Submitted October 15, 2016

The impact of epoch lengths on moderate-to-vigorous physical activity and sedentary time estimations for adults

Ruben Brondeel [1,2,3*], Jasper Schipperijn [4], Paul Kelly [5], Jacqueline Kerr [6], Yan Kestens [7,8], Basile Chaix [1,2]

Abstract

Background: Many decisions in accelerometer data collection and processing can impact physical activity indicators such as activity intensity levels. It is seldom highlighted that cut points are determined at a specific epoch length, and that shorter (or longer) epoch lengths result in very different estimates of activity intensity. In response to these knowledge gaps, using free-living data, this study will compare the impact of epoch length on estimates of moderate-to-vigorous physical activity (MVPA), light physical activity (LPA), and sedentary behavior (SB). It will also compare the impact of epoch length when using tri-axial cut-points and uni-axial cut-points. **Methods:** The RECORD GPS study collected real-life accelerometer from 227 participants aged between 35 and 83 years, living in the French capital region, Ile-de-France. The respondents were asked to wear an accelerometer (ActiGraph GT3X+) and a GPS device (QStarz) at the right hip for 7 d. Daily activity intensity levels were estimated using four epoch lengths: 1-second, 15-second, 30-second and 60-second. Both vector magnitude and vertical axis cut-points were used to estimate MVPA, LPA and SB. Three types of locations (transport, work, leisure) were detected using GPS and survey data. **Results:** Shorter epoch lengths resulted in considerably higher estimates for MVPA and SB; and consequently in lower estimates for LPA. The impact of the epoch length was higher for vector magnitude compared to vertical axis indicators; and it was lower for time spent in transport compared to time spent at the workplace and other places. The impact of epoch length on activity intensity estimates is in most cases larger than the impact of other common data processing decisions (e.g. cut-point). **Conclusions:** This study investigates the often overlooked impact of epoch length on activity intensity estimates. The findings support the call for further standardization of the data processing decisions in physical activity research. The authors advice to use the cut-points in combination of the data processing decisions, such as epoch length, as used for the respective calibration studies.

[1] *Inserm, UMR-S 1136, Pierre Louis Institute of Epidemiology and Public Health, Nemesis team, Paris, France*
[2] *Sorbonne Universités, UPMC Univ Paris 06, UMR-S 1136, Pierre Louis Institute of Epidemiology and Public Health, Nemesis team, Paris, France*
[3] *EHESP School of Public Health, Rennes, France*
[4] *Department of Sport Sciences and Clinical Biomechanics, University of Southern, Denmark*
[5] *Physical Activity for Health Research Centre (PAHRC) University of Edinburgh, UK*
[6] *Department of Family Medicine & Public Health, University of San Diego California*
[7] *Département de médecine sociale et préventive, Université de Montréal, Montreal, QC, Canada*
[8] *Centre de Recherche du Centre Hospitalier Universitaire de Montréal (CRCHUM), Montreal, QC, Canada*
***Corresponding author**: Ruben Brondeel, ruben.brondeel@iplesp.upmc.fr, +33 1 44 73 86 54, Address: Faculté de Médecine Saint-Antoine, 27 rue Chaligny, UMR-S 1136, 75012 Paris, France

1. Introduction

Accelerometer data are being used increasingly in physical activity research [1] and recently, also in large epidemiological studies with strong data on health outcomes. [2] Linking accelerometer data to health outcomes is a very important development for the field since this can inform future physical activity guidelines. [3] Accelerometer data are considered 'objective' and have some important advantages compared to self-report data. However, there are several decisions in the data collection and processing of accelerometer studies that can affect the estimates of physical activity indicators.

Several studies have shown the importance of the following decisions in accelerometer data collection protocols: the model of the device, the location of the device (e.g. waist), the sampling rate of the observations (between 30 and 100 hz), and the sample range. [4,5] Also in the data processing phase many decisions can affect the physical activity indicators. In the following paragraphs, these decisions are discussed for count-based measures of activity intensity, such as moderate-to-vigorous physical activity (MVPA), based on a waist-worn ActiGraph accelerometer.

Raw accelerometer data are processed into indicators of activity intensity in three steps. Depending on the device and the software used to transform the data, the details of the procedures might be slightly different.

- Step 1: In general, the acceleration signal is first filtered for accelerations outside the normal human activity frequency bandwidth. For older populations, a low frequency extension (LFE) filter has been proposed, to better capture lower intensity activities that are important for this population. LFE-filters should also be used when applying cut-points calibrated with older versions of the ActiGraph, since these older devices were more sensitive than the more recent versions.[4]
- Step 2: The remaining signal is transformed into counts, which express the number of times the acceleration surpasses as certain threshold during an epoch (a predefined time unit). This threshold is undisclosed for the ActiGraph devices due to commercial considerations.
- Finally, these counts are transformed into indicators of activity intensity by using cut-points for the counts per epoch. Cut-points have been calibrated in laboratory research for specific accelerometer devices and specific placement on the body. For example, Freedson's cut-point for MVPA is 1952 counts on the vertical axis per 60-second epoch measured by a CSA device (earlier version of the ActiGraph) worn at the waist.[6]

The final step in the data processing is data reduction, i.e. removing data from episodes during which the participants did not wear the device (non-wear time); and subsequently, removing data from days for which not enough wear time is detected (non-valid days). The non-wear time of the device is detected by prolonged periods with zero counts; for example, 60 consecutive min interrupted by maximum 2 min with counts between 1 and 100.[7] A valid observation day is defined by a minimum number of hours of wear time; for example, 10 h.[8]

All these decisions can have an impact on the activity intensity indicators, with differences due to different cut-points being the most frequently investigated issue.[9,10] A less studied data processing decision is epoch length. The early accelerometer studies and cut point calibration studies were focused on 60-second epochs. However, with new memory capacity, and new versions of ActiLife, came the ability to record and process data in shorter epochs.[11]

To measure the physical activity of children, there is a consensus on the use of shorter epochs to capture the more intermittent nature of the physical activity of children compared to adults.[12] Specific cut-points for the physical activity of children have therefore been developed and calibrated on 15-second epoch data and studies have used epochs as small as 2 seconds.[4,12]

Studies on adult populations have also collected and used accelerometer data with shorter epoch lengths.[13] There are currently no validated cut-points available for data with epochs shorter than 60 seconds for adults, except for the 15-second epoch cut-point for older adults, recently proposed by Evenson et al.[14] Therefore, the existing 60-second epoch cut-points are adapted to the shorter epochs; for example, the 30-second epoch cut-point would be the 60-second epoch cut-point divided by two. It has previously been noted that shorter epoch lengths may provide significantly different estimates of daily MVPA and SB, mainly in studies on child populations,[4,12,15,16] and also in some studies based on adult populations.[11,17] In a sample of 102 overweight post-menopausal women, Gabriel et al.[17] found an extra 16 min of MVPA per day when adapting the Troiano[18] cut-point to 10-second epochs compared to the original 60-second epoch cut-point (equivalent to a 58% increase in the estimated min of MVPA when using 10-second epochs). Orme et al.[11] found similar results for slightly younger sample of men and women (n = 267) using the same cut-point, reporting an extra 16 min of MVPA per day when using 5-second epochs compared to 60-second epochs (equivalent to a 62% increase in the estimated min of MVPA). Gabriel et al. also found 1 h and 39 min more SB time per day when using the 10-second compared to 60-second epochs (equivalent to a 19% increase). The results for studies on children presented similar results, with a clearly higher estimation of MVPA and SB when using shorter epoch lengths.[12,15]

It is seldom highlighted that cut points are determined at a specific epoch length, and that shorter (or longer) epoch lengths might result in very different indicators of activity intensity. In response to these knowledge gaps, using free-living data, this study will compare the impact of epoch length on estimates of SB, LPA and MVPA. It will compare the impact of epoch length for both tri-axial cut-points and uni-axial cut-points. Finally, we will compare the impact of epoch length on physical activity accumulated in different domains (workplace, transport, leisure), to illustrate the differential impact of the epoch length on activity intensity measures given the activity patterns of individuals.

2. Methods

2.1 Population

As previously described in detail, the RECORD participants were recruited during preventive health checkups in 2007–2008 and 2011–2013, born between 1928–1978, and resided at baseline in 112 municipalities of the Ile-de-France Paris region.[19–22] In the second wave of the study,[23,24] after undergoing a medical checkup and completing computerized questionnaires at the IPC Medical Centre,[25,26] 410 individuals were invited to participate in the RECORD GPS Study.[27] Written informed consent was obtained from all participants. The RECORD GPS Study was approved by the French Data Protection Authority.

2.2 Data collection procedures

The recruitment was guided by using a standardized recruitment form. Participants wore a BT-Q1000XT GPS (QStarz) and a GT3X+ accelerometer (Actigraph) on the right hip on a dedicated elastic belt for the recruitment day and seven additional days, all day long from the time of waking up until bedtime. The participants completed a travel diary to report their activity places over 7-8 days, each time with arrival and departure times. After linear interpolation of the missing data, the GPS data were analyzed with an algorithm (ArcGIS Python script) that identified all of the activity locations of the participants (any activity at a stationary location) from the accumulation of GPS points over 7 days.[28] The algorithm automatically uploaded the history of visits to places into the Mobility Web Mapping application. As previously described,[27] this information and the travel diary was then used for the prompted recall survey conducted during a phone call.[25] The survey operator could report activity locations and trips undetected by the algorithm and could modify/remove detected visits to locations that were inaccurate or incorrect. Of the 410 invited, 247 subjects agreed to participate. Nine participants abandoned the study, data collection failed for 4 participants due to protocol failure, and 7 had worn the ActiGraph GT3X (as opposed to the GT3X+) which did not allow us to obtain the raw accelerometer data. Of the 227 with accelerometer data available, 224 participants had at least 1 observation day with 10 h of accelerometer wear time or more, which was considered as the minimum wear time for a valid day. Non-wear time was defined as proposed by Choi et al.,[29] based on standard filter 60-second epoch counts. Supplementary material S1 gives an overview of the technical characteristics of the device and the decision making in the data processing. The participants were adults aged 35 to 83 years (median: 58 years old) and 87 participants were women (39%).

2.3 Measures (MVPA time and SB time)

Table 1 provides an overview of the different measures compared in this article. The accelerometer recorded the acceleration on the vertical, antero-posterior and medio-lateral axis. From the raw accelerometer data, counts per second were extracted in ActiLife 6. The 1-second epoch counts were summed up for the larger epoch lengths: 15-second, 30-second and 60-second epochs. Then, we estimated for each epoch the intensity level by grouping the counts per epoch into three categories: MVPA above the MVPA cut-point, SB below the SB cut-point and LPA between the cut-points.

For both the MVPA and SB cut-points, tri-axial and uni-axial variants were applied. Tri-axial cut-points are based on the vector magnitude which summarizes movement of all three axes (vertical, antero-posterior and medio-lateral axes) by taking the square root of the sum of the squared counts of each axis. The Sasaki cut-point for MVPA (in the ActiLife software known as Freedson 2011) is 2691 vector magnitude counts per minute (CPM).[30] The Aguilar- Farías cut-point for SB based on the vector magnitude is 200 CPM.[31] This cut-point has been calibrated for older adults and is included here for illustrative reasons only. No vector magnitude cut-point for adults was found in the literature. The uni-axial cut-points are based on the vertical axis counts. The Freedson cut-point for MVPA is 1952 CPM;[6] the Matthews SB cut-point is 100 CPM on the vertical axis.[32] To adapt the cut-points to shorter epochs, the original cut-point value was divided by the appropriate factor. For example, the 1-second adaption of the Sasaki cut-point was 45 counts per second, or 2691 divided by 60. The Sasaki and Aguilar- Farías cut-points were calibrated with newer ActiGraph devices (GT3X); and the Freedson and Matthews cut-points were calibrated with older devices. Therefore, the latter measures are used in combination with the low frequency extension filter since counts calculated with this filter better approximate the counts obtained from the older devices.[4]

From the GPS-enhanced recall survey described above, we could identify the time spent at leisure-time, occupation and transport locations. Transport time is defined as the time spent between two activity places (i.e., the places visited by the participant for which a function can be identified such as a residence, workplace or shop); occupation time as the time spent at the work place; and leisure-time is the time spent at activity places other than the work place.

Medians, interquartile ranges (IQR), means and 95% confidence intervals (CI) were calculated for the accumulated time spent in MVPA, LPA and SB per day for the three physical activity domains combined and separately. The mean and CI estimates are corrected for the hierarchical structure of the dataset (observation days nested within people) by linear mixed models with no independent variables. These models were estimated with the 'lme4' package[33] in R. To compare 1-second and 60-second activity intensity estimations, variables were constructed at the individual level by subtracting the 60-second estimation from the 1-second estimation.

3. Results

Table 2 presents the repartitioning of the total wear time into MVPA, LPA and SB for 4 different epoch lengths. The differences between 60-second and 1-second epochs were considerable for the Sasaki MVPA measure (47 $min{\cdot}d^{-1}$, confidence interval (CI): 45-49 $min{\cdot}d^{-1}$); the Freedson MVPA measure (19 $min{\cdot}d^{-1}$, CI: 18-21 $min{\cdot}d^{-1}$); the Aguilar- Farías SB measure (162 $min{\cdot}d^{-1}$, CI: 157-168 $min{\cdot}d^{-1}$); and the Matthews SB measure (152 $min{\cdot}d^{-1}$, CI: 146-158 $min{\cdot}d^{-1}$).

Since both SB and MVPA were estimated to be higher with shorter epoch lengths, the complimentary time spent in LPA is lower. The estimated LPA is 209 $min{\cdot}d^{-1}$ (CI: 202-217 $min{\cdot}d^{-1}$) lower when using 1-second epochs compared to 60-second epochs using the vector magnitude cut-points and 171 $min{\cdot}d^{-1}$ (CI: 165-178 $min{\cdot}d^{-1}$) lower when using vertical axis cut-points.

Therefore, the relative time spent doing LPA changed drastically when using 1s epoch compared to 60-second epochs.

The impact of epoch lengths on moderate-to-vigorous physical activity and sedentary time estimations for adults — 4/9

Table 1. The original (60-s epochs) and adapted cut-points for measures of MVPA, LPA and SB; based on tri-axial (vector magnitude) and uni-axial (vertical axis) data

	Cut-points in CPM by epoch length			
	1-s	15-s	30-s	60-s
Vector magnitude				
MVPA [1]	> 45	> 673	> 1345	> 2690
LPA	3.3 – 45	50 – 673	100 – 1345	200 – 2690
SB [2]	< 3.3	< 50	< 100	< 200
Vertical axis				
MVPA [3]	> 33	> 488	> 976	> 1951
LPA	1.7 – 33	25 – 488	50 – 976	100 – 1951
SB [4]	< 1.7	< 25	< 50	< 100

MVPA: moderate-to-vigorous physical activity; LPA: light physical activity; SB: sedentary behavior; CPM: counts per minute; [1] MVPA cut-point calibrated by Sasaki et al. (2011) for adults; [2] SB cut-point calibrated by Aguilar et al. (2014) for older adults; [3] MVPA cut-point calibrated by Freedson et al. (1998) for young adults; [4] SB cut-point calibrated by Treuth et al (2004) for adolescent girls and confirmed by Matthews et al. (2008) for adults.

The estimated LPA was 63 and 60 percentage lower for LPA calculated with the vector magnitude and vertical axis cut-points respectively. The estimated SB was 34 and 29 percentage higher for 1-second epochs compared to 60-second epochs calculated with the Aguilar- Farías and Matthews cut-points respectively. The estimated time spent doing MVPA was 96 and 46 percentage higher calculated with the Sasaki and Freedson cut-points respectively.

Finally, we compare the impact of epoch length on the Sasaki MVPA measure at different locations (workplace, transport, leisure), to illustrate the differential impact of the epoch length on activity intensity measures given the activity patterns of individuals. The people in this sample spent on average 125 min on transport per day, 164 min at the workplace and 568 min at other locations including home (leisure time). Table 3 presents the median and mean $min \cdot d^{-1}$ spent doing MVPA by epoch length and by location type using the Sasaki cut-point. Supplementary Material 2 presents all six activity intensity measures included in this article by location type.

The Sasaki MVPA estimates were higher for 1-second epochs compared to 60-second epochs within each physical activity domain, but the increases were very unequal. The 47 extra $min \cdot d^{-1}$ in the Sasaki MVPA measure using the 1-second epochs (equal to 96% of the 60-second estimate) were mainly due to higher estimates of leisure-time physical activity (+34 $min \cdot d^{-1}$ or +169%) and occupations physical activity (+9 $min \cdot d^{-1}$ or +155%) and to a lower extent to a higher estimate of transport-related physical activity (+4 $min \cdot d^{-1}$ or +18%).

This resulted in very different estimates of the importance of each physical activity domain. Using the original cut-points (60-second epochs), transport had the highest contribution to total physical activity, followed by leisure-time and occupational physical activity. Using the 1-second Sasaki cut-point, the attribution of transport-related MVPA was estimated 19 percentage points lower and the attribution of leisure-time MVPA 15 percentage points higher compared to the original cut-points.

Table 4 compares the impact of using 1-second epoch lengths on activity intensity indicators with the impact of other decisions in the data processing. The difference in the MVPA between using 1-second and 60-second epochs was 47 min for tri-axial measure and 19 min for the uni-axial measure; whereas the largest impact from the other data processing decisions was 9 min for the filter choice when using the tri-axial measure. The impact of the cut-point (Sasaki vs Freedson) was 6 min. The findings for LPA and SB were similar to those for MVPA, with substantially higher impact of using 1-second epochs compared to the other data processing decisions.

4. Discussion

There was a consistent and large impact of shorter epoch lengths on activity intensity measures MVPA, LPA and SB throughout the results. All the cut-points presented in this paper were calibrated on 60-second epochs. Using 1-second epochs instead of 60-second epochs increased the estimated time doing MVPA and SB for both vector magnitude cut-points and vertical axis cut-points. Consequently, LPA was significantly lower using 1-second epochs.

To put the magnitude of the results into perspective, we compared the impact of epoch lengths to those of other data processing decision. The estimated differences between 60-second and 1-second epochs were substantially larger than for the cut-points (vector magnitude versus vertical axis cut-points), the filter choice (standard versus LFE filter), the non-wear algorithm (Choi algorithm versus Troiano algorithm), and the number of hours to identify valid days (10hrs versus 8hrs). Other cut-points or data reduction decisions could have been chosen, for which the differences may have been greater. Nevertheless, the results do indicate a very substantial impact of the epoch length compared to other data processing decisions.

As for the use of different cut-points,[34] one could try to find a formula to transform activity intensity estimations based on shorter epoch lengths into 60-second epoch estimations.

The impact of epoch lengths on moderate-to-vigorous physical activity and sedentary time estimations for adults — 5/9

Table 2. Estimated time spent doing MVPA, LPA and SB by epoch length: vector magnitude and vertical axis measures

	Epoch length (median and IQR)				Epoch length (mean and 95% CI)			
	1-s	15-s	30-s	60-s	1-s	15-s	30-s	60-s
Vector magnitude								
MVPA [1]	90 (65; 120)	60 (39; 86)	50 (31; 76)	42 (23; 67)	96 (91; 100)	66 (63; 70)	57 (53; 61)	49 (45; 52)
LPA	114 (87; 148)	248 (192; 310)	284 (223; 357)	322 (252; 402)	121 (115; 126)	255 (245; 265)	293 (282; 304)	330 (318; 342)
SB [2]	640 (551; 731)	533 (443; 627)	503 (410; 598)	476 (382; 569)	635 (623; 646)	529 (516; 542)	500 (487; 513)	472 (458; 486)
Vertical axis								
MVPA [3]	57 (38; 81)	46 (26; 69)	41 (21; 64)	36 (16; 60)	62 (59; 66)	51 (48; 55)	47 (44; 50)	43 (39; 46)
LPA	106 (80; 140)	198 (152; 257)	236 (182; 301)	275 (214; 348)	113 (108; 118)	207 (199; 216)	246 (236; 255)	284 (273; 295)
SB [4]	685 (595; 771)	598 (506; 687)	564 (470; 654)	529 (434; 623)	676 (666; 687)	592 (580; 604)	558 (546; 570)	524 (511; 537)

Number of valid observation days = 1389; Number of persons = 224; Mean wear time during valid observation days = 858 min (= 14 h 18 min); VM: vector magnitude, VA: vertical axis; MVPA: moderate-to-vigorous physical activity; LPA: light physical activity; SB: sedentary behavior; IQR: interquartile range; 95% CI: 95% confidence interval; [1] MVPA cut-point calibrated by Sasaki et al. (2011) for adults; [2] SB cut-point calibrated by Aguilar et al. (2014) for older adults; [3] MVPA cut-point calibrated by Freedson et al. (1998) for young adults; [4] SB cut-point calibrated by Treuth et al (2004) for adolescent girls and confirmed by Matthews et al. (2008) for adults.

This would then make it possible to compare studies that used different epoch lengths. However, the impact of epoch length on activity intensity levels was substantially different for three types of locations, i.e. workplace, transport and leisure-time locations. This indicates that the estimated daily activity intensity of people with different activity patterns will be impacted differentially by epoch length. Therefore, it is unlikely that a transformation formula could be found without extensive information on activity patterns; if one could be found at all.

These results by location also imply that using 15-second epochs for children and 60-second epochs for adult populations make it almost impossible to investigate the transition from childhood to adulthood in terms of physical activity. Imagine a follow-up study of adolescents into adulthood that also takes into account the locations of physical activity; which is important information for policy and interventions. And imagine that in reality, the persons' physical activity behaviors did not change during the follow-up period. In that case, the importance of different locations (e.g. transport) to the total physical activity levels will change simply by using 15-second epoch cut-point for adolescents and 60-second epoch cut-point for adults; even under the assumption that the respective cut-points correctly measure the total activity intensity levels. Moreover, considering the recent availability of a 15-second epoch cut-point for older adults,[14] there seems an urgent need for a validated 15-second epoch cut-point for adults if we want to evaluate correctly the important transitions into adulthood and older adulthood in terms of physical activity.

Recently, tri-axial accelerometers became the norm. This evolution might lead to an increasing use of vector magnitude instead of vertical axis cut-points. The impact of using shorter epoch lengths on activity intensity measures was the highest for the vector magnitude measures compared to vertical axis measures; with the most distinct differences for the MVPA measures. One reason may be the calculation of the vector magnitude counts. Where the sum of 60 1-second epoch counts on the vertical axis is equal to the 60-second epoch counts; the sum of vector magnitude counts is not equal to the vector magnitude of the sum of the counts. The within minute variability, which is the reason for the impact of the epoch length, may therefore be amplified when using vector magnitude measures. (See Supplementary material S3 for more details of why the epoch length impacts activity intensity measures.) Comparing a vector magnitude measure with vertical axis measures, Banda et al.[15] did not find a higher impact of the epoch length on MVPA and SB for vector magnitude measures. However, the vector magnitude measure had a much larger estimate for MVPA and a lower estimate for SB compared to the vertical axis measures, when applying the epoch lengths used at during the calibration studies. This may indicate that the vector magnitude measures were not comparable to the vertical axis measures, confounding the impact of the epoch length on the estimated activity intensity.

A greater impact of the epoch length for vector magnitude measures does not indicate that these measures are worse than vertical axis measures. Comparative studies including a golden standard for physical activity energy expenditure are needed to investigate which epoch length results in a measure

Table 3. Daily minutes spent doing MVPA by epoch length and domain using Sasaki's vector magnitude cut-point

	Epoch length (median and IQR)				Epoch length (mean and 95% CI)			
	1-s	15-s	30-s	60-s	1-s	15-s	30-s	60-s
Total	90 (65; 120)	60 (39; 86)	50 (31; 76)	42 (23; 67)	96 (91; 100)	66 (63; 70)	57 (53; 61)	49 (45; 52)
Leisure	47 (26; 74)	24 (13; 44)	18 (9; 32)	13 (6; 25)	54 (51; 58)	33 (30; 35)	26 (23; 28)	20 (18; 22)
Transport	21 (7; 41)	18 (5; 38)	17 (4; 37)	16 (3; 36)	27 (25; 30)	25 (23; 27)	24 (22; 26)	23 (21; 25)
Occupation	0 (0; 22)	0 (0; 12)	0 (0; 9)	0 (0; 6)	14 (12; 17)	9 (7; 11)	7 (6; 9)	6 (4; 7)

Number of valid observation days = 1389; Number of persons = 224; Mean wear time during valid observation days = 858 min (= 14 h 18 min); MVPA: moderate-to-vigorous physical activity; IQR: interquartile range; CI: confidence intervals.

that approximates most accurately the activity intensity levels; and studies linking accelerometer data to health outcomes are needed to determine which epoch length results into the measure with the highest explicative power. These results do however indicate that failing to use epoch lengths in a standardized way may impact the comparability of studies using vector magnitude measures even more than those using vertical axis measures.

A strength of this study was the use of raw accelerometer data collected in free-living conditions to examine the impact of the epoch length; which to our knowledge is a first for adults, and only recently preceded for children.[15] Using raw data prevents reintegrating epochs in larger epochs; a practice that introduces biases independent of the epoch length.[35] Free-living data is important to investigate real-life activity patterns, instead of activities with constant activity intensities in laboratory studies. Constant activity intensities might lead to a very different impact of the epoch length. An indication can be found in the estimated activity intensity levels by location; where the impact of epoch length was much lower for transport-time (arguably more constant activities) compared to other activities. More research is however needed to conclude on this.

The sample used in this study is not representative for the background population: people aged 35-83 years old, residing in Ile-de-France. The MVPA, SB and LPA levels reported in this paper might therefore be higher or lower than the population levels. The sample includes less women (37%, compared to 53% in the population), more older people (48% above 60 compared to 31%) and more people from the inner city of Paris (26% compared to 19%).

For this study, we made the assumption that 60-second epoch measures are the reference measures because 60-second epochs were used in the calibration studies; while there is no information on the validity of the shorter epoch lengths in combination with the cut-points. This assumption does not contradict that shorter epoch lengths could be beneficial for future cut-points for adults; similar to those proposed for children[36] and recently for older adults.[14]

This study supports earlier calls for more harmonizing data processing.[37] We would suggest a fairly simple practice: using the measures as they were conceived. The decisions during the data processing are often presented as independent of each other. However, it makes sense to use cut-points in combination with a package of data collection and processing decisions equal to (or as similar as) the decisions made during the calibration study, including the epoch length. For example, the Sasaki and the Freedson cut-points for MVPA were calibrated in very similar studies;[30] so the sometimes large differences in estimates of MVPA presented in this study were surprising. However, when using the epoch length used in the calibration studies (60-s) and the LFE-filter for the Freedson measure to better approximate the CSA device, the difference in estimated MVPA at its lowest, even though still substantial (6 min per day).

Another practice that could help harmonizing the results of accelerometer studies is more detailed and structured reporting of the data processing decisions, especially in calibration studies. This would make it easier for users that are less informed on the technical details, to use accelerometer data in a standardized manner.[13,38] Supplementary Material 1 could be a starting point for this. In the harmonization of the measures, there is also a role for the data processing software. Instead of leaving each decision up to the user, a package of decisions, including epoch length, could be offered to help the user in avoiding making mistakes.

Next to harmonizing measures, there is still much work needed to identify the best measures of physical activity. At the moment, we feel that there is no consensus on the best measure. A good measure needs to be comprehensible enough for a wider public to be used in interventions and communication campaigns, and it needs to be relevant in relation to health outcomes. Recently, new data processing techniques have been proposed,[39] revisiting the data processing with machine learning techniques that could allow for reducing the number of arbitrary decisions by using raw data instead of count data. This type of measures uses more complex models to predict activity intensity compared to the relative simple cut-point

The impact of epoch lengths on moderate-to-vigorous physical activity and sedentary time estimations for adults — 7/9

Table 4. Impact of data processing decisions on physical activity indicators (means and CI's)

	Original measure	Epochs (1-s)	Filter choice [a]	Valid days limit (8 h)	Non wear (Troiano)
Vector magnitude					
MVPA[1]	49 (45; 52)	96 (91; 100)	58 (54; 61)	49 (46; 53)	49 (45; 53)
LPA	330 (318; 342)	121 (115; 126)	353 (340; 365)	333 (321; 345)	325 (313; 337)
SB [2]	472 (458; 486)	635 (623; 646)	440 (426; 454)	444 (431; 456)	462 (448; 476)
Vertical axis					
MVPA [3]	43 (39; 46)	62 (59; 66)	40 (36; 43)	43 (40; 47)	43 (39; 46)
LPA	284 (273; 295)	113 (108; 118)	253 (243; 263)	287 (276; 298)	280 (269; 291)
SB [4]	524 (511; 537)	676 (666; 687)	558 (546; 571)	496 (485; 508)	513 (500; 526)

Number of valid observation days = 1389; Number of persons = 224; VM: vector magnitude, VA: vertical axis; MVPA: moderate-to-vigorous physical activity; LPA: light physical activity; SB: sedentary behavior; IQR: interquartile range; 95% CI: 95% confidence interval; [1] MVPA cut-point calibrated by Sasaki et al. (2011) for adults; [2] SB cut-point calibrated by Aguilar et al. (2014) for older adults; [3] MVPA cut-point calibrated by Freedson et al. (1998) for young adults; [4] SB cut-point calibrated by Treuth et al (2004) for adolescent girls and confirmed by Matthews et al. (2008) for adults; [a] LFE-filter for tri-axial measures and standard filter for uni-axial measures.

measures and avoid certain data processing decisions such as epoch length; and might, therefore, prove to be more reliable measures. It will be important, however, to introduce these new measures enabling standardized use by the end users.

5. Conclusion

This study investigates the often overlooked impact of epoch length on activity intensity measures, based on free-living data of adults living in the French capital region, Ile-de-France. The results show a great impact of the epoch length on all three activity intensity measures considered: moderate-to-vigorous physical activity, light physical activity and sedentary behavior. Concrete indications are given on how to standardize the use of activity intensity measures based on accelerometer data; which will become only more important with the expected introduction of new types of measures in the near future.

Acknowledgments

The RECORD GPS Study was supported by the INPES (National Institute for Prevention and Health Education), the Ministry of Ecology (DGITM), CERTU (Centre for the Study of Networks, Transport, Urbanism, and Public constructions), ARS (Health Regional Agency) of Ile-de-France, STIF (Ile-de-France Transport Authority), the Ile-de-France Regional Council, RATP (Paris Public Transport Operator), and DRIEA (Regional and Interdepartmental Direction for Equipment and Planning). The authors thank the following partners from the funding institutions: Pierre Arwidson, Nadine Asconchilo, Annette Gogneau, Colette Watellier, Yasmina Baaba, Mélani Alberto, Christelle Paulo, Anne-Eole Meret-Conti, Cédric Aubouin, Benoît Kiéné, Hélène Pierre, Sophie Mazoué, John Séraphin, and Ivan Derré.

References

[1] Bassett DR, Troiano RP, McClain JJ, Wolff DL. Accelerometer-based physical activity: total volume per day and standardized measures. Med Sci Sports Exerc. 2015;47(4):833–8. Med Sci Sports Exerc.

[2] Lee IM, Shiroma EJ. Using accelerometers to measure physical activity in large-scale epidemiological studies: issues and challenges. Br J Sports Med. 2014;48(3):197–201.

[3] World Health Organization. Assessing national capacity for the prevention and control of noncommunicable diseases: report of the 2010 global survey; 2012. Available from: `http://www.who.int/chp/knowledge/national_prevention_ncds/en/`.

[4] Cain KL, Sallis JF, Conway TL, Van Dyck D, Calhoon L. Using accelerometers in youth physical activity studies: a review of methods. J Phys Act Health. 2013;10(3):437–50.

[5] Hildebrand M, VT VANH, Hansen BH, Ekelund U. Age group comparability of raw accelerometer output from wrist- and hip-worn monitors. Med Sci Sports Exerc. 2014;46(9):1816–24.

[6] Freedson PS, Melanson E, Sirard J. Calibration of the Computer Science and Applications, Inc. accelerometer. Med Sci Sports Exerc. 1998;30(5):777–81.

[7] Troiano RP. Large-scale applications of accelerometers: new frontiers and new questions. Med Sci Sports Exerc. 2007;39(9):1501.

[8] Tudor-Locke C, Johnson W, Katzmarzyk P. Accelerometer-Determined Steps per Day in US Adults. Med Sci Sports Exerc. 2009;41(7):1384–1391.

[9] Aadland E, Steene-Johannessen J. The use of individual cut points from treadmill walking to assess free-living moderate to vigorous physical activity in obese subjects by accelerometry: is it useful? BMC Med Res Method. 2012;12:172.

[10] TROST S, LOPRINZI P, MOORE R, PFEIFFER K. Comparison of Accelerometer Cut Points for Predicting Activity Intensity in Youth. Med Sci Sports Exerc. 2011;43(7):1360–1368.

[11] Orme M, Wijndaele K, Sharp SJ, Westgate K, Ekelund U, Brage S. Combined influence of epoch length, cut-point and bout duration on accelerometry-derived physical activity. Int J Behav Nutr Phys Act. 2014;11(1):34.

[12] McClain JJ, Abraham TL, Brusseau J T A, Tudor-Locke C. Epoch length and accelerometer outputs in children: comparison to direct observation. Med Sci Sports Exerc. 2008;40(12):2080–7.

[13] Heil DP, Brage S, Rothney MP. Modeling physical activity outcomes from wearable monitors. Med Sci Sports Exerc. 2012;44(1 Suppl 1):S50–60.

[14] Evenson KR, Wen F, Herring AH, Di C, LaMonte MJ, Tinker LF, et al. Calibrating physical activity intensity for hip-worn accelerometry in women age 60 to 91 years: The Women's Health Initiative OPACH Calibration Study. Prev Med Rep. 2015;2:750–756. Prev Med Rep.

[15] Banda JA, Haydel KF, Davila T, Desai M, Bryson S, Haskell WL, et al. Effects of Varying Epoch Lengths, Wear Time Algorithms, and Activity Cut-Points on Estimates of Child Sedentary Behavior and Physical Activity from Accelerometer Data. PLoS ONE. 2016;11(3):e0150534.

[16] Colley RC, Harvey A, Grattan KP, Adamo KB. Impact of accelerometer epoch length on physical activity and sedentary behaviour outcomes for preschool-aged children. Health Rep. 2014;25(1):3–9.

[17] Gabriel KP, McClain JJ, Schmid KK, Storti KL, High RR, Underwood DA, et al. Issues in accelerometer methodology: the role of epoch length on estimates of physical activity and relationships with health outcomes in overweight, post-menopausal women. Int J Behav Nutr Phys Act. 2010;7:53.

[18] Troiano RP, Berrigan D, Dodd KW, Masse LC, Tilert T, McDowell M. Physical activity in the United States measured by accelerometer. Med Sci Sports Exerc. 2008;40(1):181–8.

[19] Brondeel R, Weill A, Thomas F, Chaix B. Use of healthcare services in the residence and workplace neighbourhood: the effect of spatial accessibility to healthcare services. Health Place. 2014;30:127–33. Health Place.

[20] Chaix B, Bean K, Daniel M, Zenk SN, Kestens Y, Charreire H, et al. Associations of Supermarket Characteristics with Weight Status and Body Fat: A Multilevel Analysis of Individuals within Supermarkets (RECORD Study). PLoS ONE. 2012;7(4):e32908. PLoS ONE.

[21] Chaix B, Simon C, Charreire H, Thomas F, Kestens Y, Karusisi N, et al. The environmental correlates of overall and neighborhood based recreational walking (a cross-sectional analysis of the RECORD Study). Int J Behav Nutr Phys Act. 2014;11(1):20.

[22] Van Hulst A, Thomas F, Barnett TA, Kestens Y, Gauvin L, Pannier B, et al. A typology of neighborhoods and blood pressure in the RECORD Cohort Study. J Hypertens. 2012;30(7):1336–46.

[23] Chaix B, Kestens Y, Bean K, Leal C, Karusisi N, Meghiref K, et al. Cohort profile: residential and non-residential environments, individual activity spaces and cardiovascular risk factors and diseases–the RECORD Cohort Study. Int J Epidemiol. 2012;41(5):1283–92.

[24] Perchoux C, Kestens Y, Thomas F, Van Hulst A, Thierry B, Chaix B. Assessing patterns of spatial behavior in health studies: Their socio-demographic determinants and associations with transportation modes (the RECORD Cohort Study). Soc Sci Med. 2014;119:64–73.

[25] Chaix B, Kestens Y, Perchoux C, Karusisi N, Merlo J, Labadi K. An interactive mapping tool to assess individual mobility patterns in neighborhood studies. Am J Prev Med. 2012;43(4):440–50.

[26] Leal C, Bean K, Thomas F, Chaix B. Multicollinearity in Associations Between Multiple Environmental Features and Body Weight and Abdominal Fat: Using Matching Techniques to Assess Whether the Associations are Separable. Am J Epidemiol. 2012;175(11):1152–1162.

[27] Chaix B, Kestens Y, Duncan S, Merrien C, Thierry B, Pannier B, et al. Active transportation and public transportation use to achieve physical activity recommendations? A combined GPS, accelerometer, and mobility survey study. Int J Behav Nutr Phys Act. 2014;11(1):124.

[28] Thierry B, Chaix B, Kestens Y. Detecting activity locations from raw GPS data: a novel kernel-based algorithm. Int J Health Geogr. 2013;12(14):14.

[29] Choi L, Liu Z, Matthews CE, Buchowski MS. Validation of accelerometer wear and nonwear time classification algorithm. Med Sci Sports Exerc. 2011;43(2):357–64.

[30] Sasaki JE, John D, Freedson PS. Validation and comparison of ActiGraph activity monitors. J Sci Med Sport. 2011;14(5):411–6.

[31] Aguilar-Farías N, Brown WJ, Peeters GM. ActiGraph GT3X+ cut-points for identifying sedentary behaviour in older adults in free-living environments. J Sci Med Sport. 2014;17(3):293–9. J Sci Med Sport.

[32] Matthews CE, Chen KY, Freedson PS, Buchowski MS, Beech BM, Pate RR, et al. Amount of time spent in sedentary behaviors in the United States, 2003-2004. Am J Epidemiol. 2008;167(7):875–81.

[33] Douglas B, Martin M, Ben B, Steve W. Fitting Linear Mixed-Effects Models Using lme4. J Stat Softw. 2015;67(1):1–48.

[34] Brazendale K, Beets MW, Bornstein DB, Moore JB, Pate RR, Weaver RG, et al. Equating accelerometer estimates among youth: The Rosetta Stone 2. J Sci Med Sport. 2016;19(3):242–9. Journal of science and medicine in sport / Sports Medicine Australia.

[35] Kim Y, Beets MW, Pate RR, Blair SN. The effect of reintegrating Actigraph accelerometer counts in preschool children: comparison using different epoch lengths. J Sci Med Sport. 2013;16(2):129–34.

[36] Evenson KR, Catellier DJ, Gill K, Ondrak KS, McMurray RG. Calibration of two objective measures of physical activity for children. J Sports Sci. 2008;26(14):1557–65.

[37] Wijndaele K, Westgate K, Stephens SK, Blair SN, Bull FC, Chastin SF, et al. Utilization and Harmonization of Adult Accelerometry Data: Review and Expert Consensus. Med Sci Sports Exerc. 2015;47(10):2129–39.

[38] Ainsworth B, Cahalin L, Buman M, Ross R. The current state of physical activity assessment tools. Prog Cardiovasc Dis. 2015;57(4):387–95.

[39] Staudenmayer J, Pober D, Crouter S, Bassett D, Freedson P. An artificial neural network to estimate physical activity energy expenditure and identify physical activity type from an accelerometer. J Appl Physiol. 2009;107(4):1300–7.

3.2 Article 2: Using GPS, GIS, and accelerometer data to predict transportation modes

SPECIAL COMMUNICATIONS

Methodological Advances

Using GPS, GIS, and Accelerometer Data to Predict Transportation Modes

RUBEN BRONDEEL[1,2,3], BRUNO PANNIER[4], and BASILE CHAIX[1,2]

[1]Institut National de la Santé et de la Recherche Médicale, UMR_S 1136, Pierre Louis Institute of Epidemiology and Public Health, Research Team in Social Epidemiology, Paris, FRANCE; [2]Sorbonne Universités, Université Pierre et Marie Curie Univ Paris 06, UMR_S 1136, Pierre Louis Institute of Epidemiology and Public Health, Research Team in Social Epidemiology, Paris, FRANCE; [3]Ecole des Hautes études en Santé Publique School of Public Health, Rennes, FRANCE; and [4]IPC Medical Centre, Paris, FRANCE

ABSTRACT

BRONDEEL, R., B. PANNIER, and B. CHAIX. Using GPS, GIS, and Accelerometer Data to Predict Transportation Modes. *Med. Sci. Sports Exerc.*, Vol. 47, No. 12, pp. 2669–2675, 2015. **Introduction**: Active transportation is a substantial source of physical activity, which has a positive influence on many health outcomes. A survey of transportation modes for each trip is challenging, time-consuming, and requires substantial financial investments. This study proposes a passive collection method and the prediction of modes at the trip level using random forests. **Methods**: The RECORD GPS study collected real-life trip data from 236 participants over 7 d, including the transportation mode, global positioning system, geographical information systems, and accelerometer data. A prediction model of transportation modes was constructed using the random forests method. Finally, we investigated the performance of models on the basis of a limited number of participants/trips to predict transportation modes for a large number of trips. **Results**: The full model had a correct prediction rate of 90%. A simpler model of global positioning system explanatory variables combined with geographical information systems variables performed nearly as well. Relatively good predictions could be made using a model based on the 991 trips of the first 30 participants. **Conclusions**: This study uses real-life data from a large sample set to test a method for predicting transportation modes at the trip level, thereby providing a useful complement to time unit-level prediction methods. By enabling predictions on the basis of a limited number of observations, this method may decrease the workload for participants/researchers and provide relevant trip-level data to investigate relations between transportation and health. **Key Words:** PHYSICAL ACTIVITY, ACTIVE TRANSPORT, PASSIVE DATA COLLECTION, MACHINE LEARNING, RECORD COHORT STUDY, FRANCE

Physical activity has a positive influence on several health outcomes, such as obesity, cardiovascular health problems, depression, and certain cancers (15,36,39). Active transportation modes, such as walking, biking, and public transport, represent a substantial source of physical activity (27,28). However, reliably assessing the use of transportation modes has proven challenging (9,11,12), thereby hindering the study of the relation between transportation and physical activity. Self-reported measures of the use of transportation modes are prone to memory biases (3). Short trips, especially walking trips, tend to be underreported. Moreover, the time spent during car trips tends to be underreported, whereas the time spent in public transport tends to be exaggerated.

Using objective measurements using accelerometers or global positioning system (GPS) receivers is useful to overcome some of these issues. These devices can, in theory, register the spatial location and body movements of participants over several days. The difficulty lies in transforming the raw data into qualitative trip information, such as the transportation modes used or the departure and arrival locations of each trip.

One approach used in transportation sciences is to perform a so-called GPS-based prompted recall survey, i.e., using information derived from GPS receivers to prompt participant recall (32,38). Using this approach, GPS and accelerometer data are first collected. The departure and arrival points (in space and time) of each trip are then identified by detecting the activity places, i.e., the places visited

Address for correspondence: Ruben Brondeel, M.Sc., UMR_S 1136, Faculté de Médecine Saint-Antoine, 27 Rue Chaligny, 75012 Paris, France; E-mail: Ruben.Brondeel@iplesp.upmc.fr.
Submitted for publication November 2014.
Accepted for publication May 2015.
Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.acsm-msse.org).

0195-9131/15/4712-2669/0
MEDICINE & SCIENCE IN SPORTS & EXERCISE®

DOI: 10.1249/MSS.0000000000000704

by the participant for which a function can be identified such as a residence, workplace or shop. One technique to identify departure and arrival points involves manually segmenting the trips with geographical information systems (GIS) (31). Another approach is to apply algorithms that identify the departure and arrival points of trips on the basis of the raw GPS data (33), as conducted in the RECORD GPS study. Finally, the resulting information is verified and data on the transportation mode in each trip are collected via phone or Internet recall surveys with the participants (3,11,12). Combining device and survey data, the memory bias, and social desirability bias in survey data are reduced by the objective measures. With this approach, information derived concerning trips using the manual processing or automatic algorithms is completed with the survey information. Such GPS-based prompted recall surveys can be performed either at the end of the observation period or on a daily basis during this period (1,32); the latter method is useful for reducing memory biases.

More recently, SenseCam, a camera worn around the neck that takes pictures at regular intervals or when triggered by imbedded sensors, has been suggested to improve data collection of daily activities including trips (6,16,30). Pictures are then used to identify transportation modes or other trip characteristics.

To obtain high-quality data using these approaches, a substantial investment from both participants and research teams is required. In the RECORD GPS study, in which we performed a complete mobility survey for an observation period of 7 d, a research assistant was often able to survey only one participant per day (the entire process included the preparation, the survey, and entering the data into the application). Using SenseCam is likely to be even more burdensome, as research assistants must code all photographs. The time and cost investments required for data collection strongly limit the number of participants, whereas the burden on participant limits the extent of the remainder of the survey.

Therefore, researchers have developed algorithms to predict transportation modes on the basis of device data and sometimes on a limited number of survey items (18,20). Most of these algorithms designed to recognize modes consider short periods (time units) ranging from 1 to 60 s. These algorithms sometimes use sliding windows to optimize the prediction for a given unit using the information from one or more previous and subsequent time frame units. In addition to transportation modes, certain classifications take into account body posture (including lying, sitting, standing, etc.) or household activities. Classifications in these algorithms are based on criteria-based methods, machine learning (such as random forests, support vector machine, and Bayesian network), and probability methods (such as fuzzy logic and multinomial regression) (20).

A smaller number of detection methods, such as the present one, uses trips or trip stages (parts of trips made by a single transportation mode) as the prediction level. These methods first segment the data into trips and activity places and then predict the transportation mode for each trip. This additional step of segmenting the data into trips is an obvious drawback compared with time unit prediction methods. However, trips are meaningful units in behavioral and transportation sciences when analyzing transport-related issues. For example, when studying physical activity associated with the use of public transport, the walking distance required to travel to a train or bus is more important than the physical activity needed during the actual use of these modes. These types of research questions therefore must be addressed at the trip level, thereby making prediction models at the trip level complementary to prediction models at the time unit level.

The present study does not address the segmentation of trips process (algorithms are available for this first step (33)) but rather focuses on transportation mode detection. The aim of this study was to construct an algorithm, building on passive data collection methods that reduce the burden of work for both respondents and research teams. The approach should yield reliable predictions of the transportation mode used at the trip level, which could reduce the time required for the mobility survey or even allow researchers to avoid it completely. We propose a method based on random forests to predict transportation modes at the trip level.

METHODS

Population. As previously described in detail, the RECORD participants were recruited during preventive health checkups in 2007–2008 and 2011–2013, born between 1928 and 1978, and resided at baseline in 112 municipalities of the Île-de-France Paris region (5,7,13,34). In the second wave of the study (8,26), after undergoing a medical checkup and filling computerized questionnaires at the IPC Medical Centre (10,23), 410 individuals were invited to participate in the RECORD GPS study (9), of which 247 subjects agreed to participate. Nine participants abandoned the study, and data collection failed for two participants, thereby yielding a final participation and completion rate of 57.6% (n = 236). A written informed consent was obtained from all participants. The RECORD GPS study was approved by the French Data Protection Authority.

Data collection procedures. The recruitment was guided using a standardized recruitment form. Participants wore a BT-Q1000XT GPS (QStarz) and a GT3X+ accelerometer (ActiGraph) on the right hip with a dedicated elastic belt for the recruitment day and seven additional days, all day long from the time of waking up until bedtime. The participants completed a travel diary to report their activity places over 7 to 8 d, each time with arrival and departure times.

Using a GIS-based Python language algorithm (33) to assess the GPS data, we identified the sequence of activity places for each participant and, consequently, the departure and arrival times of trips between these places. The algorithm automatically uploaded the history of visits to places into the electronic survey application. As previously described (9), this information and the travel diary were then used for

the prompted recall survey conducted during a phone call (10). This procedure resulted in the observation of 7425 trips for 236 participants.

Measures. During the survey, participants reported a chronological sequence of transportation modes for each trip. For modeling purposes, this information was coded into a transportation mode variable consisting of four categories: "walking" (i.e., only walking), "bicycle," "private motorized," and "public transport." When both walking and another transportation mode were sequentially used within a trip, the nonwalking mode was attributed to the trip. We excluded 96 trips with two or more nonwalking modes because they could not be attributed to the mutually exclusive categories of modes required to perform the comparison and there were not enough trips with each combination of two nonwalking modes to define additional categories.

The random forests method is able to use a large variety of variables as predictors of the outcome of interest. However, because the aim of the study is also to lower the burden for researchers, we only used predictors that are relatively easy to define, such as GPS and accelerometer variables, GIS variables that require only standard data, and seven simple survey questions.

The accelerometer recorded the acceleration on three axes for each 5-s epoch or period during the trip. We used both the standard filter and a low-frequency extension filter (37), as implemented in the ActiLife software. The optional low-frequency extension filter extends the lower end of the filter, which is useful for example when processing the data of people who move slowly. On the basis of the raw accelerations obtained with these two filtering approaches, we estimated for each epoch 1) the number of footsteps taken (ActiLife software), 2) the energy expenditure calculated from activity counts and participant weight based on the Sasaki and Freedson equation (29), 3) whether moderate-to-vigorous physical activity (MVPA) was performed (29), and 4) whether the participant was sedentary during the epoch (22). We aggregated these time unit data at the trip level. To capture a maximum of relevant information, we derived standard measures of central tendency (i.e., mean and median) and measures of dispersion (i.e., SD, minimum, maximum, 10th and 90th percentiles). On the basis of the accelerometer data, the accelerations at each of the three axes separately, the number of steps taken, MVPA, sedentary time, and energy expenditure in kilocalories were aggregated in this way. In addition, we calculated the total number of steps taken, the number of MVPA epochs, the number of sedentary epochs, and total energy expenditure for each trip. We also determined the percentage of epochs that were characterized sedentary or MVPA. Each of these variables was calculated for both accelerometer filters.

Every 5 s, the GPS device registered the position coordinates (i.e., latitude, longitude, and elevation), speed, and the following three indicators of the quality of the observation: horizontal, vertical, and positional dilution of precision (HDOP, VDOP, and PDOP, respectively). To derive the summary values described earlier, only the good-quality observations (HDOP < 6, VDOP < 7, PDOP < 8) were retained (9) for the aggregation of time-unit observations at the trip level. GPS observations were determined to be valid, invalid (high dilution of precision), or missing (less than three satellites in view). On average, 27% of GPS observations were missing and 1.5% of the existing observations were invalid. The distribution of potential GPS data points across these three categories provides information on the circumstances of the trips (e.g., underground public transport, tunnel, high buildings). To capture this trip characteristic, the total number of GPS observations, number of valid GPS observations, percentage of valid GPS observations among recorded observations, and percentage of valid GPS observations relative to the maximum number of observations (including missing ones) were included in the model.

On the basis of the GPS data and geographical information on the street network provided by the National Geographic Institute, four distance measures between the departure and arrival points of each trip were calculated, as follows: the straight line distance, the shortest walking distance following the street network, the shortest street network distance by car, and the map-matched distance. The latter distance is based on the most likely route taken by the participant derived by projecting the GPS data points onto the street network (35). These four distance measures and their combination provide complementary information to differentiate between alternative transportation modes. For example, for two trips for which the shortest distance by car would be the same, a difference in the shortest walking distance could add information to differentiate between motorized and nonmotorized transport. Speed measures were calculated on the basis of these distance measures. The GIS was also used to determine whether the residence and the departure and arrival points of each trip were inside or outside the Paris inner city. All geographic calculations were conducted with Python scripts for ArcGIS 10.1. The administrative files of the study provided the sex and age of the participants. During phone call interviews, it was recorded whether the participant possessed a car, bicycle, motorbike, driving license, or public transport pass. Supplemental Digital Content 1 provides an overview of the variables used in the prediction model (see Table, Supplemental Digital Content 1, Overview of 170 predictors used in the random forest models, http://links.lww.com/MSS/A549).

Statistical analysis. We used random forests to predict the transportation mode of each trip (among four possible modes). The random forests method (4) is based on the decision tree method. Decision trees classify data into groups in subsequent steps, each time searching for the feature that best differentiates the group into consideration (branch). To obtain better generalizability, the random forests method adds two sources of randomness to the simple decision tree method and repeats the process a large number of times, thereby resulting into a forest of decision trees. The first source of randomness consists of considering only a random subsample

TABLE 1. Observed and predicted number of trips with each transportation mode.

	Observed			
Predicted	Walking	Bicycle	Private Motorized	Public
Walking	3010	59	229	76
Bicycle	6	115	1	1
Motorized	107	26	2565	112
Public	35	13	65	909

n = 7329.

of the explanatory variables in the definition of each knot of the trees. Secondly, for each tree, only a random subsample of the observations (the trips in our case) is used. Predictions are obtained from each tree for the data not used to grow the tree (so-called out-of-bag data). Finally, a forest prediction of the transportation mode is obtained for each trip as the majority of the tree predictions that were derived when the corresponding trip was out-of-bag. A forest is evaluated on the prediction error rate, in our case, the percentage of trips for which the mode has been wrongly predicted. Regarding missing values, we attributed the median value or the modal value to the corresponding observations for continuous or categorical variables, respectively. All analyses were performed using R with the "randomForest" package (24).

RESULTS

Among the 7329 trips retained for the analyses, 43.1% of the trips were made by walking, 2.9% were with a bike, 39.0% relied on a private motorized vehicle, and 15.0% relied on public transport. The median duration of a trip was 15 min (interdecile range, 3–61 min).

A first forest was grown on the full data set of 7329 trips with all 170 variables. The model had an overall error rate of 10.0% and specific error rates of 4.7% for walking, 46.0% for biking, 10.3% for private motorized transport, and 17.2% for public transport. Table 1 cross-tabulates the observed versus the predicted number of trips for each mode.

The overall error rate was relatively low, but the error rate was larger for the modes with a lower number of trips, such as bicycle or public transport use. When minimizing the overall error rate, classification methods favor precision in the categories with a greater number of observations over precision in the categories with a lower number of observations (25). When interested in greater precision for the smaller categories, the majority-vote-prediction rule can be weighted by the inverse of the probability of belonging to a category. This method greater penalizes the decision rule for mistakes in smaller categories. Growing a random forest using this method, the error rate for the prediction of "bicycle" and "public transport" dropped to 16.9% and 12.8%, respectively. The error rate for the larger categories ("walking" and "private motorized") rose to 14.4% and 19.9%, respectively. The overall error rate rose to 16.4%.

TABLE 2. Error rates (%) of models considering only a subset of the explanatory variables.

	Accelerometer	GPS	GPS/GIS	Accelerometer + GPS	Accelerometer + GPS/GIS
Overall	17.7	17.6	11.6	12.1	10.6
Walking	12.3	7.3	5.6	4.9	4.9
Bicycle	66.2	52.1	51.2	57.3	49.3
Motorized	15.1	14.4	11.3	12.8	10.6
Public	31.0	49.4	21.9	22.4	19.5

No. of trees in each forest = 1000; n = 7329.

The importance of the source of information (accelerometer, GPS, or GPS/GIS data) was then evaluated using separate forests grown with only the respective subsamples of variables (Table 2). The overall error rate for the forest with only the accelerometer variables was 17.7%. The overall error rates for the forests with GPS variables only and GPS/GIS variables only were 17.6% and 11.6%, respectively. Interestingly, the latter error rate was thus not markedly higher than the error rate of the full model (10.0%).

To mimic a study in which participant and trip data are used to predict modes for subsequent trips, forests were grown on the basis of the first 5, 10, 20, 30, 40, 50, 100, 150, and 200 participants. These forests were evaluated by using the prediction error rates for subsequently observed participants (Table 3). A model based on the first five participants (143 trips) yielded a prediction error rate of 28% for the other 231 participants (7187 trips). The overall error rate dropped and then stabilized when at least 30 participants were used to grow the forest (991 trips). The error rates for transportation modes with a larger number of trips were relatively small even for the model based on only a few participants. The gain in prediction quality was relatively small when including additional participants (i.e., more than 30 individuals) in the model. For the transportation modes with a small number of trips, the error rate was high in models with few participants, and it dropped relatively slowly. The reduction in the error rates became negligible only when including more than 50 participants.

DISCUSSION

Main results. When using the data of all participants, the random forest correctly predicted the transportation mode in 90.0% of the trips. This is comparable with the prediction rates found in studies that made predictions at the time-unit level. Ellis et al. (17) have reported prediction rates of 89.8% and 91.9% (depending on the method) when using random forests to predict five different modes for units of 1 min on the basis of GPS and accelerometer data. Using 1-s units,

TABLE 3. Error rates (%) of the predictions from models based on a limited number of participants.

	Overall	Walking	Bicycle	Private Motorized	Public	n
First 5	28.0	7.4	100.0	16.1	95.5	143
First 10	17.0	5.8	85.3	21.6	21.2	298
First 20	15.2	4.6	95.9	14.4	28.4	630
First 30	13.9	4.1	79.5	14.1	26.2	991
First 40	13.7	4.5	81.4	12.6	26.5	1340
First 50	13.0	5.0	57.7	13.0	24.9	1639
First 100	13.6	5.0	64.1	14.3	24.7	3280
First 150	12.9	4.5	64.6	10.8	27.2	4757
First 200	10.2	4.7	61.7	7.6	15.4	6261

No. of trees in each forest = 1000; n = number of trips; total number of trips = 7329.

Feng and Timmermans (18) have found a prediction rate of approximately 90% for eight modes using a Bayesian Belief Network Model using GPS, accelerometer, and survey data.

Few studies addressing mode prediction at the trip or trip-stage level have been reported. Gong et al. (19) and Chen et al. (14) have yielded prediction rates of 79.1% and 82.6%, respectively, in two New York–based studies using a step-by-step algorithm. Other work attempting to predict modes at the level of trip stages (unimodal components of trips) have used more complex strategies. For example, Kohla et al. (21) have used time unit-level detection of walking stages within trips to further segment the trips into trip stages. The nonwalking modes were then identified. Multinomial logistic regression yielded a prediction rate of 80%.

The prediction rates found in the present study are within the range of those reported in the aforementioned studies, which is promising for future applications of the method. However, it is difficult to compare the performance of our algorithm with those of previous studies. Most of these models relied on relatively small convenience samples or scripted/controlled travel behavior data collections (in which participants are asked to follow a specific itinerary with a specific mode). Models based on controlled data to predict activity modes are less generalizable and less apt to predict real-life data (2,16); the same may be expected for the prediction of transportation modes. In contrast, small convenience samples might lack some variety, and they do not represent the relative importance of the different categories well. Because the size of the categories influences the overall prediction rate, these overall rates are not easily comparable between studies. More studies are required to compare the different prediction methods in the same context, with the same quality of data and the same choice of categories (21).

Importantly, we found that the method differentiated between public and private motorized transport well. Additional analysis of these two categories only (not reported) indicated that the highest predictive variables were "possessing a car" (survey), "proportion of valid GPS observations among all possible (including observed and missing) observations," and "possessing a public transport pass." The findings suggest that these indicators that are not always included in published models may be of particular interest. However, it must be kept in mind that the public transport system is particularly well served in Paris and that these variables may have a different predictive contribution in other settings.

Testing trees grown on the data of various numbers of participants enabled us to evaluate the predictive performance of the algorithm for data collected later (i.e., to understand from how many participants detailed mode data should be collected to make reliable predictions using less detailed data). When using no more than 30 participants, the overall prediction rate for the remaining 206 participants was 86.1%. This observation shows that data on a relatively small number of participants can provide valuable information on a much larger data set. However, prediction models based on less than 30 participants displayed poor performance for the mode categories with the fewest trips. To limit the number of participants required to grow the random forest, oversampling the categories with the fewest trips or participants could be considered.

The approach of collecting limited data from the context in which one is willing to make predictions to build a prediction model contrasts with pretrained models (i.e., prediction models trained on data from a different context). Pretrained models are considerably less expensive because no preliminary data collection is required in each particular context. However, pretrained models are less well adapted to the specific context of interest. It can be expected that the optimal set of variables and thresholds of variables used to differentiate transportation modes vary between different contexts. Further studies are required to compare pretrained and same-context prediction models and then determine whether the extra effort of preliminary data collection yields a significant improvement in prediction quality.

In previous studies, it has been argued that accelerometer data may enhance the predictive power of a model for transportation modes, especially concerning trips with frequent missing GPS data values (e.g., during subway use) (18). We found relatively good prediction rates from an accelerometer data-only model. However, we noted only a small increase in prediction rates when including accelerometer data in the GPS/GIS model, which may be attributable to our study design in which observation units represented trips rather than time units, as applied in most previous studies. The indicators associated with GPS data (proportion of invalid or missing data, dispersion of speed throughout the trip) are possibly more informative in the trip-level models than in a time unit-level approach, thus rendering it less useful to also consider accelerometer data.

Strengths and limitations. The algorithm of imputation of transportation modes developed in our study was relatively accurate, using a combination of GPS and GIS data processed with algorithms, travel diaries, and a phone prompted recall survey. However, the preparation of the survey and difficulties to contact some of the participants by phone proved to be a bottleneck in the data collection process, thereby causing delays between the device data collection and the survey for a median period of 17 d. This delay very likely led to memory bias in identification of activity places and transportation modes, despite the information available to prompt participant recall. Our prediction method proved to be convenient to implement and reliable compared with the results of previous studies. This prediction method can be easily adapted to a different study context, and the explanatory variables used to grow the random forest can be selected depending on the available information. In our approach, the prediction model was accurate because it was constructed on data obtained from the same population for which the predictions were made. To obtain this context specificity and the ability to select the set of locally available variables, one must conduct a preliminary data collection to adequately train the model. The duration

of this learning phase depends on the complexity of the prediction (i.e., the number of categories of the outcome and especially the number of observations in the smallest categories). Importantly, our work demonstrated that a fairly short-term learning phase is sufficient for adequate predictions. When adapting this methodology, data collection for a limited number of participants could include techniques such as a system of survey of modes and activity locations (if not too burdensome for the participants) or the SenseCam methodology. The extra burden on the participants during this learning phase could be compensated for by reducing the amount of data collected in other parts of the data collection process or reducing the number of observation days per participant.

Compared with a time unit-level prediction method, predictions at the trip level provide less detailed information. However, as trip-level data are useful in transportation sciences and behavioral sciences, a trip-level prediction method has some interesting advantages over time unit-level prediction methods. First, information on the entire trip can be used to derive predictors, such as quantification of the intratrip variability in GPS and accelerometer indicators (e.g., speed or acceleration) and summaries of the GPS data quality. Second, a trip-level method is more parsimonious in the number of predictions made. Because only one prediction per trip is required, the method allows for more participants and more observation days per participant in the model. In this RECORD GPS study, 7329 trips were observed for 236 participants and 1647 observation days. Given 12 $h \cdot d^{-1}$ of observations, a 5-s window approach would yield more than 14 million predictions, while a 1-min window approach would yield nearly 1.2 million predictions. Modeling this number of predictions would require a very high computational time. For large-scale studies with 1000 participants or more, time unit predictions would therefore require high performance computing. Finally, time unit-level models also model the data at activity places and must include activity mode categories in the model, which may reduce the quality of the overall prediction. In conclusion, we do not argue that a trip-level method is better than a unit-level method, although it does provide researchers with a valid alternative to address a large number of research questions.

A clear limit of our proposed mode detection algorithm is that its application requires data segmented into trips because the present algorithm was intended to be a complement of another trip segmentation algorithm that we commonly use in our studies (33). Moreover, it should be emphasized that the use of an algorithm of mode detection at the time unit level (e.g., min) would also require the application of a second algorithm to derive coherent information on the mode(s) used at the trip level.

This method is inappropriate for trips with multiple transportation modes. In our study, we observed 1.3% of multimodal trips (comprising more than one nonwalking mode), and we excluded them to train the prediction model. The model predicted one of the two modes for 99% of these trips. Depending on the application of this method and the proportion of multimodal trips in the study area, this limitation may be problematic and may provide an argument for the use of more advanced prediction methods that segment trips into trip stages and impute the corresponding modes (38).

Finally, it should be kept in mind that any mode prediction algorithm will have a certain error rate. In specific circumstances, researchers may want to collect more accurate data on modes for each trip. Although using SenseCam in addition to GPS receivers is useful to obtain an accurate criterion for validating algorithms, we argue that wearable cameras are too intrusive and the corresponding data are too burdensome to process to permit data collection across a large sample size. In that case, combining GPS data collection with the use of a GPS-based prompted recall mobility survey may represent a feasible option to derive accurate trip-level data.

CONCLUSIONS

This study is one of the first to use real-life data from a relatively large and diverse sample to test a prediction method for transportation modes. The approach uses a trip-level model, thereby rendering the application more convenient for subsequent application in a variety of transportation or behavioral study designs. This method could improve future data collection processes by decreasing the workload for both participants and researchers and providing relevant data to investigate the relation between transportation and health.

The authors thank the following partners from the funding institutions: Pierre Arwidson, Nadine Asconchilo, Annette Gogneau, Colette Watellier, Yasmina Baaba, Mélani Alberto, Christelle Paulo, Anne-Eole Meret-Conti, Cédric Aubouin, Benoît Kiéné, Hélène Pierre, Sophie Mazoué, John Séraphin, and Ivan Derré.

The RECORD GPS study was supported by the INPES (National Institute for Prevention and Health Education), the Ministry of Ecology (DGITM), CERTU (Centre for the Study of Networks, Transport, Urbanism, and Public constructions), ARS (Health Regional Agency) of Ile-de-France, STIF (Ile-de-France Transport Authority), the Île-de-France Regional Council, RATP (Paris Public Transport Operator), and DRIEA (Regional and Interdepartmental Direction for Equipment and Planning).

The provision of financial support does not, in any way, infer or imply endorsement of the research findings by any agency.

The authors declare no conflict of interest.

The results of the present study do not constitute endorsement by the American College of Sports Medicine.

REFERENCES

1. Auld J, Williams CA, Mohammadian AK, Nelson PC. An automated GPS-based prompted recall survey with learning algorithms. *J Transport Lett*. 2009;1(1):59–79.
2. Bastian T, Maire A, Dugas J, et al. Automatic identification of physical activity types and sedentary behaviors from triaxial accelerometer: laboratory-based calibrations are not enough. *J Appl Physiol (1985)*. 2015;118(6):716–22.
3. Bohte W, Maat K. Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: a large-scale application in the Netherlands. *Transport Res C Emer*. 2009;17(3):285–97.

4. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5–32.
5. Brondeel R, Weill A, Thomas F, Chaix B. Use of healthcare services in the residence and workplace neighbourhood: the effect of spatial accessibility to healthcare services. *Health Place*. 2014; 30:127–33.
6. Carlson JA, Jankowska MM, Meseck K, et al. Validity of PALMS GPS scoring of active and passive travel compared with SenseCam. *Med Sci Sports Exerc*. 2015;47(3):662–7.
7. Chaix B, Bean K, Daniel M, et al. Associations of supermarket characteristics with weight status and body fat: a multilevel analysis of individuals within supermarkets (RECORD study). *PLoS One*. 2012;7(3):e32908.
8. Chaix B, Kestens Y, Bean K, et al. Cohort profile: residential and non-residential environments, individual activity spaces and cardiovascular risk factors and diseases—the RECORD Cohort study. *Int J Epidemiol*. 2012;41(5):1283–92.
9. Chaix B, Kestens Y, Duncan S, et al. Active transportation and public transportation use to achieve physical activity recommendations? A combined GPS, accelerometer, and mobility survey study. *Int J Behav Nutr Phys Act*. 2014;11(1):124.
10. Chaix B, Kestens Y, Perchoux C, Karusisi N, Merlo J, Labadi K. An interactive mapping tool to assess individual mobility patterns in neighborhood studies. *Am J Prev Med*. 2012;43(4):440–50.
11. Chaix B, Meline J, Duncan S, et al. Neighborhood environments, mobility, and health: towards a new generation of studies in environmental health research. *Rev Epidemiol Sante Publique*. 2013; 61(3 Suppl):S139–45.
12. Chaix B, Meline J, Duncan S, et al. GPS tracking in neighborhood and health studies: s step forward for environmental exposure assessment, a step backward for causal inference? *Health Place*. 2013;21:46–51.
13. Chaix B, Simon C, Charreire H, et al. The environmental correlates of overall and neighborhood based recreational walking (a cross-sectional analysis of the RECORD study). *Int J Behav Nutr Phys Act*. 2014;11(1):20.
14. Chen C, Gong H, Lawson C, Bialostozky E. Evaluating the feasibility of a passive travel survey collection in a complex urban environment: lessons learned from the New York City case study. *Transport Res A Pol*. 2010;44(10):830–40.
15. de Nazelle A, Nieuwenhuijsen MJ, Anto JM, et al. Improving health through policies that promote active travel: a review of evidence to support integrated health impact assessment. *Environ Int*. 2011;37(4):766–77.
16. Ellis K, Godbole S, Chen J, Marshall S, Lanckriet G, Kerr J. Physical activity recognition in free-living from body-worn sensors. In: *Proceedings of the 4th International SenseCam and Pervasive Imaging Conference*; 18–19 November, 2013; San Diego (CA): ACM; 2013. pp. 88–9.
17. Ellis K, Godbole S, Marshall S, Lanckriet G, Staudenmayer J, Kerr J. Identifying active travel behaviors in challenging environments using GPS, accelerometers, and machine learning algorithms. *Front Public Health*. 2014;2:36.
18. Feng T, Timmermans HJ. Transportation mode recognition using GPS and accelerometer data. *Transport Res C Emer*. 2013; 37:118–30.
19. Gong H, Chen C, Bialostozky E, Lawson CT. A GPS/GIS method for travel mode detection in New York City. *Comput Environ Urban*. 2012;36(2):131–9.
20. Gong L, Morikawa T, Yamamoto T, Sato H. Deriving personal trip data from GPS data: a literature review on the existing methodologies. *Procd Soc Behv*. 2014;138:557–65.
21. Kohla B, Meschik M, Gerike R, Sammer G, Hössinger R, Unbehaun W. A new algorithm for mode detection in travel surveys: mobile technologies for activity—travel data collection and analysis. In: Rasouli S, Timmermans HJP, editors. *Mobile Technologies for Activity-Travel Data Collection and Analysis*. Hershey (PA): IGI Global; 2014. pp. 134–51.
22. Kozey-Keadle S, Libertine A, Lyden K, Staudenmayer J, Freedson PS. Validation of wearable monitors for assessing sedentary behavior. *Med Sci Sports Exerc*. 2011;43(8):1561–7.
23. Leal C, Bean K, Thomas F, Chaix B. Multicollinearity in the associations between multiple environmental features and body weight and abdominal fat: using matching techniques to assess whether the associations are separable. *Am J Epidemiol*. 2012;175(11):1152–62.
24. Liaw A, Wiener M. Classification and regression by random forest. *R News*. 2002;2(3):18–22.
25. Lin WJ, Chen JJ. Class-imbalanced classifiers for high-dimensional data. *Brief Bioinform*. 2013;14(1):13–26.
26. Perchoux C, Kestens Y, Thomas F, Van Hulst A, Thierry B, Chaix B. Assessing patterns of spatial behavior in health studies: their socio-demographic determinants and associations with transportation modes (the RECORD Cohort study). *Soc Sci Med*. 2014;119:64–73.
27. Rissel C, Curac N, Greenaway M, Bauman A. Physical activity associated with public transport use—a review and modelling of potential benefits. *Int J Environ Res Public Health*. 2012;9(7):2454–78.
28. Sahlqvist S, Song Y, Ogilvie D. Is active travel associated with greater physical activity? The contribution of commuting and non-commuting active travel to total physical activity in adults. *Prev Med*. 2012;55(3):206–11.
29. Sasaki JE, John D, Freedson PS. Validation and comparison of ActiGraph activity monitors. *J Sci Med Sport*. 2011;14(5):411–6.
30. Shen L, Stopher PR. Using SenseCam to pursue "ground truth" for global positioning system travel surveys. *Transport Res C Emer*. 2014;42:76–81.
31. Southward EF, Page AS, Wheeler BW, Cooper AR. Contribution of the school journey to daily physical activity in children aged 11–12 years. *Am J Prev Med*. 2012;43(2):201–4.
32. Stopher PR, Collins A, editors. Conducting a GPS prompted recall survey over the internet. In: *The 84th Annual Meeting of the Transportation Research Board*; 9–13 January, 2005; Washington (DC).
33. Thierry B, Chaix B, Kestens Y. Detecting activity locations from raw GPS data: a novel kernel-based algorithm. *Int J Health Geogr*. 2013;12(14).
34. Van Hulst A, Thomas F, Barnett TA, et al. A typology of neighborhoods and blood pressure in the RECORD Cohort study. *J Hypertens*. 2012;30(7):1336–46.
35. Velaga NR, Quddus MA, Bristow AL. Developing an enhanced weight-based topological map-matching algorithm for intelligent transport systems. *Transport Res C Emer*. 2009;17(6):672–83.
36. Wanner M, Gotschi T, Martin-Diener E, Kahlmeier S, Martin BW. Active transport, physical activity, and body weight in adults: a systematic review. *Am J Prev Med*. 2012;42(5):493–502.
37. Wanner M, Martin BW, Meier F, Probst-Hensch N, Kriemler S. Effects of filter choice in GT3X accelerometer assessments of free-living activity. *Med Sci Sport Exerc*. 2013;45(1):170–7.
38. Wolf J, Oliveira M, Thompson M. Impact of underreporting on mileage and travel time estimates: Results from global positioning system-enhanced household travel survey. *Transp Res Record*. 1854;2003:188–98.
39. Xu H, Wen LM, Rissel C. The relationships between active transport to work or school and cardiovascular health or body weight: a systematic review. *Asia Pac J Public Health*. 2013;25(4):298–315.

3.3 Article 3: Associations of socioeconomic status with transport-related physical activity: combining a household travel survey and accelerometer data using random forests

Journal of Transport & Health 3 (2016) 287–296

Contents lists available at ScienceDirect

Journal of Transport & Health

journal homepage: www.elsevier.com/locate/jth

Associations of socioeconomic status with transport-related physical activity: combining a household travel survey and accelerometer data using random forests

Ruben Brondeel [a,b,c,*], Bruno Pannier [d], Basile Chaix [a,b]

[a] Inserm, UMR_S 1136, Pierre Louis Institute of Epidemiology and Public Health, Nemesis team, Paris, France
[b] Sorbonne Universités, UPMC Univ Paris 06, UMR_S 1136, Pierre Louis Institute of Epidemiology and Public Health, Nemesis team, Paris, France
[c] EHESP School of Public Health, Rennes, France
[d] IPC Medical Centre, Paris, France

ARTICLE INFO

Article history:
Received 4 January 2016
Received in revised form
21 April 2016
Accepted 3 June 2016
Available online 4 July 2016

Keywords:
Socially sustainable urban transport
Moderate-to-vigorous physical activity
Transport-related physical activity
Multiple imputation
Random forests
France

ABSTRACT

Background: Socioeconomic disparities in active transport have been documented in household travel surveys. However, active transport in these studies was operationalized with self-reported measures, which poorly approximate physical activity. Unfortunately, objective accelerometer data are very expensive to obtain in large-scale travel studies.
Purpose: To benefit from a large sample and objective physical activity data, this study linked a cross-sectional household travel survey with accelerometer data from a small sample to investigate the association between socioeconomic disadvantage and the daily level of transport-related moderate-to-vigorous physical activity (T-MVPA) in an adult population (35–83 years).
Methods: Accelerometer data for participants' trips over 7 days from the RECORD GPS Study (7138 trips, 229 participants) were combined with information on participants' trips over 1 day from the Global Transport Survey (Enquête Globale Transport, EGT) (82084 trips, 21332 participants). Trip-level T-MVPA data from the RECORD sample were used to train a random forests prediction model, enabling the prediction of T-MVPA for each participant's trip from EGT. The associations between socioeconomic indicators and daily T-MVPA were analyzed with negative binomial regression models.
Results: An average time of 18.9 min (95% confidence interval: 18.6–19.2) of T-MVPA was found for these 35–83 year old adults. The education level had a positive association with T-MVPA. Household income had a negative association with T-MVPA, especially for those people without a motorized vehicle.
Conclusions: This study developed a methodology exporting precise sensor-based knowledge to a large survey sample to shed light on population-level socioeconomic disparities in transport-related physical activity.

1. Introduction

Physical activity is known to be protective for various health outcomes, such as obesity, cardiovascular health problems, depression, and certain cancers (de Nazelle et al., 2011; Wanner et al., 2012). The World Health Organization recommends 150 minutes of moderate-to-vigorous physical activity (MVPA) per week for 18 to 64 year old people (World Health Organisation, 2015), while the French recommendation is currently of 30 minutes of MVPA per day (Programme National Nutrition Santé, 2015). Transport-related physical activity is an important source of everyday physical activity (Besser and Dannenberg, 2005; Chaix et al., 2014a; Sahlqvist et al., 2012), and therefore an important target for health prevention authorities to encourage populations to reach the recommended levels of physical activity.

Socioeconomic status leads to disparities in transport-related physical activity (Beenackers et al., 2012). For example, a higher personal level of education has been associated with more minutes of walking for transport (Cerin et al., 2009), more trips with active transport modes (Cerin et al., 2009; Scheepers et al., 2013), and more cycling trips (Carse et al., 2013). In contrast of the finding that higher levels of

* Corresponding author at: Faculté de médecine Saint-Antoine, 27 rue Chaligny, UMR-S 1136, Nemesis team, 75012 Paris, France. Tel.: +33 1 44 73 86 54.
E-mail address: Ruben.Brondeel@gmail.com (R. Brondeel).

http://dx.doi.org/10.1016/j.jth.2016.06.002
2214-1405/

education are positively associated with active transport, higher income has been associated with fewer minutes of walking and less frequent trips with active modes (Cerin et al., 2009). These results are based on large-scale survey data, as large samples are needed to investigate social inequalities. However, surveys provide only self-reported measures of transport-related physical activity, thus imprecise measures of physical activity: e.g. the 'usual transportation mode' or the approximate 'number of minutes or trips with active transport modes'. These measures are subject to measurement error because people only imprecisely know the start and end times of trips and because they ignore the inactive time during trips with active transportation modes and the physically active time during trips with 'non-active' transportation modes (Steene-Johannessen et al., 2016).

Numerous studies have relied on accelerometers to derive objective measures of physical activity (Steene-Johannessen et al., 2016; Wijndaele et al., 2015). However, studies were less successful in linking transport behavior with physical activity because identifying trips with their exact start and end times is required to perform this linkage (Bohte and Maat, 2009; Brondeel et al., 2015; Chaix et al., 2014a). Unfortunately, study designs including trip recognition and accelerometer data collection often result in datasets with very precise measures but with limited sample sizes.

An alternative way to measure physical activity for a large number of participants is to rely on a large survey sample and then to estimate the intensity of physical activity based on previously established knowledge. The compendium of Ainsworth et al. (2011) enables this by providing an estimated physical activity level in 'metabolic equivalent of task' (MET) per minute for numerous activities. The researcher has to determine which category of the compendium relates best to each trip, given the transportation mode, duration of the trip, and intensity of use of certain active modes. However, despite the usefulness of the compendium, the accuracy of its predictions can be criticized. The measures in the compendium are based on findings in very restricted settings (mostly laboratories) (Ainsworth et al., 2011), and are not adaptable to the characteristics of trips in a specific city or country. Therefore, they may not reflect free-living physical activity in a specific study context.

In this study we present and apply a method that makes predictions for trips reported in a household travel survey based on the data from a GPS and accelerometer data collection conducted in the same geographical context (the Paris Ile-de-France region). The prediction of transport-related physical activity for the trips in the travel survey was based on a random forests model, which enabled us to use a high number of variables to improve the prediction. As a result of this innovative approach, the present study is the first analysis of the effect of socioeconomic status on transport-related MVPA (T-MVPA) in a large and representative dataset of 35 to 83 year old adults ($n=20730$).

The model studied in this paper is graphically presented (Fig. 1) in a directed acyclic graph (DAG). A similar model has been recently tested by Rachele et al. (2015), describing the relations between educational level, occupational status, household income, neighborhood disadvantage and the most frequently used transportation mode. In our study, the model was applied to T-MVPA instead of the self-reported 'most frequently used transportation mode' and interaction terms were added compared to this previous work, as indicated in the DAG by the dotted arrows (using arrow to arrow notation as suggested by Weinberg (2007)). In this model, education and household income were examined separately instead of a single socioeconomic status variable, since these two dimensions had an opposite effect on walking for transport in previous studies (Cerin et al., 2009; Turrell et al., 2014). The hypothesized interactions are based on findings of social exclusion from transport research. Socioeconomic disadvantage and transport disadvantage (e.g., spatial accessibility to public transport, ownership of a car, or walkability of streets) were found to interact and together amplify social inequalities in the number of trips per individual (Lucas, 2012).

This study aimed to investigate the associations between socioeconomic disadvantage, transport disadvantage, and transport-related physical activity for older adults (35–83 years old). It expands previous literature by relying on a precise measure of transport-related physical activity and by exploring interactions between various forms of disadvantage. It also describes a novel methodology combining the strengths of a large population dataset with precise sensor-based data (data integration approaches) that advances the field and can be applied to various research questions.

2. Methods

2.1. The global transport survey

The Global Transport Survey ('Enquête Global Transport', EGT) is a household travel survey conducted every 10 years in Île-de-France, the French capital region. The main purpose of the survey is to inform local authorities and transport planners on the mobility and transport use in Île-de-France. The latest EGT-survey was conducted in 2010 by two French transport institutions: the Ile-de-France Transport Authority (STIF) and the Regional and Interdepartmental Direction for Equipment and Planning (DRIEA).

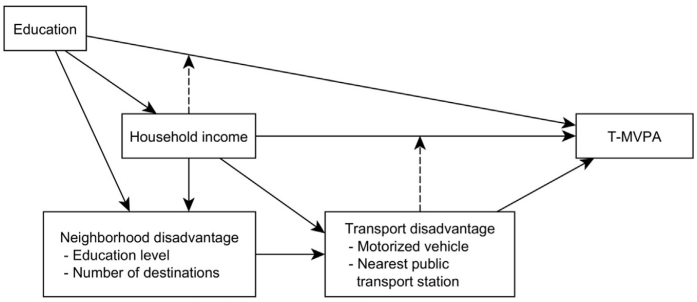


Fig. 1. Directed acyclic graph for the associations between socioeconomic indicators, neighborhood disadvantage, transport disadvantage, and transport-related moderate-to-vigorous physical activity (T-MVPA).

During face-to-face interviews with members of randomly selected households, data were collected for all the trips made during the day before the interview. We selected participants between 35 and 83 years old for the present study, yielding 82084 trips made by 21332 people. Limiting the EGT-dataset to the people within this age range prevented interpolations of physical activity outside of the age range of the RECORD Study.

2.2. The RECORD GPS Study

As previously described in detail (Brondeel et al., 2014; Chaix et al., 2014a), the participants in the RECORD Study (Residential Environment and CORonary heart Disease) were recruited during preventive health checkups in 2007–2008, and were born in 1928–1978. Every participant residing in 112 pre-selected municipalities of the Ile-de-France Paris region at baseline from the administrative files of the IPC Medical Center was invited at the health center (Chaix et al., 2012a; Van Hulst et al., 2012). The selected municipalities of the region Ile-de-France included a broad range of municipalities in median household income. In the second wave of the study (2011–2012) (Chaix et al., 2012b; Chaix et al., 2012c; Leal et al., 2012; Perchoux et al., 2014), 410 participants were invited to enter the RECORD GPS Study (Chaix et al., 2014a). Participants wore a BT-Q1000XT GPS (QStarz) and a GT3X+ accelerometer (The Actigraph) on the right hip with a dedicated elastic belt, for the recruitment day and 7 additional days, all day long from wake up to bedtime. The participants had to fill out a travel diary by reporting their activity places over the 7-8 days, each time with arrival and departure times. The GPS data were collected every 5 seconds. After linear interpolation of the missing data, the GPS data were analyzed with an algorithm (ArcGIS Python script) that identified all of the activity locations of the participants (any activity at a stationary location) from the accumulation of GPS points over 7 days (Thierry et al., 2013). Based on these outputs of the algorithm, the Mobility Web Mapping application was then used to visualize the activity and transport patterns on a map per participant per day. The Mobility Web Mapping application was designed by the University of Montreal. The application was used to survey the participants on the activity performed at each visited location and on the modes used in each trip. The survey operator could report activity locations and trips undetected by the algorithm and could modify/remove detected visits to locations that were inaccurate or incorrect. This procedure resulted in the identification of 7138 trips for 229 participants. Written informed consent was obtained from all participants. The RECORD GPS Study was approved by the French Data Protection Authority.

Participants in EGT had a considerable lower education and had a lower household income than the participants in RECORD (see Table 1). Supplementary material S1 provides a comparison of these demographic characteristics between the RECORD sample, the EGT sample and the background population (35 to 83 year old people in Ile-de-France). This comparison supports the hypothesis that the EGT sample represents the background population better than the RECORD sample. The EGT sample included more women, more young people and less people from the inner city.

2.3. Measures

All the dependent and independent variables used in the study are summarized in Table 2.

From the raw accelerometer data, the counts per minute were extracted in ActiLife 5.1. No missing data was allowed within a trip or all data were considered to be missing. There was no minimal wear time per day required. A minute of MVPA was defined as a minute during which a vector magnitude higher than 2690 (Sasaki et al., 2011) was recorded, based on the tri-axial GT3X+ accelerometer data in the RECORD GPS study. Accelerometers worn at the hip underestimate physical activity during biking trips. Therefore, all minutes during biking trips were considered as minutes of T-MVPA. This and other limitations of this measure are discussed in the Discussion section.

The following variables were defined both in the RECORD GPS and in the EGT databases (in addition to age and gender). Self-reported household income was coded as a continuous variable. Three educational levels were considered: 'no diploma of secondary education', 'diploma of secondary education or lower tertiary education', and 'diploma of higher tertiary education'. Working situation was categorized as employed, unemployed, retired, or other. Participants indicated whether a bike, a motorbike, a car, a motorized vehicle (the combination of the two previous ones) was available in their household. They indicated whether they had a public transport pass. The distance to the nearest public transport station was the distance from the residence to the nearest bus, tram, metro, or train station following the street network. Residential neighborhoods were defined as 1 km buffers around the residence following the street network; corresponding to a 10-to-15 min walk that reflects the local resources easily accessible within a 'walkable' distance (Brondeel et al., 2014; Chaix et al., 2014b; Frank et al., 2005; Karusisi et al., 2014; Troped et al., 2010; Villanueva et al., 2014). The information needed for alternative definitions such as the perceived neighborhood (Vallée et al., 2015) or the activity space (Matthews and Yang, 2013) was not available. The neighborhood educational level was the percentage of residents with a higher University degree (2010 Census of the National Institute of Statistics and Economic Studies (INSEE)) with census participants geocoded at the building level. The number of destinations in the residential neighborhood was the total number of services of different types (shops, administrative services, leisure facilities, etc.) from the 2011 Permanent Facilities Database of INSEE. We also calculated the number of street intersections (National Geographic Institute data), the area size of parks (Ile-de-France Urbanization Institute), and the population density (2010 Census) in each neighborhood. All these contextual variables were also calculated at the departure and arrival of each trip (see Table 2). ArcGIS (v10.3) automated using Python (v2.7) was used for the geographical analyses.

Based on the RECORD and EGT mobility surveys, the following variables were determined at the trip level: transportation mode, trip duration, time and day of the trip, distance covered and speed.

2.4. Statistical analysis

An overview of the dependent and independent variables in the prediction model, the multiple imputation model and the main regression model is provided in Table 2.

The RECORD GPS data were used to train a random forests prediction model for T-MVPA (see explanatory variables in Table 2) with 1000 trees and a random selection of 16 variables at each knot. The random forests model was grown with the 'randomForest' package (Liaw and Wiener, 2002) in R. Based on the prediction model and on the comparable prediction variables in EGT, we predicted the number of minutes of T-MVPA for each trip in the EGT dataset. The predicted values were summed up per day, resulting in a daily time of T-MVPA in minutes per person.

The associations between the disadvantage variables and the predicted T-MVPA time were analyzed with a negative binomial regression model using the 'MASS' package in R (Venables and Ripley, 2002). The time variable could be considered as continuous and analyzed with a regular linear regression. However, given the left-censored distribution of the variable (i.e. 0 as the absolute minimum and many observation equal 0 or close to 0), we preferred the negative binomial regression that is adapted to

Table 1
Educational level and household income of the EGT sample and the RECORD GPS sample.

	EGT[a]	RECORD
Educational level		
No diploma of secondary education (%)	40	28
Diploma of secondary education or lower tertiary education[b] (%)	26	30
Diploma of higher tertiary education[c] (%)	33	42
Household income (mean)	3377	4393
Sample size	21,332	229

[a] EGT: 'Enquête globale transport'.
[b] Lower tertiary education: two years or less of University education.
[c] Higher tertiary education: three years or more of University education.

count variables with overdispersion (a high variance compared to the mean). There were missing values on 8 independent variables for 24% of the respondents, of which 6% had more than 1 missing value. Therefore, multiple imputations were performed with the 'mice' package in R (Van Buuren and Groothuis-Oudshoorn, 2011). This method enabled us to analyze the data under the hypothesis that the unobserved values are randomly distributed given the observed data (Little and Rubin, 1989). To account for the non-linear and interaction effects in the imputation process, random forests methods were also used for the multiple imputations of explanatory variables in EGT. Five

Table 2
Overview of the variables used in the negative binomial regression model (NB), the multiple imputation model (MI) and the random forests prediction model (RF).

	NB	MI	RF
T-MVPA			
Daily minutes of T-MVPA[a]	X	X	
Minutes of T-MVPA per trip[a]			X
Socioeconomic disadvantage			
Household income[b]	X	X	X
Personal education level[b]	X	X	X
Transport disadvantage			
Street network distance to nearest public transport station from residence[c]	X	X	X
Street network distance to nearest train station[c]			X
Street network distance to nearest metro station[c]			X
Street network distance to nearest tram station[c]			X
Street network distance to nearest bus station[c]			X
A motorized vehicle available in the household[b]	X	X	X
A car available in the household[b]			X
A motorbike available in the household[b]			X
In possession of a public transport pass[b]		X	X
Other personal variables			
Age[b]	X	X	X
Gender[b]	X	X	X
Work situation (employed, unemployed, retired, other)[b]	X	X	X
Other residential neighborhood characteristics			
Educational level in the residential neighborhood[d]	X	X	X
Number of destinations in the residential neighborhood[d]	X	X	X
Number of intersections in the area[d]		X	X
Area size of parks in the area[d]		X	X
Population density in the area[d]		X	X
Address located in Paris, or in the counties adjacent to the city center, or in the counties non-adjacent to the city center		X	X
Personal daily transport behavior			
Minutes in transport per day[e]		X	
Minutes in transport walking per day[e]		X	
Minutes in transport by bike per day[e]		X	
Minutes in transport by private motorized vehicle per day[e]		X	
Minutes in public transport per day[e]		X	
Number of trips per day[e]		X	
Number of trips by walking per day[e]		X	
Number of trips by bike per day[e]		X	
Number of trips by private motorized vehicle per day[e]		X	
Number of trips by public transport day[e]		X	
Trip characteristics			
Transportation mode[f]			X
Duration of the trip in minutes[f]			X
Time of the day at departure[f]			X
Day of the week at departure[f]			X
Rush hour or not at departure: from 8am to 11am and from 4 pm to 7 pm[f]			X
Straight-line distance from departure address to arrival address[f]			X
Speed based on duration and straight-line distance[f]			X
Trip departure and arrival location characteristics (2 separate set of variables)			
Distance to nearest train station[c]			X
Distance to nearest metro station[c]			X
Distance to nearest tram station[c]			X
Distance to nearest bus station[c]			X
Distance to nearest public transport station[c]			X
Educational level in the area[d]			X
Number of intersections in the area[d]			X
Number of destinations in the area[d]			X
Area size of parks in the area[d]			X
Population density in the area[d]			X

Table 2 (*continued*)

	NB	MI	RF
T-MVPA			
Address located in the city center or not (i.e., in Paris as opposed the other parts of Ile-de-France Region)			X
Day of the EGT mobility survey: week or weekend		X	

[a] Accelerometry information in RECORD or predicted time in EGT.
[b] RECORD and EGT questionnaire.
[c] Shortest street network distance determined with ArcGIS from the residence or from the departure/arrival of each trip geocoded at the address level in RECORD or at the center of a 100 m square in EGT.
[d] The area around the residence or departure or arrival point of each trip was defined with ArcGIS as a 1 km buffer following the street network, and information was aggregated at the level of this buffer.
[e] Information from the mobility survey in EGT.
[f] Information from the mobility survey in RECORD and in EGT; T-MVPA: transport-related moderate-to-vigorous physical activity.

imputation datasets were constructed though an iterative process using 100 trees for every imputed variable at each iteration. One imputed dataset was retained every five iterations (25 iterations overall). The convergence of the imputations was checked with plots of the means and standard deviations over the iterations.

In the analysis of the determinants of T-MVPA, the interaction terms of interest were plotted in graphs based on the coefficients and on the variance-covariance matrix from the regression model. The code for these plots was based on the library 'effects' in R (Fox, 2003), but adapted to the negative binomial regression. The script for all the analyses with R (v3.2.2) (R Core Team, 2014) can be found in Supplementary material S2.

3. Results

The random forests prediction model for T-MVPA was very accurate, predicting 67% of the variance in T-MVPA in RECORD. The three most important variables in predicting trip-level T-MVPA were transportation mode, distance and duration of the trip (see Supplementary material S2). Applying this model to the EGT trips and summing up the predicted minutes of T-MVPA by day, we found a mean predicted time of T-MVPA of 18.9 minutes (95% confidence interval (CI): 18.6–19.2) per participant per day (interquartile range: 5, 28). The mean T-MVPA times for the levels of education 'no diploma of secondary education', 'diploma of secondary education or lower tertiary education', and 'diploma of higher tertiary education' were respectively of 17.5, 18.5, and 21.0 min per day (descriptive data, unadjusted). Household income was negatively associated with the daily T-MVPA time (Incidence Risk Ratio=0.98 for a change in income of 1000€, 95% CI: 0.97–0.99). Regarding transportation disadvantage, participants who had access to a motorized vehicle (i.e., a car or motorbike) in the household had a mean daily T-MVPA time of only 16.8 minutes while their counterparts who had no vehicle had 28.9 min of T-MVPA per day. The distance to the nearest public transport station was negatively associated with the daily T-MVPA time (Incidence Risk Ratio=0.72 for a change in distance of 1 km, 95% CI: 0.65-0.78). No difference between men and women was noted. Finally, older people had slightly less daily T-MVPA (Incidence Risk Ratio=0.98 for a change in age of 10 years, 95% CI: 0.97–1.00).

The results of the multiple negative binomial regression (Table 3) confirmed the bivariate analyses, while adding nuance by introducing interaction effects. Figs. 2 and 3 represent two interaction effects. Household income had a negative association with T-MVPA for all three categories of education level (Fig. 2). The interaction effect was statistically significant (Wald-test for pooled regression results (Van Buuren and Groothuis-Oudshoorn, 2011): $P=0.041$), but there was no clear gradient in the strength of the association between income and T-MVPA between the different education levels. Furthermore, household income had a negative association with T-MVPA for both those with and without a motorized vehicle available in the household (Fig. 3). However, the association was much stronger for those without a motorized vehicle.

The distance to the nearest public transport station had a negative association with T-MVPA for all levels of income. The interaction effect with income was small and does not alter the interpretation of the results. Two interaction effects between education level and transport disadvantage (availability of a motorized vehicle and access to public transport) were tested. Including these into the model did not change the interpretation of the results nor did it improve the model in statistical terms ($P=0.180$). For the sake of parsimony, these two interaction terms were excluded from the final model.

To facilitate the interpretation of the associations between socio-economic factors and T-MVPA, Table 4 provides information on the associations between educational level and household income on one hand and the mean number of trips and the mean duration of trips by transportation mode on the other hand. From these descriptive data, the positive association of educational level with T-MVPA may be attributable to some extent to the number of walking and public transport trips. Higher educated people had more walking and public transport trips. This is attenuated but not completely counterbalanced by the longer duration of walking and public transport trips of lower educated people. The negative association of income with T-MVPA may also be attributable to some extent to the number of walking and public transport trips and to the duration of the walking trips. People with higher income had less and shorter walking trips, and less public transport trips. From the descriptive data, biking trips had little or no impact on both associations.

Table 3
Associations between socioeconomic or transport disadvantage and daily T-MVPA (negative binomial regression).

Predictor	IRR	95% CI
Socioeconomic disadvantage		
Education level		
No diploma of secondary education	1.00	Referent
Diploma of secondary education or lower tertiary education [a]	1.06	1.01, 1.10
Diploma of higher tertiary education [b]	1.12	1.07, 1.17
Household income (/1000 euros)	0.97	0.94, 1.00
Interaction Education –Income		
No secondary education - income	1.00	Referent
Secondary or lower tertiary education - income	0.99	0.96, 1.02
Higher tertiary education - income	1.00	0.98, 1.03
Transport disadvantage		
Motorized vehicle available in household		
No motorized vehicle	1.00	Referent
Motorized vehicle	0.65	0.61, 0.68
Nearest public transport (km)	1.01	0.93, 1.11
Interactions Socioeconomic - Transport		
Motorized vehicle - income		
No motorized vehicle	1.00	Referent
Motorized vehicle	1.02	1.00, 1.05
Nearest public transport - income	0.95	0.90, 1.00
Neighborhood disadvantage		
Educational level	1.26	1.11, 1.43
Number of destinations (/1000)	1.12	1.10, 1.14
Other		
Age (10y)	0.96	0.94, 0.98
Gender		
Female	1.000	Referent
Male	1.02	0.99, 1.05
Work situation		
Employed	1.000	Referent
Unemployed	1.02	0.95, 1.09
Retired	1.09	1.03, 1.14
Other	0.98	0.92, 1.04
(intercept)	24.19	22.71, 25.77

Abbreviations: CI, confidence interval; IRR, incidence rate ratio; MVPA, moderate-to-vigorous physical activity.
[a] Lower tertiary education: two years or less of University education.
[b] Higher tertiary education: three years or more of University education.

4. Discussion

4.1. Main results

Our study suggests that transport-related physical activity is a major source of physical activity for the population in the Ile-de-France region. On average, the participants had 18.9 min of MVPA per day. The international recommendation of 30 min of MVPA per day (including all sources of physical activity) was attained by 23% of participants through their transport behavior alone.

The model showed a negative association of household income with T-MVPA and a positive relation of educational level with T-MVPA. Understanding the mechanisms underlying these associations is very important to efficiently target subpopulations in physical activity interventions. It has been argued that lower educated people have symbolic and affective predispositions that promote car use over active transport (e.g., car use perceived as a marker of wealth) (Beirão and Sarsfield Cabral, 2007; Scheepers et al., 2013). Instead of psychological explanations, other studies have established a link between lower educational levels and material obstacles to healthy behavior including physical activity (Brunello et al., 2016; Chaix et al., 2014b). These obstacles are situated within diverse domains of the social life: e.g., the residential environment (e.g. walking possibilities) or the workplace (e.g. parking facilities at work) or the local organization of transport (e.g. bus frequency) (Dalton et al., 2013; Delbosc and Currie, 2011). Further research is needed to fully understand the motivations and obstacles of people with a lower level of education and a high income to participate in active transport, and to confirm the observed patterns of associations in other geographical contexts and other populations such as children going to school or younger adults. However, the results clearly show that education and income should be considered separately when studying transport-related physical activity or mobility in general, instead of using a combined measure of socioeconomic status.

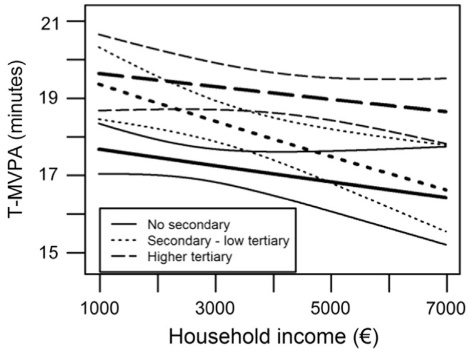


Fig. 2. Moderating effect of household income on the relationship of personal education level to daily minutes of transport-related moderate-to-vigorous physical activity (T-MVPA). Confidence intervals and predicted values are represented for each category of education level for all levels of household income within the studied range.

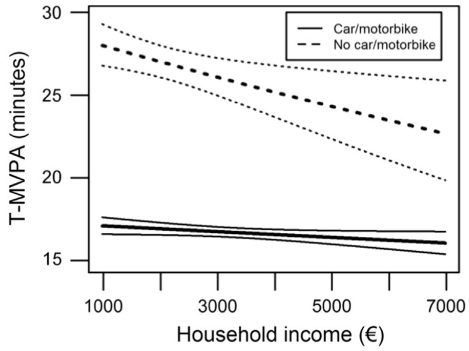


Fig. 3. Moderating effect of the availability of a motorized vehicle in the household on the relationship of household income to daily minutes of transport-related moderate-to-vigorous physical activity (T-MVPA). Confidence intervals and predicted values are represented for the availability of a vehicle (yes/no) for all levels of household income within the studied range.

Table 4
Mean number of trips and mean duration of trips in the EGT sample and in the RECORD GPS sample.

	Mean number of trips per person				Mean duration of trips (min)			
	W	B	PM	PT	W	B	PM	PT
Education[a]								
Level 1	1.2	0.0	1.8	0.4	14.2	25.0	22.4	52.8
Level 2	1.2	0.0	2.2	0.5	13.1	24.6	21.9	50.8
Level 3	1.4	0.1	2.0	0.7	12.5	20.1	22.5	45.9
Household income								
Less than 2000 €	1.5	0.0	1.3	0.7	14.0	19.8	22.8	49.8
2000 to 4000 €	1.2	0.0	2.1	0.5	13.3	23.6	22.0	50.4
4000 € or more	1.2	0.1	2.3	0.5	12.2	22.2	22.7	47.1

W: walking; B: biking; PM: private motorized (car/motorbike); PT: public transport; min: minutes.

[a] Education: No diploma of secondary education, Diploma of secondary education or lower tertiary education (2 years or less of University education); Diploma of higher tertiary education (three years or more of University education).

The availability of a motorized vehicle largely moderated the association between household income and T-MVPA. The negative association of household income with T-MVPA was much stronger within the group of people with no motorized vehicle available. This might reflect the influence of the distance from the residence to important places such as work or services. For the higher income groups, this distance is typically shorter than for the lower income groups. So, people with long trips to cover and no accessibility to a motorized vehicle are constrained to use more active transport modes, including public transport.

4.2. Strengths and limitations

Hopefully, technical advances will enable researchers in the future to both assess the trips of study participants and objectively measure physical activity in these trips for large samples of people. Until then, we believe that predicting transport-related physical activity (here T-MVPA) by applying precise knowledge derived from sensor data to large survey datasets has several advantages over the use of approximate self-reported measures (e.g., on the use of active transport) combined with information from a physical activity compendium. Compared to approximate self-reported measures, T-MVPA enables the comparison to the WHO health recommendations; it allows one to take into account the specific intensity of physical activity of active modes in the study territory of interest; and it includes the physical activity during trips with 'non-active' modes (e.g., walk to or from a car, use of stairs in public transport). This is especially important in regions with a relatively high use of public transport, such as in the French capital region Île-de-France. Daily T-MVPA is a useful variable from a public health perspective since it encompasses the influences of the transportation mode, the number of trips, and the duration of trips instead of just one of these indicators. Moreover, compared to the use of a compendium, the prediction of T-MVPA is based on sensor data from the same geographical context. Finally, the use of an underlying prediction model enables the use of numerous variables to individualize the physical activity intensity for each participant's profile. Therefore, it can be expected that the predictions are of much better quality than if standard compendium values were applied to trips, even though a comparative study is needed to examine this.

This study provided a sophisticated model including direct, moderated, and mediated associations between socioeconomic disadvantage and T-MVPA. Especially the moderated associations presented in this study show the need for a conceptual thinking that goes beyond basic associations applied to everyone when investigating social disparities in T-MVPA. Unfortunately, we could not test other variables of the built environment (e.g., the width of sidewalks) than those that were examined, or related to other individual dimensions (time available, behavioral preferences, etc.) to further understand and explain the associations between education, income, and T-MVPA.

Combining two datasets from the same geographical setting – a large-scale survey and a smaller dataset with detailed sensor measures – could be a pragmatic approach to address a large range of research questions where large data collections with detailed measures are too expensive. Given a good prediction model with variables available in both datasets, this method could provide a relatively inexpensive option for research questions where large-scale survey data are necessary (e.g., when investigating population disparities as in our case). Further methodologic work is needed to evaluate different machine learning methods. The random forests method was preferred for this study, since it explained a high percentage of the variation (67%) compared to two other machine learning methods: support vector machines (42%, using the 'svm' function in the R package 'e1071' (Meyer et al., 2015)) and neural networks (45%, using the 'mlp' function in the R package 'RSNNS' (Bergmeir and Benitez, 2012)). Secondly, the random forests method does not rely on parameters of the distribution of the outcome variable. Therefore, it cannot predict values outside the range of the input data, which is particularly important for a left-censored variable (i.e. 0 as a strict minimum value) such as T-MVPA. A limitation of the random forests method is its complexity, making it hard to interpret the relations between the predictive variables and the outcome.

The cut point for MVPA used in this study is not without limitation. The cut point aims to identify body movements that require an energy expenditure of three MET (metabolic equivalent of task) (Sasaki et al., 2011). The cut point is not age-specific, whereas research has found that the energy expenditure is higher for older people than younger people when performing the same physical task (Hall et al., 2013). The cut point will therefore have to be age-specific in future research. Also, the cut point has been established during laboratory tests and might therefore poorly correspond to three METs in free-living conditions.

An important limit to this study is the lack of a total daily MVPA measure (e.g., including leisure physical activity). A lack of transport-related physical activity could be compensated by leisure-time physical activity. And even though this compensation mechanism was documented neither by Hearst et al. (2013) for walking time nor by Sahlqvist (2012) for self-reported physical activity, more studies in this domain are needed.

Finally, for biking trips, an accelerometer at the hip usually underestimates T-MVPA. Therefore, we had to use an estimate of biking physical activity from the compendium of Ainsworth (2011). A drawback of this is that all minutes of biking trips were considered to be physically active, disregarding stops over the way. The impact on the results is probably small with around 6.2% of T-MVPA obtained from cycling in this population. A slight overestimation of this small share of T-MVPA probably only led to a minor overestimation of the daily T-MVPA. For studies with cycling as the focus, other types of accelerometer devices (such as the VitaMove system used in the RECORD MultiSensor Study) or other ways to carry the accelerometer are recommended.

5. Conclusions

This study is, to our knowledge, the first to use a large dataset to estimate the association between socioeconomic disadvantage and T-MVPA. It gives insights on the relationships between socioeconomic disadvantage and daily transport-related physical activity, which is a relatively large part of the daily physical activity of the adult population in the Ile-de-France region. An important finding for future interventions on active transport is that both the expected positive association with education and a negative association with income were document. More research is needed to understand the exact motivations and obstacles leading to social disparities in transport-related physical activity.

Acknowledgements

The RECORD GPS Study was supported by INPES (National Institute for Prevention and Health Education); the Ministry of Ecology (DGITM); CEREMA (Centre for the Study of and Expertise on Risks, the Environment, Mobility, and Urbanism); ARS (Health Regional Agency) of Ile-de-France; STIF (Ile-de-France Transport Authority); the Ile-de-France Regional Council; RATP (Paris Public Transport Operator); and DRIEA (Regional and Interdepartmental Direction for Equipment and Planning). The authors thank the following partners

from the funding institutions: Pierre Arwidson, Nadine Asconchilo, Annette Gogneau, Colette Watellier, Yasmina Babaa, Laurent Jardinier, Tristan Guilloux, Mélani Alberto, Christelle Paulo, Anne-Eole Meret-Conti, Cédric Aubouin, Benoît Kiéné, Hélène Pierre, Sophie Mazoué, John Séraphin, and Ivan Derré.

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.jth.2016.06.002.

References

Ainsworth, B.E., Haskell, W.L., Herrmann, S.D., Meckes, N., Bassett Jr., D.R., Tudor-Locke, C., Greer, J.L., Vezina, J., Whitt-Glover, M.C., Leon, A.S., 2011. 2011 Compendium of Physical Activities: a second update of codes and MET values. Med. Sci. Sport Exer 43, 1575–1581.
Beenackers, M.A., Kamphuis, C.B., Giskes, K., Brug, J., Kunst, A.E., Burdorf, A., van Lenthe, F.J., 2012. Socioeconomic inequalities in occupational, leisure-time, and transport related physical activity among European adults: a systematic review. Int. J Behav. Nutr. Phys. Act. 9, 116.
Beirão, G., Sarsfield Cabral, J.A., 2007. Understanding attitudes towards public transport and private car: A qualitative study. Transp. Policy 14, 478–489.
Bergmeir, C., Benitez, J.M., 2012. Neural networks in R Using the Stuttgart Neural Network Simulator: RSNNS. J Stat. Softw. 46, 1–26.
Besser, L.M., Dannenberg, A.L., 2005. Walking to public transit: steps to help meet physical activity recommendations. Am. J Prev. Med. 29, 273–280.
Bohte, W., Maat, K., 2009. Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: a large-scale application in the Netherlands. Transp. Res. C-Emer 17, 285–297.
Brondeel, R., Pannier, B., Chaix, B., 2015. Using GPS, GIS, and Accelerometer Data to Predict Transportation Modes. Med. Sci. Sport Exer 47, 2669–2675.
Brondeel, R., Weill, A., Thomas, F., Chaix, B., 2014. Use of healthcare services in the residence and workplace neighbourhood: The effect of spatial accessibility to healthcare services. Health Place 30C, 127–133.
Brunello, G., Fort, M., Schneeweis, N., Winter-Ebmer, R., 2016. The causal effect of education on health: what is the role of health behaviors? Health Econ. 25, 314–336.
Carse, A., Goodman, A., Mackett, R.L., Panter, J., Ogilvie, D., 2013. The factors influencing car use in a cycle-friendly city: the case of Cambridge. J. Transp. Geogr. 28, 67–74.
Cerin, E., Leslie, E., Owen, N., 2009. Explaining socio-economic status differences in walking for transport: an ecological analysis of individual, social and environmental factors. Soc. Sci. Med. 68, 1013–1020.
Chaix, B., Bean, K., Daniel, M., Zenk, S.N., Kestens, Y., Charreire, H., Leal, C., Thomas, F., Karusisi, N., Weber, C., Oppert, J.-M., Simon, C., Merlo, J., Pannier, B., 2012a. Associations of supermarket characteristics with weight status and body fat: a multilevel analysis of individuals within supermarkets (RECORD Study). Plos One 7, e32908.
Chaix, B., Kestens, Y., Bean, K., Leal, C., Karusisi, N., Meghiref, K., Burban, J., Fon Sing, M., Perchoux, C., Thomas, F., Merlo, J., Pannier, B., 2012b. Cohort Profile: Residential and non-residential environments, individual activity spaces and cardiovascular risk factors and diseases–The RECORD Cohort Study. Int. J Epidemiol. 41, 1283–1292.
Chaix, B., Kestens, Y., Duncan, S., Merrien, C., Thierry, B., Pannier, B., Brondeel, R., Lewin, A., Karusisi, N., Perchoux, C., Thomas, F., Meline, J., 2014a. Active transportation and public transportation use to achieve physical activity recommendations? A combined GPS, accelerometer, and mobility survey study. Int. J Behav. Nutr. Phy 11, 124.
Chaix, B., Kestens, Y., Perchoux, C., Karusisi, N., Merlo, J., Labadi, K., 2012c. An interactive mapping tool to assess individual mobility patterns in neighborhood studies. Am. J Prev. Med. 43, 440–450.
Chaix, B., Simon, C., Charreire, H., Thomas, F., Kestens, Y., Karusisi, N., Vallee, J., Oppert, J.M., Weber, C., Pannier, B., 2014b. The environmental correlates of overall and neighborhood based recreational walking (a cross-sectional analysis of the RECORD Study). Int. J Behav. Nutr. Phys. 11, 20.
Dalton, A.M., Jones, A.P., Panter, J.R., Ogilvie, D., 2013. Neighbourhood, route and workplace-related environmental characteristics predict adults' mode of travel to work. Plos One 8, e67575.
de Nazelle, A., Nieuwenhuijsen, M.J., Anto, J.M., Brauer, M., Briggs, D., Braun-Fahrlander, C., Cavill, N., Cooper, A.R., Desqueyroux, H., Fruin, S., Hoek, G., Panis, L.I., Janssen, N., Jerrett, M., Joffe, M., Andersen, Z.J., van Kempen, E., Kingham, S., Kubesch, N., Leyden, K.M., Marshall, J.D., Matamala, J., Mellios, G., Mendez, M., Nassif, H., Ogilvie, D., Peiro, R., Perez, K., Rabl, A., Ragettli, M., Rodriguez, D., Rojas, D., Ruiz, P., Sallis, J.F., Terwoert, J., Toussaint, J.F., Tuomisto, J., Zuurbier, M., Lebret, E., 2011. Improving health through policies that promote active travel: a review of evidence to support integrated health impact assessment. Environ. Int. 37, 766–777.
Delbosc, A., Currie, G., 2011. Transport problems that matter – social and psychological links to transport disadvantage. J. Transp. Geogr. 19, 170–178.
Fox, J., 2003. Effect displays in R for generalised linear models. J Stat. Softw. 8, 1–27.
Frank, L.D., Schmid, T.L., Sallis, J.F., Chapman, J., Saelens, B.E., 2005. Linking objectively measured physical activity with objectively measured urban form: findings from SMARTRAQ. Am. J Prev. Med. 28, 117–125.
Hall, K.S., Howe, C.A., Rana, S.R., Martin, C.L., Morey, M.C., 2013. METs and accelerometry of walking in older adults: standard versus measured energy cost. Med. Sci. Sport. Exerc. 45, 574–582.
Hearst, M.O., Sirard, J.R., Forsyth, A., Parker, E.D., Klein, E.G., Green, C.G., Lytle, L.A., 2013. The relationship of area-level sociodemographic characteristics, household composition and individual-level socioeconomic status on walking behavior among adults. Transp. Res. Part A Policy Pract. 50, 149–157.
Karusisi, N., Thomas, F., Meline, J., Brondeel, R., Chaix, B., 2014. Environmental conditions around itineraries to destinations as correlates of walking for transportation among adults: the RECORD cohort study. Plos One 9, e88929.
Leal, C., Bean, K., Thomas, F., Chaix, B., 2012. Multicollinearity in the associations between multiple environmental features and body weight and abdominal fat: using matching techniques to assess whether the associations are separable. Am. J Epidemiol. 175, 1152–1162.
Liaw, A., Wiener, M., 2002. Classification and regression by randomforest. R News 2, 18–22.
Little, R.J.A., Rubin, D.B., 1989. The Analysis of social science data with missing values. Sociol. Methods Res. 18, 292–326.
Lucas, K., 2012. Transport and social exclusion: Where are we now? Transp. Policy 20, 105–113.
Matthews, S.A., Yang, T.C., 2013. Spatial polygamy and contextual exposures (SPACEs): promoting activity space approaches in research on place and health. Am. Behav. Sci. 57, 1057–1081.
Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., 2015. e1071: Misc functions of the department of statistics, probability theory group (Formerly: E1071). TU Wien. R package version 1, 6–7.
Perchoux, C., Kestens, Y., Thomas, F., Van Hulst, A., Thierry, B., Chaix, B., 2014. Assessing patterns of spatial behavior in health studies: Their socio-demographic determinants and associations with transportation modes (the RECORD Cohort Study). Soc. Sci. Med. 119, 64–73.
Programme National Nutrition Santé, 2015. Manger Bouger - Que veut dire bouger?.
R Core Team, 2014. R: A language and environment for statistical computing, in: R Foundation for Statistical Computing, V., Austria (Ed.).
Rachele, J.N., Kavanagh, A.M., Badland, H., Giles-Corti, B., Washington, S., Turrell, G., 2015. Associations between individual socioeconomic position, neighbourhood disadvantage and transport mode: baseline results from the HABITAT multilevel study. J Epidemiol. Commun. Health.
Sahlqvist, S., Song, Y., Ogilvie, D., 2012. Is active travel associated with greater physical activity? The contribution of commuting and non-commuting active travel to total physical activity in adults. Prev. Med. 55, 206–211.
Sasaki, J.E., John, D., Freedson, P.S., 2011. Validation and comparison of ActiGraph activity monitors. J. Sci. Med. Sport 14, 411–416.
Scheepers, E., Wendel-Vos, W., van Kempen, E., Panis, L.I., Maas, J., Stipdonk, H., Moerman, M., den Hertog, F., Staatsen, B., van Wesemael, P., Schuit, J., 2013. Personal and environmental characteristics associated with choice of active transport modes versus car use for different trip purposes of trips up to 7.5 km in The Netherlands. Plos One 8, e73105.
Steene-Johannessen, J., Anderssen, S.A., van der Ploeg, H.P., Hendriksen, I.J., Donnelly, A.E., Brage, S., Ekelund, U., 2016. Are self-report measures able to define individuals as physically active or inactive? Med. Sci. Sport Exer 48, 235–244.
Thierry, B., Chaix, B., Kestens, Y., 2013. Detecting activity locations from raw GPS data: a novel kernel-based algorithm. Int. J Health Geogr., 12.
Troped, P.J., Wilson, J.S., Matthews, C.E., Cromley, E.K., Melly, S.J., 2010. The built environment and location-based physical activity. Am. J Prev. Med. 38, 429–438.
Turrell, G., Hewitt, B., Haynes, M., Nathan, A., Giles-Corti, B., 2014. Change in walking for transport: a longitudinal study of the influence of neighbourhood disadvantage and individual-level socioeconomic position in mid-aged adults. Int. J Behav. Nutr. Phys. Act. 11, 151.
Vallée, J., Le Roux, G., Chaix, B., Kestens, Y., Chauvin, P., 2015. The 'constant size neighbourhood trap' in accessibility and health studies. Urban Stud. 52, 338–357.

Van Buuren, S., Groothuis-Oudshoorn, K., 2011. mice: Multivariate Imputation by Chained Equations in R. J Stat. Softw. 45, 1–67.
Van Hulst, A., Thomas, F., Barnett, T.A., Kestens, Y., Gauvin, L., Pannier, B., Chaix, B., 2012. A typology of neighborhoods and blood pressure in the RECORD Cohort Study. J Hypertens. 30, 1336–1346.
Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with S, Fourth Edition Springer, New York.
Villanueva, K., Knuiman, M., Nathan, A., Giles-Corti, B., Christian, H., Foster, S., Bull, F., 2014. The impact of neighborhood walkability on walking: does it differ across adult life stage and does neighborhood buffer size matter? Health Place 25, 43–46.
Wanner, M., Gotschi, T., Martin-Diener, E., Kahlmeier, S., Martin, B.W., 2012. Active transport, physical activity, and body weight in adults: a systematic review. Am. J Prev. Med. 42, 493–502.
Weinberg, C.R., 2007. Can DAGs clarify effect modification? Epidemiology 18, 569–572.
Wijndaele, K., Westgate, K., Stephens, S.K., Blair, S.N., Bull, F.C., Chastin, S.F., Dunstan, D.W., Ekelund, U., Esliger, D.W., Freedson, P.S., Granat, M.H., Matthews, C.E., Owen, N., Rowlands, A.V., Sherar, L.B., Tremblay, M.S., Troiano, R.P., Brage, S., Healy, G.N., 2015. Utilization and harmonization of adult accelerometry data: review and expert consensus. Med. Sci. Sport Exer. 47, 2129–2139.
World Health Organisation, 2015. Factsheet on physical activity.

3.4 Article 4: Simulating the impact of transport mode shifts on transport-related physical activity

Submitted September 2016

Simulating the impact of transport mode shifts on transport-related physical activity

Ruben Brondeel [1,2,3*], Yan Kestens [4], Basile Chaix [1,2]

Abstract

Background: Physical inactivity is widely recognized as one of the leading causes of mortality. Transport interventions have been implemented to increase physical activity, but intervention evaluations have been limited by the lack of data. This study develops a simulation approach to evaluate the potential impact of transport mode shifts on physical activity estimated by accelerometer data. **Methods:** Scenarios were designed and tested based on the Global Transport Survey (n = 21332) and the RECORD GPS Study (n = 229) from the French region of Paris. The scenarios included promoting walking, biking, or public transport and discouraging private motorized modes. Random forest models were used to predict the likelihood that each trip was made with an alternative mode and to evaluate the impact of the mode shifts on physical activity. **Results:** Promoting walking and discouraging private motorized modes were the most effective scenarios, with a gain of 6 minutes of moderate to vigorous physical activity (MVPA) per day for the most ambitious scenarios. Promoting biking or public transport was less effective (3 minutes of MVPA), due to a low prevalence of biking trips and reverse effects of public transport replacing walking trips. Inequalities by educational level in transport-related physical activity were relatively large, and were increased by the simulated transport mode shifts. **Conclusions:** Successful transport interventions may contribute to increase physical activity in adults. The simulations suggest that public transport should be explicitly promoted as an alternative for private motorized transport, to limit reverse effects.

Keywords

Active transport — Simulation study — Machine learning — Data integration — Health inequalities — RECORD Cohort Study — France

[1] *Inserm, UMR-S 1136, Pierre Louis Institute of Epidemiology and Public Health, Nemesis team, Paris, France*
[2] *Sorbonne Universités, UPMC Univ Paris 06, UMR-S 1136, Pierre Louis Institute of Epidemiology and Public Health, Nemesis team, Paris, France*
[3] *EHESP School of Public Health, Rennes, France*
[4] *Department of Social and Preventive Medicine, University of Montreal, Montreal, Quebec, Canada*
*__Corresponding author__: Ruben Brondeel, ruben.brondeel@iplesp.upmc.fr, +33 1 44 73 86 54, Address: Faculté de Médecine Saint-Antoine, 27 rue Chaligny, UMR-S 1136, 75012 Paris, France

1. Introduction

Overall physical activity levels are low worldwide, with an estimated 31% of adults physically inactive.[1] Physical inactivity is widely recognized as one of the leading causes of mortality and morbidity due to its impact on several noncommunicable diseases.[2,3] Therefore, the World Health Organization (WHO) and many governments have adopted health plans to promote regular physical activity.[2,3]

Transport activity can be an important source of regular, incidental physical activity,[4–6] making transport interventions promising for the promotion of physical activity.[7,8] From an individual perspective, the impact of interventions promoting organized recreational physical activity (e.g., sport) is likely to be higher. However, the impact of interventions promoting transport physical activity may be substantial in a Public health perspective due to the universal nature of the transport activity.[7,9]

Even if many studies have evaluated real-world transport interventions,[8,10–12] such studies remain difficult to conduct. First, transport interventions on a community scale are often costly, and the evaluation of such interventions (e.g., with before and after assessments) are challenging to design and are themselves costly when relying on assessment methodologies such as accelerometers.[13,14] Second, real-world interventions are often implemented over restricted territories or populations (e.g., one school or company), and their impact may be difficult to generalize to the larger population. Third, it is often impossible to determine what would have been the impact of the intervention if it had been implemented in a different way or with a different intensity. To address these concerns, as a complement of evaluations of real-world interventions, we developed an approach assessing the impact of hypothetical interventions through simulation.

The aim of the present study was to evaluate the impact of transport mode shifts on transport-related physical activity. We propose a simulation approach based on random forest prediction methods. The transportation modes are changed in a predefined percentage of trips. Then, the recalculated post-

intervention transport-related physical activity is compared to the pre-intervention one to evaluate the impact of the intervention at the population level. To reach precise estimates of intervention effects, we integrate detailed accelerometer data for accuracy with mobility survey data from a large sample for an improved generalizability.

People with a low educational level have lower physical activity levels.[15] Transport interventions, like other health-related interventions, have the potential to enlarge existing social inequalities.[11,16] Therefore, we also evaluated the impact of the simulated transport mode shifts on the magnitude of educational inequalities in transport-related physical activity.

2. Methods

2.1 Study Population

The Global Transport Survey ('Enquête Global Transport', EGT) is a household travel survey conducted every 10 years in Île-de-France, the French capital region. The main purpose of the survey is to inform local authorities and transport planners on mobility and transport use in Île-de-France. The last EGT survey, approved by the French Data Protection Authority, was conducted in 2010 by the Ile-de-France Transport Authority (STIF) and the Regional and Interdepartmental Direction for Equipment and Planning (DRIEA). During face-to-face interviews with members of randomly selected households, data were collected for all the trips made during the day before the interview. For this study, we selected participants between 35 and 83 years old, resulting in a dataset of 82084 trips made by 21332 people.

2.2 Measures and Definitions

There were no accelerometer data in the EGT sample. Therefore, the measure of transport-related moderate to vigorous physical activity (T-MVPA) was introduced in the dataset by the integration of the EGT and the RECORD GPS Study datasets. The RECORD GPS Study,[4,14,17,18] as a subsample of the RECORD Cohort Study,[19–24] collected mobility pattern and accelerometer data for 236 participants during 7 days, resulting in the observation of 7138 trips. All participants resided in Ile-de-France and were between 35 and 83 years old, comparable to the EGT study population. A full description of the study design can be found in Supplementary material S1. In the RECORD GPS dataset, a minute of MVPA was defined as a minute during which the 3-axis vector magnitude was higher than 2690,[25] based on the tri-axial GT3X+ accelerometer data. Accelerometers worn at the hip underestimate physical activity during biking trips. Therefore, we used an estimate of biking physical activity from the compendium of Ainsworth,[26] i.e., all minutes during biking trips were considered as minutes of T-MVPA.

The data integration consisted of predicting the T-MVPA in EGT based on the data of the RECORD GPS Study,[27] using a random forest prediction model.[28] For the data integration, 45 variables common to both datasets were used to predict the accelerometer based T-MVPA. These can be categorized as follows: trip characteristics (e.g., transportation mode, duration), personal characteristics (e.g., age and educational level), personal transport accessibility characteristics (e.g., possession of a motorized vehicle) and area transport accessibility characteristics for the residence and the departure and arrival location of each trip (e.g., distance to nearest transport station). Supplementary material S2 presents the full list of variables used for the data integration. The data integration process was previously described in detail.[27]

Three categories of educational level were considered: 'no diploma of secondary education', 'diploma of secondary education or lower tertiary education', and 'diploma of higher tertiary education'. The transportation mode variable consisted of four categories: 'walking', 'bicycle', 'private motorized', and 'public transport'. Trips with non-walking modes that also included 'walking' were categorized on the basis of the non-walking mode.

2.3 Statistical Analysis

Analyses were conducted in 2015–2016. Twelve scenarios of transportation mode shifts were considered, 3 for each of the 4 transportation modes: walking, biking, public transport, and private motorized transport. All scenarios were designed to promote more active transportation modes. So, in the private motorized scenarios, private motorized trips were changed into walking, biking, or public transport trips. For the other three modes, trips not performed by the respective mode were changed into this mode. For example, in the walking scenarios, non-walking trips were changed into walking trips.

The simulation process for all scenarios consisted of three consecutive steps. In a first step, the transportation mode for a predefined proportion of trips was changed into an alternative mode. The predefined proportion of trips was chosen in function of the prevalence of the mode under consideration. In the 'private motorized' scenarios, the percentages of private motorized trips changed into walking, biking, or public transport trips were of 10%, 20%, and 30%. In the walking scenarios, the number of non-walking trips changed into walking trips was of 10%, 30%, and 50% of the observed walking trips. The same percentages were applied in the public transport scenarios. In the biking scenarios, the percentages applied were 100%, 200%, and 300%.

Any trip of interest for the intervention could be selected to be changed, but the selection was weighted by the likelihood of a trip to be performed by the alternative mode. For the 'private motorized' scenarios, the alternative mode was the most likely alternative transportation mode for the respective trip. For the other scenarios (change to walking, biking, or public transport), trips were selected based on their likelihood to be performed by the target mode. Taking as an example the scenarios to promote walking, non-walking trips were selected for change based on the likelihood that these trips were performed by walking. In this example, the likelihood of performing these non-walking trips by walking was extracted

Table 1. Variables used to predict transportation mode (TM), duration of the trip (D), and transport-related MVPA (MVPA)

	TM	D	T-MVPA
Minutes of T-MVPA per trip [a]			X
Duration of the trip (in minutes) [b]		X	X
Speed based on duration and straight-line distance [b]			X
Transportation mode [b]	X	X	X
Trip characteristics			
Time of the day at departure [b]	X	X	X
Day of the week at departure [b]	X	X	X
Rush hour or not at departure: from 8am to 11am and from 4pm to 7pm [b]	X	X	X
Straight-line distance from departure point to arrival point [c]	X	X	X
Personal variables			
Age [c]	X	X	X
Gender [c]	X	X	X
Work situation (employed, unemployed, retired, other) [c]	X	X	X
Personal education level [c]	X	X	X
Household income [c]	X	X	X
Spatial access to public transport from the trip departure, trip arrival, and residence (3 separate sets of variables)			
Street network distance to nearest public transport station from residence [d]	X	X	X
Street network distance to nearest train station [d]	X	X	X
Street network distance to nearest metro station [d]	X	X	X
Street network distance to nearest tram station [d]	X	X	X
Street network distance to nearest bus station [d]	X	X	X
Other contextual characteristics at the trip departure, trip arrival, and residence (3 separate sets of variables)			
Educational level in the area [e]	X	X	X
Number of destinations in the area [e]	X	X	X
Number of intersections in the area [e]	X	X	X
Surface of parks in the area [e]	X	X	X
Population density in the area [e]	X	X	X
Place located in Paris, in the other counties adjacent to Paris, or in the other counties non-adjacent to Paris	X	X	X

[a] Accelerometry information in RECORD or predicted time in EGT; [b] Information from the mobility survey in RECORD and in EGT; [c] RECORD and EGT questionnaires; [d] Shortest street network distance determined with ArcGIS from the residence or from the departure/arrival of each trip geocoded at the address level in RECORD or at the center of a 100 m square in EGT; [e] The area around the residence or departure or arrival point of each trip was defined with ArcGIS as a 1 km buffer following the street network, and information was aggregated at the level of this buffer; MVPA: moderate to vigorous physical activity.

from a random forest model predicting the transportation mode. We had to rescale this average predicted probability of walking in these non-walking trips to the pre-specified level of change in the scenario of interest (e.g., to 10%, 30%, or 50%). This rescaling relied on a transformation of the probabilities to the logit scale and then back to the probability scale to avoid probabilities out of the [0; 1] range. These transformed probabilities enabled us to draw random samples of the trips selected for change, weighted by the likelihood for the trip to be performed by the alternative mode given the predictor variables.

In a second step, the duration of the trip was predicted for the trips to which a new transportation mode was attributed in step 1. The prediction was based on a random forest model for the duration of trips in the EGT data. In a final step, the MVPA was predicted for the trips with a changed transportation mode and duration with the same model than the one used for the data integration, i.e., a random forest model for MVPA based on the RECORD GPS data. The simulation of each scenario was repeated 100 times to avoid random sampling error in the results. Table 1 presents all the variables used in the three random forest models for the transportation mode, duration, and T-MVPA. The scripts for all the analyses with R (v3.3.0)[29] can be found in Supplementary material S3.

3. Results

Table 2 presents the predicted daily T-MVPA for the observed mobility patterns, the scenarios of the transport mode shifts, the impact of the shifts on T-MVPA, and the impact on the inequalities in T-MVPA by educational level. The mean T-MVPA predicted for the observed transportation modes was 19.0 min per day (CI = 18.9; 19.0). The mean T-MVPA was of 17.5 (CI = 17.5; 17.6), 18.6 (CI = 18.6; 18.7) and 21.1 (CI = 21.0; 21.1) minutes per day for people with a low, medium, and high educational level respectively.

Each simulated intervention increased the mean T-MVPA. The walking scenarios increased T-MVPA to 19.8 minutes per day for a 10% increase in the number of walking trips and to 25.0 minutes for a 50% increase. The biking scenarios increased T-MVPA to 20.0 and 22.2 minutes per day for increases of 100% and 300%. The public transport scenarios increased T-MVPA to 19.4 and 21.3 minutes per day for increases of 10% and 50%. The private motorized scenarios increased T-MVPA levels to 20.7 minutes and 25.1 minutes per day for decreases in the number of private motorized trips by respectively 10 and 30%.

Almost all scenarios increased the absolute inequalities in T-MVPA between educational groups. Only the absolute inequalities between the medium and high educational level

Table 2. Predicted T-MVPA for observed and simulated mobility patterns

	Changes by simulation			T-MVPA	T-MVPA by Education			Abs. Diff. in T-MVPA [a]			Rel. Diff. in T-MVPA [b]		
	% Cat	% Tot	N	Total	E1	E2	E3	E2-1	E3-1	E3-2	E2-1	E3-1	E3-2
Observed				19.0	17.5	18.6	21.1	1.1	3.5	2.4	1.06	1.20	1.13
Walking	10	3.4	2754	19.8	18.3	19.5	21.9	1.2	3.6	2.5	1.06	1.20	1.13
	30	10.1	8261	22.2	20.6	21.9	24.4	1.3	3.8	2.5	1.06	1.19	1.12
	50	16.8	13761	25.0	23.3	24.8	27.4	1.5	4.1	2.6	1.07	1.18	1.10
Biking	100	1.5	1205	20.0	18.4	19.7	22.4	1.3	4.0	2.7	1.07	1.22	1.14
	200	2.9	2420	21.1	19.2	20.8	23.8	1.6	4.6	3.0	1.08	1.24	1.14
	300	4.4	3634	22.2	20.1	21.9	25.1	1.8	5.0	3.2	1.09	1.25	1.15
Pub. Trans.	10	1.4	1174	19.4	17.9	19.1	21.5	1.2	3.6	2.4	1.07	1.20	1.12
	30	4.3	3527	20.4	18.8	20.1	22.4	1.3	3.6	2.3	1.07	1.19	1.12
	50	7.1	5863	21.3	19.8	21.1	23.4	1.4	3.6	2.3	1.07	1.18	1.11
Priv. Mot.	10	5.1	4159	20.7	19.1	20.4	22.9	1.3	3.8	2.5	1.07	1.20	1.12
	20	10.1	8309	22.8	21.1	22.6	25.1	1.5	4.0	2.5	1.07	1.19	1.11
	30	15.2	12478	25.1	23.3	25.0	27.3	1.7	4.0	2.3	1.07	1.17	1.09

Total number of persons = 21332; Total number of trips = 82084; % Cat = the change in transportation modes for the scenario is expressed as a percentage of the observed number of trips in the category that is eligible for change (for walking for example, it refers to 10%, 30%, or 50% of non-walking trips that can be changed to walking); % Tot = the change in transportation modes for the scenario is expressed as a percentage of the total number of trips in the sample; N = number of trips changed by the intervention; T-MVPA = transport-related moderate-to-vigorous physical activity expressed in minutes per day; E1 = no diploma of secondary education; E2 = Diploma of secondary education or two years or less of University education; E3 = Diploma of three years or more of University education; Pub. Trans. = public transport; Priv. Mot. = private motorized transport; a Difference between the two categories of educational level in T-MVPA; b Ratio of T-MVPA between 2 categories of educational level.

decreased for the public transport and private motorized scenarios. The ratios indicated that the relative differences between the educational levels were rather small, with overall minor decreases after the walking, public transport, and private motorized scenarios. The highest increases in absolute and relative differences between educational levels were noted for the biking scenarios.

Table 3 presents the T-MVPA by transportation mode, before and after the four simulated mode shifts. Most of the observed T-MVPA was related to walking (6.8 minutes, CI = 6.8; 6.8) and public transport (6.8 minutes, CI = 6.8; 6.8), followed by private motorized transport (4.2 minutes, CI = 4.2; 4.2) and biking (1.2 minutes, CI = 1.2; 1.2). For each simulation, Table 3 shows the intended positive effects and the unintended reverse effects, i.e., the loss of physical activity for the transportation modes that were not promoted.

4. Discussion

4.1 Main results

This study underlines the importance of transport for reaching daily physical activity levels. People between 35 and 83 years old residing in the Ile-de-France region had an average of 19 minutes of daily T-MVPA. The impact of the transport mode shifts was an increase of MVPA by 6 minutes per day for the most ambitious scenarios promoting walking or discouraging private motorized transport. The impact for the most ambitious scenarios promoting biking or public transport was about 3 min of MVPA per day. These findings confirm the importance of transport interventions in the promotion of physical activity.

The most efficient scenarios were promoting walking and discouraging private motorized modes (i.e., car or motorbike). A reason may be that, for a given distance, walking is the most active mode and private motorized the least active mode. The impact of biking scenarios was relatively modest, due to a very low frequency of biking trips in our population (1% of the total number of trips).

Transport interventions to promote physical activity traditionally focus on walking and biking.[12] Recently, public transport has been found to also contribute significantly to population physical activity levels.[4,12,27,30] The results in this study confirm this finding, with a mean of 6.8 minutes of daily MVPA related to public transport with the observed patterns of mobility, corresponding to 36% of the total daily T-MVPA. However, public transport interventions had a lower impact on total T-MVPA than expected by the observed increase in T-MVPA. This is partly due to reverse effects on T-MVPA when promoting public transport. Promoting public transport also induces changing trips previously performed by walking or biking to public transport, which lowers the physical activity during these trips. These results suggest that public transport should explicitly be promoted as an alternative for private motorized transport exclusively, to limit these reverse effects.

Clear absolute differences in T-MVPA by educational level were observed and such differences were larger after the transport mode shifts (although relative inequalities in T-MVPA did not increase, except in the biking scenarios). The differential impact of the mode shifts by educational level

Table 3. Impact of transportation mode shifts on T-MVPA per day: tabulations by transportation mode

	Observed	Simulated transportation mode shifts			
		Promoting Walking 30%	Promoting Biking 200%	Promoting PT 30%	Discouraging PM 20%
Total T-MVPA	19.0	+3.2	+2.1	+1.4	+3.8
T-MVPA in walking trips	6.8	+4.9	-0.3	-0.3	+1.9
T-MVPA in biking trips	1.2	-0.3	+2.8	-0.2	+0.1
T-MVPA in public transport trips	6.8	-0.7	-0.3	+2.2	+2.8
T-MVPA in private motorized trips	4.2	-0.7	-0.1	-0.3	-1.0

T-MVPA: transport-related moderate-to-vigorous physical activity expressed in min per day; PT: public transport; PM: private motorized transport

was completely due to the characteristics of the trips (e.g., length of the trip). In real-life interventions, uptake of, access to, and compliance with the intervention are likely to be different according to the educational group.[16] Some of these factors were accounted for in our model defining the probability of change (e.g., work situation, geographic location of the residence, spatial access to services and public transport). However, since it is unlikely we included all factors contributing to a weaker impact of an intervention among the low educated, we can expect that educational disparities after an intervention would likely be larger than predicted in this study, unless social disparities are considered in the intervention design. Too often, social disparities are neglected in the design and evaluation of active transport interventions.[31]

4.2 Strengths and limitations

Simulation studies cannot replace intervention studies. There are many unknown variables in a real-life setting that cannot be simulated such as the adaptation of participants to an intervention, the longer term effects, unintended changes in other variables that are important for the outcome, etc. However, we argue that well-designed simulations can be complementary to intervention studies. Simulations have the great advantage of being very cost-efficient, while allowing for the comparison of a multitude of intervention scenarios.

The results of the simulation will only make sense if it is well-designed. Due to a clear and strong causal link between the transportation mode and physical activity, this type of simulations can provide important and reliable information on the expected range of the impact. For example, the results shed light on the intended and unintended effects of transport interventions on T-MVPA.

The most important limitation of this simulation is that we had to assume that transport scenarios could only affect the transportation mode used in a trip, but that it could not influence the choice to make a trip or not or the destination of the trip itself. For example, a successful transport intervention could motivate people to choose a destination further away than the current destination if this further destination has more to offer.

The collection of accelerometer measures of transport-related physical activity is very expensive for large samples. Survey measures of physical activity are prone to memory biases and they only approximate physical activity by indirect indicators (e.g. 'minutes of walking during one week'). We therefore used data integration to add an accelerometer measure of transport-related physical activity to the large survey dataset.[16,27] Even though the prediction model had a high accuracy, real measured accelerometer data would definitely be preferable. The advantage of using this outcome over survey data is the detail it provides. For example, public transport trips are clearly not fully inactive periods of time, since the person has to walk to and from the public transport station. The predicted measure of MVPA used in this study can capture this type of physical activity, making it an interesting measure for large-scale studies for which real accelerometer data is unavailable.

5. Conclusion

This study shows that transport mode shifts can have a significant impact on daily MVPA, even at a population level. The simulated mode shifts with the highest impact were those promoting walking or discouraging car use, compared to the scenarios promoting biking or public transport. The lower impact of the latter two strategies may be attributable to the low prevalence of biking and to the reverse effects of promoting public transport (decreasing the number of car trips but also walking trips). Public transport should explicitly be promoted as an alternative for private motorized transport to limit these reverse effects. The simulations also showed that interventions may increase the absolute inequalities in transport-related physical activity by educational level, which should be anticipated during the design of interventions.

Acknowledgments

The RECORD GPS Study was supported by INPES (National Institute for Prevention and Health Education); the Ministry of Ecology (DGITM); CEREMA (Centre for the Study of and Expertise on Risks, the Environment, Mobility, and Urbanism); ARS (Health Regional Agency) of Ile-de-France; STIF (Ile-de-France Transport Authority); the Ile-de-France Regional Council; RATP (Paris Public Transport Operator); and DRIEA (Regional and Interdepartmental Direction for Equipment and Planning). The authors thank the following partners from the funding institutions: Pierre Arwidson, Nadine Asconchilo, Annette Gogneau, Colette Watellier, Yasmina Babaa, Laurent Jardinier, Tristan Guilloux, Mélani Alberto,

Christelle Paulo, Anne-Eole Meret-Conti, Cédric Aubouin, Benoît Kiéné, Hélène Pierre, Sophie Mazoué, John Séraphin, and Ivan Derré. Ruben Brondeel conceived and designed this particular study, drafted the manuscript and analyzed and interpreted the data. Yan Kestens critically revised the manuscript. Basile Chaix conceived and designed the study, supervised the acquisition of data and the present work, and critically revised the manuscript. No financial disclosures were reported by the authors of this paper.

References

[1] Hallal P, Andersen L, Bull F, Guthold R, Haskell W, Ekelund U, et al. Global physical activity levels: Surveillance progress, pitfalls, and prospects. Lancet. 2012;380(9838):247–257.

[2] WHO. Assessing national capacity for the preventing and control of noncommunicable diseases; 2012.

[3] Kohl H, Craig C, Lambert E, Inoue S, Alkandari J, Leetongin G, et al. The pandemic of physical inactivity: Global action for public health. Lancet. 2012;380(9838):294–305.

[4] Chaix B, Kestens Y, Duncan S, Merrien C, Thierry B, Pannier B, et al. Active transportation and public transportation use to achieve physical activity recommendations? A combined GPS, accelerometer, and mobility survey study. Int J Behav Nutr Phys Act. 2014;11(1):124.

[5] Besser L, Dannenberg A. Walking to Public Transit Steps to Help Meet Physical Activity Recommendations. Am J Prev Med. 2005;29(4):273–280.

[6] Sahlqvist S, Song Y, Ogilvie D. Is active travel associated with greater physical activity? The contribution of commuting and non-commuting active travel to total physical activity in adults. Prev Med. 2012;55(3):206–11.

[7] Reynolds R, McKenzie S, Allender S, Brown K, Foulkes C. Systematic review of incidental physical activity community interventions. Prev Med. 2014;67:46–64.

[8] Scheepers C, Wendel-Vos G, den Broeder J, van Kempen E, van Wesemael P, Schuit A. Shifting from car to active transport: A systematic review of the effectiveness of interventions. Transp Res Part A. 2014;70:264–280.

[9] Brockman R, Fox K. Physical activity by stealth? The potential health benefits of a workplace transport plan. Public Health. 2011;125(4):210–6.

[10] Arnott B, Rehackova L, Errington L, Sniehotta FF, Roberts J, Araujo-Soares V. Efficacy of behavioural interventions for transport behaviour change: systematic review, meta-analysis and intervention coding. Int J Behav Nutr Phys Act. 2014;11:133.

[11] Ogilvie D, Egan M, Hamilton V, Petticrew M. Promoting walking and cycling as an alternative to using cars : systematic review. BMJ. 2004;.

[12] Petrunoff N, Rissel C, Wen LM. The effect of active travel interventions conducted in work settings on driving to work: A systematic review. J Transp Health. 2016;3(1):61–76.

[13] Bohte W, Maat K. Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: a large-scale application in the Netherlands. Transp Res Part C. 2009;17(3):285–297.

[14] Brondeel R, Pannier B, Chaix B. Using GPS, GIS, and Accelerometer Data to Predict Transportation Modes. Med Sci Sport Exer. 2015;47(12):2669–2675.

[15] Beenackers M, Kamphuis C, Giskes K, Brug J, Kunst A, Burdorf A, et al. Socioeconomic inequalities in occupational , leisure-time , and transport related physical activity among European adults : A systematic review. Int J Behav Nutr Phy Act. 2012;9(116).

[16] Lorenc T, Petticrew M, Welch V, Tugwell P. What types of interventions generate inequalities? Evidence from systematic reviews. J Epidemiol Commun H. 2013;67:190–193.

[17] Chaix B, Meline J, Duncan S, Merrien C, Karusisi N, Perchoux C, et al. GPS tracking in neighborhood and health studies: a step forward for environmental exposure assessment, a step backward for causal inference? Health Place. 2013;21:46–51.

[18] Chaix B, Kestens Y, Duncan DT, Brondeel R, Méline J, El Aarbaoui T, et al. A GPS-based methodology to analyze environment–health associations at the trip level: case-crossover analyses of built environment effects on walking. Am J Epidemiol. (In press);.

[19] Chaix B, Kestens Y, Bean K, Leal C, Karusisi N, Meghiref K, et al. Cohort profile: Residential and nonresidential environments, individual activity spaces and cardiovascular risk factors and diseases-The RECORD cohort study. Int J Epidemiol. 2012;41(5):1283–1292.

[20] Havard S, Reich BJ, Bean K, Chaix B. Social inequalities in residential exposure to road traffic noise: an environmental justice analysis based on the RECORD Cohort Study. Occup Environ Med. 2011;68(5):366–74.

[21] Chaix B, Bean K, Daniel M, Zenk SN, Kestens Y, Charreire H, et al. Associations of Supermarket Characteristics with Weight Status and Body Fat: A Multilevel Analysis of Individuals within Supermarkets (RECORD Study). PLoS ONE. 2012;7(4):e32908.

[22] Leal C, Bean K, Thomas F, Chaix B. Multicollinearity in Associations Between Multiple Environmental Features and Body Weight and Abdominal Fat: Using Matching Techniques to Assess Whether the Associations are Separable. Am J Epidemiol. 2012;175(11):1152–1162.

[23] Chaix B, Jouven X, Thomas F, Leal C, Billaudeau N, Bean K, et al. Why socially deprived populations have a faster resting heart rate: Impact of behaviour, life course anthropometry, and biology - the RECORD Cohort Study. Soc Sci Med. 2011;73(10):1543–1550.

[24] Brondeel R, Weill A, Thomas F, Chaix B. Use of healthcare services in the residence and workplace neighbourhood: the effect of spatial accessibility to healthcare services. Health Place. 2014;30:127–133.

[25] Sasaki JE, John D, Freedson PS. Validation and comparison of ActiGraph activity monitors. J Sci Med Sport. 2011;14(5):411–6.

[26] Ainsworth B, Haskell W, Herrmann S, Meckes N, Bassett Jr D, Tudor-Locke C, et al. 2011 Compendium of Physical Activities: a second update of codes and MET values. Med Sci Sport Exer. 2011;43(8):1575–1581.

[27] Brondeel R, Pannier B, Chaix B. Associations of socioeconomic status with transport-related physical activity: combining a household travel survey and accelerometer data using random forests. J Transp Health. In press;.

[28] Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.

[29] Team RC. R: A language and environment for statistical computing. R Foundation for Statistical Computing A Vienna, editor; 2014. Available from: `http://www.R-project.org/`.

[30] Rissel C, Curac N, Greenaway M, Bauman A. Physical Activity Associated with Public Transport Use—A Review and Modelling of Potential Benefits. Int J Env Res Public Heal. 2012;9(12):2454–2478.

[31] Ogilvie D, Foster C, Rothnie H, Cavill N, Hamilton V, Fitzsimons C, et al. Interventions to promote walking: Systematic review. BMJ. 2007;334(7605):1204–1207.

4. Discussion

In this discussion, we first summarize the main results reported in the articles and compare them to findings in previous research. The results are organised into two sections: methodological findings and developments, and empirical findings. After the summary of results, we discuss the strengths and limitations of this PhD work. Then, we discuss the contributions of this work to the literature in light of the broader research context and indicating the most relevant topics for future research. Finally, we present our conclusions based on this PhD work.

4.1 Summary of results

4.1.1 Methodological findings and developments

Transport-related moderate-to-vigorous physical activity (T-MVPA) accumulated per day was the principal indicator used to measure transport-related physical activity throughout this PhD work. Each of the four articles has contributed to the measurement of MVPA or T-MVPA measurements.

The first research objective in this work was to *investigate the bias introduced in accelerometer-based measurements of physical activity by using short epoch lengths.* The corresponding article (Article 3.1) evaluated the impact of the epoch length on

the estimation of accumulated MVPA per day when using the Sasaki tri-axial cut point.[22] The MVPA estimation using 60-s epochs was the reference measure, as the Sasaki cut point was originally calibrated based on 60-s epoch data. We found that shorter epochs resulted in considerably larger estimates of MVPA levels compared to 60-s epochs. For example, the estimated MVPA was 96 minutes per day using 1-s epochs; whereas using 60-s epochs resulted in an estimated 49 minutes per day.

This was the first study to investigate the impact of epoch length on a tri-axial measure of MVPA in adults. Previous studies based on uni-axial data from adults[18,19] also reported larger estimates of MVPA when using shorter epoch lengths. When we compared the results for the Sasaki tri-axial cut point to the Freedson uni-axial cut point in our own data, the tri-axial indicator of MVPA was more susceptible to the epoch length than the uni-axial indicator. A previous study used a tri-axial cut point (validated with 15-s epochs) for accelerometer data from children,[17] and also found higher estimates when using shorter epoch lengths. A larger impact of the epoch length for tri-axial cut points compared to uni-axial cut points is an important finding for future research, as it is likely that more studies will be based on tri-axial data in the future.

The second objective of this study was to *develop a prediction method for the transport mode that enables more cost- and time-efficient data collection for transport-related physical activity.* A precise observation of the mobility patterns, i.e. the departure and arrival times and locations and the transport modes, is necessary to measure T-MVPA. Collecting such mobility data is very work intensive, it is therefore difficult or impossible to do this collection for large study samples. This study relied on previous work from Thierry et al.[32] for the detection of departure and arrival times and locations. In the corresponding article (Article 3.2), we developed a methodology to automatize the detection of transport modes. An algorithm was based on a random forest prediction model, which correctly predicted 90% of the transportation modes in

the RECORD GPS dataset. This result is comparable to or better than results from previous studies.[23–27]

The third objective of this study was to *develop a data integration method combining an accelerometer dataset with a survey dataset in order to provide accelerometer-based estimates of physical activity for a large sample.* Instead of optimising the data collection process for large samples (such as the automatic detection of transportation modes used for the previous objective), estimating T-MVPA for a large sample was made possible by applying data integration. This approach relied on two datasets. The first was the RECORD GPS dataset, a relatively small dataset with very detailed information. The second dataset was the EGT dataset, a large dataset from a sample representative of the background population. This dataset included accurate mobility data, but no accelerometer data. We built a random forest prediction model for T-MVPA at the trip level based on the RECORD GPS dataset, which explained 67% of the variance in this variable. The model was then applied to the EGT dataset at the trip level, after which the accumulated T-MVPA per day was calculated.

The last methodological objective in this PhD work was to *develop a simulation technique that enables estimation of the changes in transport-related physical activity due to transport mode shifts.* Based on the T-MVPA measures obtained through the data integration, a simulation approach enabled the estimation of changes in T-MVPA associated with shifts in the transport modes observed. For each of the four transportation modes, three scenarios of active transport promotion were constructed. Random forest prediction models were used to select the trips for which the transport mode had to be changed, then to predict the new duration of the trip, and finally to predict the T-MVPA per trip. The accumulated T-MVPA per day was calculated in each scenario and compared to the original estimate.

4.1.2 Empirical findings

The fifth objective of this thesis was to *analyse the population distribution of transport-related physical activity for adults living in Ile-de-France, with a specific focus on social inequalities.* Based on integration of the RECORD and EGT datasets, we were able to analyse the T-MVPA for a population between 35 and 83 years old residing in Ile-de-France. The average daily T-MVPA was 18.9 minutes (95% confidence interval: 18.6; 19.2 minutes).

To our knowledge, this PhD work is the first to use an accelerometer-based MVPA indicator for transport-related physical activity. Previous research comparing survey-based and accelerometer-based measures of total MVPA found that the two measures had low levels of correlation, and the estimated average minutes of MVPA substantially different than the survey estimate.[66,67] Therefore, we did not compare the above findings to previous research.

A way to interpret the results of physical activity studies is to make a comparison with the WHO recommendations (150 minutes of MVPA per week, or 30 minutes per day for most of the days during the week). A population average of 19 minutes of T-MVPA indicates a considerable contribution to total physical activity of transport. However, the WHO recommendations are themselves based on survey studies and are still waiting for an update based on accelerometer based studies.[10] It is therefore not clear to what extent accelerometer-based MVPA measures can be compared to the WHO guidelines.

Social inequalities in transport-related physical activity were investigated as the associations between household income or education level, and T-MVPA per day. People with the highest educational level (diploma of three or more years at university level) spent more time doing T-MVPA per day than those with a medium educational level (diploma of secondary education or maximum 2 years at university level); who, in

turn, spent more time in T-MVPA than the people with the lowest educational level (no secondary diploma). In contrast to the positive association of educational level, household income had a negative association with time spent doing T-MVPA. People with a higher household income did less T-MVPA per day.

In previous studies, a higher personal level of education has been associated with more minutes of walking for transport,[28] more trips with active transport modes,[28,29] and more cycling trips.[30] In contrast with the finding that higher levels of education are positively associated with active transport, higher income has been associated with fewer minutes of walking and less frequent trips with active modes.[28]

The sixth and final objective of this study was to *use simulations to analyse the impact of transport mode shifts on the population average of transport-related physical activity and the social inequalities in transport-related physical activity.* The simulations focused in each scenario on one of the four transport modes. Before the shifts of transport mode (i.e. the observed data), walking accounted for an average 6.8 minutes of T-MVPA per day, biking for 1.2 minutes, public transport for 6.8 minutes, and private motorised transport for 4.2 minutes per day.

Promoting walking and discouraging private motorised transport seemed to have the highest impact on physical activity, increasing the average minutes spent in T-MVPA per day by 6 minutes per day for the most ambitious plans. Biking scenarios had a relatively low impact because even the very ambitious plans, such as increasing the number of biking trips by 300%, resulted in relatively small changes in absolute numbers. However, even though there was a low population-level impact from biking scenarios, the individual impact of one or more extra biking trips per day at the individual level was large; especially if the trips were previously performed by motorised transport.

Although public transport was an important source of T-MVPA per day, the impact

of the public transport scenarios was relatively low. This was largely due to reverse effects. When promoting public transport, the relative contribution of the public transport trips to T-MVPA increased; but simultaneously, the contributions of the other three categories, which involved active transport, decreased because fewer trips were undertaken using these modes. These reverse effects were present in all scenarios, but especially large for public transport scenarios. The trips most likely to be changed to public transport trips included private motorised trips, but also included walking and biking trips.

The simulated transport mode shifts further enlarged inequalities in T-MVPA by educational level. The simulations did not take into account certain important variables that lead to intervention-generated inequalities, such as access to the intervention, uptake of the intervention, and compliance with the intervention.[120] Therefore, the induced inequalities in T-MVPA due to real interventions are likely to be even larger than those predicted from this study.

4.2 Strengths and limitations

A key strength of this PhD work was the use of two datasets: the RECORD GPS dataset and the EGT dataset. Both datasets were collected from the same background population and complemented each other: the RECORD dataset included very detailed accelerometer and GPS measurements, while the EGT dataset was a good representation of the background population. The datasets included a large number of common variables, so they could be integrated at the trip-level.

However, the two datasets were not collected within the same study, which lead to the following limitations. First, the information obtained from the questionnaires might have been different in the two sources due to slightly different questions used, for example categories of education level had to be recoded; and due to different protocols,

for example surveys were conducted at the health centres in the RECORD Study and collected at home in the EGT study. The main difference between the two datasets was the data collection methods used in the mobility survey. The RECORD GPS Study used a prompted recall mobility survey enhanced with GPS data and the Mobility Web Mapping application; whereas the EGT study used a regular mobility survey. With no prompted recall possible, the EGT study collected only information related to the day before the survey was administered, to prevent memory bias. It is likely that the two protocols led to slightly different results. Second, there was no estimation of the precision of the data integration prediction models. The R^2 for the prediction model was high, indicating a good level of accuracy of the predictions from the RECORD GPS observations. The predictions from the EGT sample were likely less precise; but how much less precise could not be tested.

The analyses of T-MVPA presented in the article on data integration and social inequalities (Article 3.3) indicated social inequalities in T-MVPA by education level and income. However, the variables available in this study allowed little analyses of the mechanisms through which differences in educational level and income lead to social inequalities in T-MVPA. There was limited information on the personal motivations to make a trip, to choose a particular transportation mode, or to choose the destination. In previous survey-based research, these choices were influenced by factors related to social inequalities; for example, the built environment,[121,122] the perceived built environment,[123,124] and affective factors related to car ownership.[125]

A limit to the RECORD GPS dataset was the lack of information on the trip stages, i.e. unimodal segments of the trip; which was available in the EGT dataset. Particularly, the information on walking trip stages during non-walking trips could have improved the data integration. This information can, however, be partly introduced in the data integration model by other variables such as the distance to the closest transport station. Future research could determine whether detailed information on trip stages

has enough influence on the estimated physical activity levels to make it important to be collected in a mobility survey, given the extra burden that this collection would place on the participants and the researchers.

Using information at the trip stage level would also enable better use of the multimodal trips in the analysis. In the prediction models based on the RECORD GPS dataset, the trips performed by 2 or more non-walking modes were excluded because the different categories, for example trips by both car and bike, were too rare. In the analysis models based on the EGT dataset, however, it was possible to include these multimodal trips. Since the accumulation of T-MVPA was the main indicator, missing even a single trip could substantially alter the outcome. Therefore, multimodal trips were used for the EGT-based models, and the main transport mode for these trips was defined by the longest trip stage. This information was not available for the RECORD GPS dataset.

The study is based on data from people residing in the French capital region, Ile-de-France. Paris is a very centralised city, with a high concentration of jobs in the city centre.[126] It is also a relatively old city with no car-oriented development.[126] The results might therefore not be generalizable to regions with different population densities, transport infrastructures, and socio-geographical inequalities.

4.3 Contributions to the literature

This work is a contribution to the literature by evaluating accelerometer based measures of physical activity and discussing the associated data processing; by developing new methodologies to incorporate these measures in large-scale transport studies; by developing a simulation technique for transport mode shifts; and by reporting empirical findings related to the level of transport-related physical activity undertaken by 35 to 83 year old people living in Ile-de-France.

Accelerometers have been available for health research for more than two decades,[43] making more 'objective' measures of physical activity possible. However, accelerometers measure acceleration and not the activity type or the activity intensity. Several decisions in the data processing can have a substantial impact on the measurement of physical activity.

An often overlooked aspect of data processing is the epoch length. Some studies had already shown the importance of epoch lengths in the calculation of physical activity indicators based on uni-axial data;[19] nevertheless, many different lengths have been used in previous research.[81] The corresponding article (Article 3.1) contributed to the literature by pointing out that the impact of epoch lengths is too important to be overlooked, and that this impact is even greater for the new indicators based on tri-axial data. This is an important finding given that tri-axial devices are becoming the norm.

These findings indicate that a lot of work remains to be done in the development and standardisation of accelerometer indicators. Count based indicators are very prominent in the literature. Large-scale accelerometer studies linking physical activity to health outcomes are becoming feasible; and will become very important to inform future updates to the WHO guidelines. It is therefore important that researchers start to use comparable indicators of physical activity, including a standardised epoch length.

With counts readily provided by software, count based indicators of activity intensity are often used because they are simple to calculate; surprisingly simple compared to the complexity of the technology used for data collection.[127] The indicators are easy to use, however, they are arguably not very exact. The cut points are not adapted to the fact that people have very different levels of physical fitness. Separate cut points have therefore been developed for children and older adults; but even within these age groups, the intensity level can be very variable. Recent studies have therefore explored the possibility of using raw accelerometer data and machine learning techniques.[128]

Until these methods have been fully developed, it is important to standardise the existing methods so that studies are comparable.

A second contribution to the literature, in the article on the prediction of transportation modes (Article 3.2), was the development of a model to predict transportation modes, which can be used for data collections in transport and health studies. This article added theoretical insights related to automatic transport mode detection, such as the value of trip-level prediction models. Compared to time-unit prediction models, for example a prediction for every 5 seconds, trip-level prediction models, which use only one prediction for the whole trip, seem less informative and less detailed. However, our results showed the importance of trip-level variables, for example the maximum speed reached during the trip, for the accuracy of the predictions. The trip-level model does not replace time-unit models, but seems to complement this approach. More research is needed to investigate how to combine the two approaches within one prediction model. A first step in this direction is taken by Kohla et al.[27], but improvements on prediction accuracy can be expected by new developments in machine learning techniques.

In the article on data integration (Article 3.3), we further explored data integration methodologies, this time applying them to the T-MVPA indicator. The data integration methodology improves regular prediction methods by using datasets originating from the same population. Even though further research is needed to quantify the added value of using datasets from the same population, it can be expected that the prediction accuracy is considerably higher in this case. The technique was applied to a physical activity indicator (T-MVPA) for this study, but could be used in other contexts where good prediction models are possible.

A final contribution to the methodology in this field was the use of simulations to investigate the impact of transport mode shifts on physical activity. By testing a range of scenarios of shifts in transportation modes, it was possible to investigate the changes in T-MVPA. This study has provided tools to better understand the impact

of transport modes on physical activity and to anticipate approximately the results of future transport interventions on health outcomes. These tools include random forest machine learning, enabling a large degree of flexibility in the models. Some aspects of the simulations need further clarification and more complexity in the scenarios might result in better approximations of real transport interventions. In particular, the simulations of transport mode shifts did not take into account that real interventions would not only impact the transport mode, but also the destinations of the trips and the number of trips per person.

The empirical results of this study included the estimation of population T-MVPA levels and the investigation of social inequalities in T-MVPA. This study was the first to investigate accelerometer based T-MVPA on a large sample (n = 21332). For a population between 35 and 83 years old, residing in Ile-de-France, we found a mean of 19 minutes of T-MVPA per day. The results underlined the importance of transport in the accumulation of physical activity for this population of adults and older adults. As discussed before, most previous studies - and consequently the WHO recommendations - are based on survey data. There is therefore no real reference to compare these findings to. In the RECORD dataset, T-MVPA accounted for 47% of total daily MVPA. The RECORD dataset is not representative of the background population, however, this percentage indicates that at least a considerable part of the total MVPA is due to transport. Future studies that combine accelerometer data with precise mobility data - such as the MobiliSense project - are needed to more precisely investigate the relative contribution of transport to total physical activity.

Social inequalities were found in T-MVPA for both education level and the household income. Taking other variables into account, people with a higher educational level did more T-MVPA, whereas people with a higher income did less T-MVPA. These results underline the importance of using separate indicators when studying social inequalities in the field of physical activity. Using a composite socio-economic indicator would hide

the inverse relationships seen in relation to education and income. The findings in this study related to social inequalities are important, but are only a first step. More research is needed to understand the mechanisms through which factors such as income and education lead to social inequalities in (transport-related) physical activity.

4.4 Conclusion

This work was based on two very complementary datasets. The RECORD GPS study dataset is, to our knowledge, the largest accelerometer dataset that can accurately identify transport-related physical activity. Even though large in its type, the RECORD dataset is still small in terms of the number of individual participants, and it lacks generalisability to the background population, i.e. 35-83 year olds in Ile-de-France, the French capital region. Therefore, this dataset was integrated with the EGT dataset, a household transport survey which included 21332 participants and was more representative of the background population.

We wrote four articles based on these datasets, which are the basis of this PhD work. This PhD work has contributed to the critical evaluation of the most commonly used physical activity indicators: moderate-to-vigorous physical activity (MVPA), sedentary behaviour (SB), and light physical activity (LPA). The finding that epoch length has a considerable impact on these activity intensity indicators adds to a series of studies which have evaluated the data processing decisions taken during collection and processing of accelerometer data. On the one hand, we reiterate previous calls for standardising the existing indicators of physical activity;[31] on the other hand, these findings support the development of new indicators that might avoid some, if not most, of the decisions by using raw data and machine learning algorithms.

A second article advanced methodologies for automatic identification of mobility patterns, i.e. how, when, and where people use transport in their daily life. Relying on a

previously developed algorithm to detect the departure and arrival location and time of trips,[32] this study developed a method to automatically detect the transportation mode. There remains a substantial amount of work for future research to utilise the rapidly growing amount of data sources available in this field. It will also be of interest to see what private companies such as Google, and non-profit open source initiatives will be able to establish in the near future.

In a third and fourth article, we applied a data integration approach to use accelerometer based physical activity indicators in a large sample representative of the background population. This approach allowed us to move away from the black and white division between active trips such as walking and cycling trips, and non-active trips such as car, motorbike, and public transport trips. To understand transport-related physical activity, it is necessary to account for the active episodes which occur during a non-active trip, such as walking to a bus stop; and the non-active or less active episodes during active trips, such as slow pace walking and waiting at a red light. The importance of these episodes to the total amount of transport-related walking is likely highly dependent on the context. People living in rural areas or cities with a less developed public transport system are likely to have fewer active spells during inactive trips due to their reliance on cars. However, our findings indicate that these issues cannot be ignored in future research.

In the final article, we used a simulation approach to investigate the impact of mode shifts on transport-related physical activity. This approach used an innovative machine learning based methodology to investigate the approximate impact of interventions. This simulation methodology still requires a degree of testing and fine-tuning to improve the accuracy of the inferences made. However, it has already proved to be a very powerful and flexible method to model policy scenarios.

The empirical results in this work add important information to the growing number of studies that emphasise the importance of transport, including public transport to

physical activity. This study examined the social inequalities in transport-related physical activity, and underlined the importance of both household income and educational level as complementary measures of socio-economic status in this field. More research on the mechanisms by which income and educational inequalities lead to inequalities in transport-related physical activity is needed to better inform future transport and health interventions.

Bibliography

[1] Sofi F, Capalbo A, Cesari F, Abbate R, Gensini GF. Physical activity during leisure time and primary prevention of coronary heart disease: an updated meta-analysis of cohort studies. Eur J Cardiov Prev R. 2008;15(3):247–257.

[2] Reiner M, Niermann C, Jekauc D, Woll A. Long-term health benefits of physical activity–a systematic review of longitudinal studies. BMC Public Health. 2013;13(1):813.

[3] De Bourdeaudhuij I, Verloigne M, Maes L, Van Lippevelde W, Chinapaw MJM, Te Velde SJ, et al. Associations of physical activity and sedentary time with weight and weight status among 10-to 12-year-old boys and girls in Europe: A cluster analysis within the ENERGY project. Pediatr Obes. 2013;8(5):367–375.

[4] Sallis J, Bull F, Guthold R, Heath G, Inoue S, Kelly P, et al. Progress in physical activity over the Olympic quadrennium. Lancet. 2016;6736(16):1–12.

[5] Teychenne M, Ball K, Salmon J. Physical activity and likelihood of depression in adults: A review. Prev Med. 2008;46(5):397–411.

[6] Behrens G, Leitzmann M. The association between physical activity and renal cancer: systematic review and meta-analysis. Br J Cancer. 2013;108(4):798–811.

[7] Schmid D, Leitzmann MF. Association between physical activity and mortality among breast cancer and colorectal cancer survivors: a systematic review and meta-analysis. Ann Oncol. 2014;23(Suppl 7):1–19.

[8] Thune I, Furberg A. Physical activity and cancer risk: dose-response and cancer, all sites and site-specific. Med Sci Sport Exer. 2001;33(6 Suppl):S530–S550.

[9] Wu Y, Zhang D, Kang S. Physical activity and risk of breast cancer: A meta-analysis of prospective studies. Breast Cancer Res Tr. 2013;137(3):869–882.

[10] WHO. Assessing national capacity for the preventing and control of noncommunicable diseases; 2012.

[11] Ogilvie D, Foster C, Rothnie H, Cavill N, Hamilton V, Fitzsimons C, et al. Interventions to promote walking: Systematic review. BMJ. 2007;334(7605):1204–1207.

[12] Pucher J, Dill J, Handy S. Infrastructure, programs, and policies to increase bicycling: an international review. Prev Med. 2010;50(S1):S106–25.

[13] Sahlqvist S, Song Y, Ogilvie D. Is active travel associated with greater physical activity? The contribution of commuting and non-commuting active travel to total physical activity in adults. Prev Med. 2012;55(3):206–11.

[14] Morgan PJ, Young MD, Smith JJ, Lubans DR. Targeted Health Behavior Interventions Promoting Physical Activity: A Conceptual Model. Exerc Sport Sci Rev. 2016;44(2):71–80.

[15] Bird E, Baker G, Mutrie N, Ogilvie D, Sahlqvist S, Powell J. Behavior Change Techniques Used to Promote Walking and Cycling : A Systematic Review. Heal psych. 2013;32(8):829–838.

[16] Reis RS, Salvo D, Ogilvie D, Lambert E, Goenka S, Brownson R. Series Physical Activity 2016: Progress and Challenges Scaling up physical activity interventions worldwide: stepping up to larger and smarter approaches to get people moving for the Lancet Physical Activity Series 2 Executive Committee*. Lancet. 2016;6736(16):1–12.

[17] Banda J, Haydel K, Davila T, Desai M, Bryson S, Haskell W, et al. Effects of Varying Epoch Lengths, Wear Time Algorithms, and Activity Cut-Points on Estimates of Child Sedentary Behavior and Physical Activity from Accelerometer Data. PLoS One. 2016;11(3):e0150534.

[18] Gabriel KP, McClain J, Schmid K, Storti K, High R, Underwood D, et al. Issues in accelerometer methodology: the role of epoch length on estimates of physical activity and relationships with health outcomes in overweight, post-menopausal women. Int J Behav Nutr Phy Act. 2010;7:53.

[19] Orme M, Wijndaele K, Sharp SJ, Westgate K, Ekelund U, Brage S. Combined influence of epoch length, cut-point and bout duration on accelerometry-derived physical activity. Int J Behav Nutr Phy Act. 2014;11(1):34.

[20] Bohte W, Maat K. Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: a large-scale application in the Netherlands. Transp Res Part C. 2009;17(3):285–297.

[21] Chaix B, Kestens Y, Duncan S, Merrien C, Thierry B, Pannier B, et al. Active transportation and public transportation use to achieve physical activity recommendations? A combined GPS, accelerometer, and mobility survey study. Int J Behav Nutr Phys Act. 2014;11(1):124.

[22] Sasaki JE, John D, Freedson PS. Validation and comparison of ActiGraph activity monitors. J Sci Med Sport. 2011;14(5):411–6.

[23] Ellis K, Godbole S, Marshall S, Lanckriet G, Staudenmayer J, Kerr J. Identifying active travel behaviors in challenging environments using GPS, accelerometers, and machine learning algorithms. Front Public Health. 2014;.

[24] Feng T, Timmermans HJP. Transportation mode recognition using GPS and accelerometer data. Transp Res Part C. 2013;37:118–130.

[25] Gong H, Chen C, Bialostozky E, Lawson CT. A GPS/GIS method for travel mode detection in New York City. Comput Environ Urban. 2012;36(2):131–139.

[26] Chen C, Gong H, Lawson C, Bialostozky E. Evaluating the feasibility of a passive travel survey collection in a complex urban environment: Lessons learned from the New York City case study. Transp Res Part A. 2010;44(10):830–840.

[27] Kohla B, Meschik M, Gerike R, Sammer G, Hössinger R, Unbehaun W. A New Algorithm for Mode Detection in Travel Surveys: Mobile Technologies for Activity- Travel Data Collection and Analysis. In: Rasouli S, Timmermans HJP, editors. Mobile Technologies for Activity-Travel Data Collection and Analysis. Hershey: IGI Global; 2014. p. 134–151.

[28] Cerin E, Leslie E, Owen N. Explaining socio-economic status differences in walking for transport: an ecological analysis of individual, social and environmental factors. Soc Sci Med. 2009;68(6):1013–20.

[29] Scheepers E, Wendel-Vos W, van Kempen E, Panis L, Maas J, Stipdonk H, et al. Personal and environmental characteristics associated with choice of active transport modes versus car use for different trip purposes of trips up to 7.5 kilometers in The Netherlands. PLoS One. 2013;8(9):e73105.

[30] Carse A, Goodman A, Mackett R, Panter J, Ogilvie D. The factors influencing car use in a cycle-friendly city: the case of Cambridge. J Transp Geogr. 2013;28:67–74.

[31] Wijndaele K, Westgate K, Stephens S, Blair S, Bull F, Chastin S, et al. Utilization and Harmonization of Adult Accelerometry Data: Review and Expert Consensus. Med Sci Sport Exer. 2015;47(10):2129–2139.

[32] Thierry B, Chaix B, Kestens Y. Detecting activity locations from raw GPS data: a novel kernel-based algorithm. Int J Heal Geogr. 2013;12(14).

[33] Schmid D, Behrens G, Keimling M, Jochem C, Ricci C, Leitzmann M. A systematic review and meta-analysis of physical activity and endometrial cancer risk. Eur J Epidemiol. 2015;30(5):397–412.

[34] Samitz G, Egger M, Zwahlen M. Domains of physical activity and all-cause mortality: Systematic review and dose-response meta-analysis of cohort studies. Int J Epidemiol. 2011;40(5):1382–1400.

[35] Kohl H, Craig C, Lambert E, Inoue S, Alkandari J, Leetongin G, et al. The pandemic of physical inactivity: Global action for public health. Lancet. 2012;380(9838):294–305.

[36] Hallal P, Andersen L, Bull F, Guthold R, Haskell W, Ekelund U, et al. Global physical activity levels: Surveillance progress, pitfalls, and prospects. Lancet. 2012;380(9838):247–257.

[37] Kalman M, Inchley J, Sigmundova D, Iannotti RJ, Tynjälä Ja, Hamrik Z, et al. Secular trends in moderate-to-vigorous physical activity in 32 countries from 2002 to 2010: a cross-national perspective. Eur J Public Heal. 2015;25 (Suppl:37–40.

[38] WHO. Physical activity and health in Europe: evidence for action; 2006.

[39] Söjström M, Oja P, Hagströmer M, Smith B, Bauman A. Health-enhancing physical activity across European Union countries: The Eurobarometer study. J Public Heal. 2006;14(5):291–300.

[40] Knuth AG, Hallal PC. Temporal trends in physical activity: a systematic review. J Phys Act Heal. 2009;6(5):548–559.

[41] Ekelund U, Tomkinson G, Armstrong N. What proportion of youth are physically active? Measurement issues, levels and recent time trends. Br J Sport Med. 2011;45(11):859–865.

[42] Steene-Johannessen J, Anderssen S, van der Ploeg H, Hendriksen I, Donnelly A, Brage S, et al. Are Self-report Measures Able to Define Individuals as Physically Active or Inactive? Med Sci Sport Exer. 2016;48(2):235–244.

[43] Bassett D, Troiano R, McClain J, Wolff D. Accelerometer-based Physical Activity: Total Volume per Day and Standardized Measures. Med Sci Sport Exer. 2015;47(4):833–838.

[44] Beenackers M, Kamphuis C, Giskes K, Brug J, Kunst A, Burdorf A, et al. Socioeconomic inequalities in occupational , leisure-time , and transport related physical activity among European adults : A systematic review. Int J Behav Nutr Phy Act. 2012;9(116).

[45] Hinrichs T, Von Bonsdorff MB, Törmäkangas T, Von Bonsdorff ME, Kulmala J, Seitsamo J, et al. Inverse effects of midlife occupational and leisure time physical activity on mobility limitation in old age - A 28-year prospective follow-up study. J Am Geriatr Soc. 2014;62(5):812–820.

[46] Holtermann A, Hansen J, Burr H, Sogaard K, Sjogaard G. The health paradox of occupational and leisure-time physical activity. Br J Sport Med. 2012;46(4):291–295.

[47] Hu GC, Chien KL, Hsieh SF, Chen CY, Tsai WH, Su TC. Occupational Versus Leisure-Time Physical Activity in Reducing Cardiovascular Risks and Mortality Among Ethnic Chinese Adults in Taiwan. Asia-Pac J Public He. 2014;26(6):604–613.

[48] Besser L, Dannenberg A. Walking to Public Transit Steps to Help Meet Physical Activity Recommendations. Am J Prev Med. 2005;29(4):273–280.

[49] de Nazelle A, Nieuwenhuijsen M, Anto JM, Brauer M, Briggs D, Braun-Fahrlander C, et al. Improving health through policies that promote active travel: a review of evidence to support integrated health impact assessment. Environ Int. 2011;37(4):766–777.

[50] Xu H, Wen LM, Rissel C. The relationships between active transport to work or school and cardiovascular health or body weight: a systematic review. Asia-Pac J Public He. 2013;25(4):298–315.

[51] Hamer M, Chida Y. Active commuting and cardiovascular risk : A meta-analytic review. Prev Med. 2008;46:9–13.

[52] Humphreys DK, Goodman A, Ogilvie D. Associations between active commuting and physical and mental wellbeing. Prev Med. 2013;57(2):135–9.

[53] Wen LM, Rissel C. Inverse associations between cycling to work , public transport , and overweight and obesity : Findings from a population based study in Australia. Prev Med. 2008;46:29–32.

[54] Reynolds R, McKenzie S, Allender S, Brown K, Foulkes C. Systematic review of incidental physical activity community interventions. Prev Med. 2014;67:46–64.

[55] Scheepers C, Wendel-Vos G, den Broeder J, van Kempen E, van Wesemael P, Schuit A. Shifting from car to active transport: A systematic review of the effectiveness of interventions. Transp Res Part A. 2014;70:264–280.

[56] Brockman R, Fox K. Physical activity by stealth? The potential health benefits of a workplace transport plan. Public Health. 2011;125(4):210–6.

[57] Keating D. Social Inequality in Population Developmental Health. An Equity and Justice Issue.. vol. 50. 1st ed.; 2016.

[58] Kaplan GA, Keil JE. Socioeconomic Factors and Cardiovascular Disease : A Review of the Literature. Circulation. 1993;88:1973–1998.

[59] WHO. Uncovering health inequalities : A path towards leaving no one behind; 2016.

[60] Wilkinson R, Pickett K. The Spirit Level: Why Equality is Better for Everyone. London: Penguin Books; 2010.

[61] Caspersen C, Powell K, Christenson G. Physical activity, exercise, and physical fitness: definitions and distinctions for health-related research. Public Heal Rep. 1985;100(2):126–31.

[62] WHO. Global recommendations on Physical Activity for Health; 2010.

[63] Heil D, Brage S, Rothney M. Modeling physical activity outcomes from wearable monitors. Med Sci Sport Exer. 2012;44(SUPPL 1):50–60.

[64] He Z, Jin L. Activity Recognition from acceleration data Based on Discrete Consine Transform and SVM. In: Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics. San Antonio, Texas, USA; 2009. .

[65] Lee P, Macfarlane D, Stewart S. Validity of the international physical activity questionnaire short form IPAQ-SF A systematic review. Int J Behav Nutr Phy Act. 2011;8.

[66] Cerin E, Cain KL, Oyeyemi AL, Owen N, Conway TL, Cochrane T, et al. Correlates of Agreement between Accelerometry and Self-reported Physical Activity. Med Sci Sport Exer. 2016;48(6):1075–84.

[67] Van Holle V, De Bourdeaudhuij I, Deforche B, Van Cauwenberg J, Van Dyck D. Assessment of physical activity in older Belgian adults: validity and reliability of an adapted interview version of the long International Physical Activity Questionnaire (IPAQ-L). BMC Public Health. 2015;15:433.

[68] Kelly P, Fitzsimons C, Baker G. Should we reframe how we think about physical activity and sedentary behaviour measurement? Validity and reliability reconsidered. Int J Behav Nutr Phy Act. 2016;13:32.

[69] Crouter S, Horton M, Bassett D. Validity of ActiGraph Child-Specific Equations during Various Physical Activities. Med Sci Sport Exer. 2013;45(7):1403–1409.

[70] Craig C, Marshall A, Sjöström M, Bauman A, Booth M, Ainsworth B, et al. International Physical Activity Questionnaire: 12-Country Reliability and Validity. Med Sci Sport Exer. 2003;35(8):1381–1395.

[71] Kozey S, Lyden K, Howe C, Staudenmayer J, Freedson P. Accelerometer Output and MET Values of Common Physical Activities. Med Sci Sport Exer. 2010;42(9):1776–1784.

[72] Ainsworth B, Cahalin L, Buman M, Ross R. The current state of physical activity assessment tools. Prog Cardiovasc Dis. 2015;57(4):387–95.

[73] Troiano R, Berrigan D, Dodd K, Mâsse L, Tilert T, Mcdowell M. Physical activity in the United States measured by accelerometer. Med Sci Sport Exer. 2008;40(1):181–188.

[74] Freedson P, Melanson E, Sirard J. Calibration of the Computer Science and Applications, Inc. accelerometer. Med Sci Sport Exer. 1998;30(5):777–781.

[75] Peach D, Van Hoomissen J, Callender H. Exploring the ActiLife filtration algorithm: converting raw acceleration data to counts. Physiol Meas. 2014;35:2359–2367.

[76] Kozey-Keadle S, Shiroma E, Freedson P, Lee IM. Impact of accelerometer data processing decisions on the sample size, wear time and physical activity level of a large cohort study. BMC Public Health. 2014;14:1210.

[77] Thiese M, Hegmann T K adn Behrens, Garg A, Porucznik C. Important Differences in Accelerometer Cut Points for Quantifying Physical Activity in a Nested Occupational Cohort. J Exerc Sport Orthop. 2014;1(1):12.

[78] Kim Y, Beets M, Welk G. Everything you wanted to know about selecting the "right" Actigraph accelerometer cut-points for youth, but. . .: A systematic review. J Sci Med Sport. 2012;15:311–321.

[79] Brazendale K, Beets M, Bornstein D, Moore J, Pate R, Weaver R, et al. Equating accelerometer estimates among youth: The Rosetta Stone 2. J Sci Med Sport. 2016;19:242–249.

[80] McClain J, Abraham T, Brusseau T, Tudor-Locke C. Epoch length and accelerometer outputs in children: Comparison to direct observation. Med Sci Sport Exer. 2008;40(12):2080–2087.

[81] Cain K, Sallis J, Conway T, Van Dyck D, Calhoon L. Using accelerometers in youth physical activity studies: a review of methods. J Phys Act Heal. 2013;10(3):437–50.

[82] Evenson K, Wen F, Herring A, Di C, LaMonte M, Fels Tinker L, et al. Calibrating physical activity intensity for hip-worn accelerometry in women age 60 to 91 years: The Women's Health Initiative OPACH Calibration Study. Prev Med. 2015;2:750–756.

[83] Colley R, Harvey A, Grattan K, Adamo K. Impact of accelerometer epoch length on physical activity and sedentary behaviour outcomes for preschool-aged children. Heal Rep. 2014;25(1):3–9.

[84] Romanzini M, Petroski EL, Ohara D, Dourado AC, Reichert FF. Calibration of ActiGraph GT3X, Actical and RT3 accelerometers in adolescents. Eur J Sport Sci. 2014;14(1):91–99.

[85] Kerr J, Duncan S, Schipperijn J. Using Global Positioning Systems in Health Research: A Practical Approach to Data Collection and Processing. Am J Prev Med. 2011;41(5):532–540.

[86] Southward E, Page A, Wheeler B, Cooper A. Contribution of the school journey to daily physical activity in children aged 11-12 years. Am J Prev Med. 2012;43(2):201–4.

[87] Ellis K, Marshall S, Chen J, Kerr J. Physical Activity Recognition in Free-living from Body-worn Sensors. In: ACM, editor. Proc. 4th Int. SenseCam Pervasive Imaging Conf. San Diego (CA); 2013. p. 88–89.

[88] Auld J, Williams C, Mohammadian A, Nelson P. An automated GPS-based prompted recall survey with learning algorithms. Transportation Letters. 2009;1(1):59–79.

[89] Shareck M, Kestens Y, Gauvin L. Examining the spatial congruence between data obtained with a novel activity location questionnaire, continuous GPS tracking, and prompted recall surveys. Int J Heal Geogr. 2013;12(40).

[90] Ainsworth B, Haskell W, Herrmann S, Meckes N, Bassett Jr D, Tudor-Locke C, et al. 2011 Compendium of Physical Activities: a second update of codes and MET values. Med Sci Sport Exer. 2011;43(8):1575–1581.

[91] Chaix B, Kestens Y, Perchoux C, Karusisi N, Merlo J, Labadi K. An Interactive Mapping Tool to Assess Individual Mobility Patterns in Neighborhood Studies. Am J Prev Med. 2012;43(4):440–450.

[92] Chaix B, Bean K, Leal C, Thomas F, Havard S, Evans D, et al. Individual/neighborhood social factors and blood pressure in the record cohort study: Which risk factors explain the associations? Hypertension. 2010;55(3):769–775.

[93] Brondeel R, Weill A, Thomas F, Chaix B. Use of healthcare services in the residence and workplace neighbourhood: the effect of spatial accessibility to healthcare services. Health Place. 2014;30:127–133.

[94] Karusisi N, Thomas F, Méline J, Brondeel R, Chaix B. Environmental conditions around itineraries to destinations as correlates of walking for transportation among adults: The RECORD cohort study. PLoS One. 2014;9(5).

[95] Chaix B, Kestens Y, Bean K, Leal C, Karusisi N, Meghiref K, et al. Cohort profile: Residential and non-residential environments, individual activity spaces and cardiovascular risk factors and diseases-The RECORD cohort study. Int J Epidemiol. 2012;41(5):1283–1292.

[96] Leal C, Bean K, Thomas F, Chaix B. Are associations between neighborhood socioeconomic characteristics and body mass index or waist circumference based on model extrapolations? Epidemiology. 2011;22(5):694–703.

[97] Robusto KM, Trost SG. Comparison of three generations of ActiGraph activity monitors in children and adolescents. J Sport Sci. 2012;30(13):1429–1435.

[98] Matthews C. Calibration for Accelerometer Output for Adults. Med Sci Sport Exer. 2005;37(11 Suppl):S512–S522.

[99] Choi L, Liu Z, Matthews CE, Buchowski MS. Validation of accelerometer wear and nonwear time classification algorithm. Med Sci Sports Exerc. 2011;43(2):357–64.

[100] Bornstein D, Beets M, Byun W, Welk G, Bottai M, Dowda M, et al. Equating accelerometer estimates of moderate-to-vigorous physical activity: In search of the Rosetta Stone. J Sci Med Sport. 2011;14(5):404–410.

[101] Anokye N, Stamatakis E. Different conceptual constructs for modelling sedentary behaviour and physical activity: the impact on the correlates of behaviour. BMC Res Notes. 2014;7(1):921.

[102] Aadland E, Steene-Johannessen J. The use of individual cut points from treadmill walking to assess free-living moderate to vigorous physical activity in obese subjects by accelerometry: is it useful? BMC Med Res Methodol. 2012;12(1):172.

[103] Ayabe M, Kumahara H, Morimura K, Tanaka H. Epoch length and the physical activity bout analysis: An accelerometry research issue. BMC Res Notes. 2013;6(1):20.

[104] Tarp J, Andersen L, Østergaard L. Quantification of Underestimation of Physical Activity During Cycling to School When Using Accelerometry. J Phys Act Heal. 2015;12:701–707.

[105] Frank L, Schmid T, Sallis J, Chapman J, Saelens B. Linking objectively measured physical activity with objectively measured urban form: Findings from SMARTRAQ. Am J Prev Med. 2005;28(2S2):117–125.

[106] Troped P, Wilson J, Matthews C, Cromley E, Melly S. The built environment and location-based physical activity. Am J Prev Med. 2010;38(4):429–438.

[107] Perchoux C, Kestens Y, Thomas F, Van Hulst A, Thierry B, Chaix B. Assessing patterns of spatial behavior in health studies: Their socio-demographic determinants and associations with transportation modes (the RECORD Cohort Study). Soc Sci Med. 2014;p. 64–73.

[108] Perchoux C, Chaix B, Brondeel R, Kestens Y. Residential buffer, perceived neighborhood, and individual activity space: New refinements in the definition of exposure areas – The RECORD Cohort Study. Health Place. 2016;40:116–122.

[109] Charreire H, Weber C, Chaix B, Salze P, Casey R, Banos A, et al. Identifying built environmental patterns using cluster analysis and GIS: Relationships with walking, cycling and body mass index in French adults. Int J Behav Nutr Phy Act. 2012;9(59).

[110] Wanner M, Martin B, Meier F, Probst-Hensch N, Kriemler S. Effects of filter choice in GT3X accelerometer assessments of free-living activity. Med Sci Sport Exer. 2013;45(1):170–177.

[111] Matthews C, Chen K, Freedson P, Buchowski M, Beech B, Pate R, et al. Amount of time spent in sedentary behaviors in the United States, 2003-2004. Am J Epidemiol. 2008;167(7):875–881.

[112] Velaga N, Quddus M, Bristow A. Developing an Enhanced Weight-Based Topological Map-Matching Algorithm for Intelligent Transport Systems. Transp Res Part C. 2009;17(6):672–683.

[113] Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.

[114] Liaw A, Wiener M. Classification and Regression by randomForest. R news. 2002;2(December):18–22.

[115] Little R, Rubin D. The Analysis of Social Science Data with Missing Values. Soc Methods Res. 1989;18:292–326.

[116] Schafer J. Multiple imputation: a primer. Stat Methods Med Res. 1999;8(1):3–15.

[117] Flexible imputation of missing data. Boca Raton, FL: Chapman & Hall/CRC; 2012.

[118] Venables W, Ripley B. Modern Applied Statistics with S. 4th ed. New York: Springer; 2002.

[119] Van Buuren S, Groothuis-Oudshoorn K. Multivariate Imputation by Chained Equations. J Stat Softw. 2011;45(3):1–67.

[120] Lorenc T, Petticrew M, Welch V, Tugwell P. What types of interventions generate inequalities? Evidence from systematic reviews. J Epidemiol Commun H. 2013;67:190–193.

[121] Badland H, Schofield G, Garrett N. Travel behavior and objectively measured urban design variables: Associations for adults traveling to work. Health Place. 2008;14:85–95.

[122] Van Dyck D, Cardon G, Deforche B, Sallis J, Owen N, De Bourdeaudhuij I. Neighborhood SES and walkability are related to physical activity behavior in Belgian adults. Prev Med. 2010;50:S74–S79.

[123] Giles-Corti R Billie adn Donovan. Socioeconomic Status Differences in Recreational Physical Activity Levels and Real and Perceived Access to a Supportive Physical Environment. Prev Med. 2002;35:601–611.

[124] Delbosc G Alexa adn Currie. Transport problems that matter – social and psychological links to transport disadvantage. J Transp Geogr. 2011;19:170–178.

[125] Steg L. Car use: lust and must. Instrumental, symbolic and affective motives for car use. Transp Res Part A. 2005;39:147–162.

[126] Pereira R, Nadalin V, Monasterio L, Albuquerque P. Urban Centrality: A Simple Index. Geogr Anal. 2013;45:77–89.

[127] Lee IM, Shiroma E. Using accelerometers to measure physical activity in large-scale epidemiological studies: issues and challenges. Br J Sport Med. 2014;48:197–201.

[128] Staudenmayer J, Pober D, Crouter S, Bassett D, Freedson P. An artificial neural network to estimate physical activity energy expenditure and identify physical activity type from an accelerometer. J Appl Physiol. 2009;107:1300–1307.

Appendices

Supplemental material for Article 1: Measuring moderate-to-vigorous physical activity for adults: the impact of the epoch length

Supplementary material 1: Technical details of data collection and processing

Data collection	
Manufacturer	ActiGraph
Type	GT3X+
Wear location	Right hip
Sampling rate	30 Hz
Sample range	± 6 G
Sensitivity	3 mg/LSB
Data processing	
Data type downloaded from device	Raw data (in G-units)
Bandpass filter	Standard filter; LFE-filter where indicated. Details not released by manufacturer
Epoch length	1-s, 15-s, 30-s and 60-s
Cut-point activity intensity	Tri-axial MVPA [1]: > 2690 VM-CPM Tri-axial LPA: 200 - 2691 VM-CPM Tri-axial SB [2]: < 200 VM-CPM Uni-axial MVPA [3]: > 1951 CPM Uni-axial LPA: 100 1951 CPM Uni-axial SB [4]: < 100 CPM
Axis	Tri-axial measures were based on vector magnitude counts; Uni-axial measures were based on vertical axis counts.
Bouts	No bouts applied
Data reduction	
Non-wear detection	1-min time intervals with consecutive zero counts for at least 90-min time window, allowing a short time intervals with nonzero counts lasting up to 2 minutes if no counts are detected during both the 30-min upstream and downstream from that interval; any nonzero counts except the allowed short interval are considered as wearing. [5] Non-wear were calculated with 60-s epoch counts; which were calculated with the standard bandwidth filter.
Valid days	Minimum 10 h wear time
Software	ActiLife 6 was used for applying the bandpass filters, calculating counts and non-wear detection. R 3.3.0 was used for the calculation of activity intensity levels per epoch and per day, and for the valid day detection.

Hz: Hertz; G: gravitational unit; mg: milli-G; LSB: least significant bit; LFE: low frequency extension; MVPA: moderate-to-vigorous physical activity; SB: sedentary behavior; LPA: light physical activity; CPM: counts per minute; VM-CPM: vector magnitude CPM; [1] MVPA cut-point calibrated by Sasaki et al. (2011) for adults; [2] SB cut-point calibrated by Aguilar et al. (2014) for older adults; [3] MVPA cut-point calibrated by Freedson et al. (1998) for young adults; [4] SB cut-point calibrated by Treuth et al (2004) for adolescent girls and confirmed by Matthews et al. (2008) for adults; [5] algorithm proposed by Choi et al. (2011)

Supplementary material 2: Daily minutes spent doing MVPA, LPA and SB by epoch length and by location type using vector magnitude and vertical axis cut-points

	Median and IQR by epoch length				Mean and 95% CI by epoch length			
	ep01s	ep15s	ep60s	ep60s	ep01s	ep15s	ep60s	ep60s
Sasaki MVPA								
Total	90	60	50	42	96	66	57	49
	(65; 120)	(39; 86)	(31; 76)	(23; 67)	(91; 100)	(63; 70)	(53; 61)	(45; 52)
Leisu	47	24	18	13	54	33	26	20
	(26; 74)	(13; 44)	(9; 32)	(6; 25)	(51; 58)	(30; 35)	(23; 28)	(18; 22)
Trans	21	18	17	16	27	25	24	23
	(7; 41)	(5; 38)	(4; 37)	(3; 36)	(25; 30)	(23; 27)	(22; 26)	(21; 25)
Occup	0	0	0	0	14	9	7	6
	(0; 22)	(0; 12)	(0; 9)	(0; 6)	(12; 17)	(7; 11)	(6; 9)	(4; 7)
Tri-axial LPA								
Total	114	248	284	322	121	255	293	330
	(87; 148)	(192; 310)	(223; 357)	(252; 402)	(115; 126)	(245; 265)	(282; 304)	(318; 342)
Leisu	77	173	202	229	83	178	206	233
	(43; 118)	(96; 247)	(112; 286)	(128; 323)	(77; 89)	(166; 189)	(193; 219)	(219; 247)
Trans	16	28	30	31	19	32	35	38
	(8; 25)	(14; 44)	(14; 48)	(15; 53)	(18; 20)	(30; 34)	(33; 38)	(36; 41)
Occup	0	0	0	0	19	45	52	59
	(0; 33)	(0; 79)	(0; 91)	(0; 102)	(16; 22)	(37; 52)	(43; 61)	(49; 68)
Aguilar SB								
Total	640	533	503	476	635	529	500	472
	(551; 731)	(443; 627)	(410; 598)	(382; 569)	(623; 646)	(516; 542)	(487; 513)	(458; 486)
Leisu	445	350	324	299	422	349	327	306
	(273; 579)	(222; 484)	(205; 456)	(187; 426)	(403; 441)	(332; 366)	(311; 344)	(290; 322)
Trans	62	50	46	45	79	68	66	64
	(26; 111)	(19; 97)	(18; 93)	(15; 91)	(74; 85)	(63; 73)	(61; 71)	(59; 69)
Occup	0	0	0	0	133	112	107	102
	(0; 316)	(0; 256)	(0; 234)	(0; 213)	(114; 152)	(96; 129)	(91; 122)	(87; 117)
Freedson MVPA								
Total	53	42	37	32	58	47	43	40
	(34; 76)	(23; 64)	(18; 60)	(14; 56)	(55; 61)	(44; 51)	(40; 47)	(36; 43)
Leisu	21	12	10	7	27	19	16	13
	(12; 35)	(6; 24)	(4; 19)	(3; 16)	(25; 29)	(17; 21)	(14; 18)	(12; 15)
Trans	17	15	15	15	24	23	22	22
	(4; 36)	(3; 35)	(3; 35)	(2; 35)	(21; 26)	(21; 25)	(20; 25)	(20; 24)
Occup	0	0	0	0	8	6	5	4
	(0; 12)	(0; 8)	(0; 6)	(0; 4)	(7; 10)	(5; 7)	(4; 6)	(3; 5)
Uni-axial LPA								
Total	73	170	206	242	78	179	216	253
	(55; 97)	(130; 222)	(160; 266)	(188; 313)	(75; 82)	(172; 187)	(207; 225)	(243; 263)
Leisu	50	120	146	172	54	127	153	180
	(27; 75)	(67; 177)	(82; 214)	(96; 252)	(50; 58)	(119; 135)	(144; 163)	(169; 191)
Trans	10	19	22	25	13	24	28	31
	(5; 17)	(10; 32)	(10; 37)	(12; 42)	(12; 14)	(22; 26)	(26; 29)	(29; 33)
Occup	0	0	0	0	11	28	35	42
	(0; 18)	(0; 46)	(0; 58)	(0; 70)	(9; 13)	(23; 34)	(29; 42)	(34; 49)
Matthew SB								
Total	724	633	597	566	715	624	591	558
	(641; 809)	(541; 718)	(508; 688)	(471; 656)	(705; 726)	(613; 635)	(580; 603)	(546; 571)
Leisu	525	438	404	377	479	414	390	366
	(306; 657)	(264; 570)	(248; 537)	(228; 510)	(457; 500)	(394; 433)	(371; 409)	(348; 384)
Trans	71	61	58	54	89	78	75	72
	(31; 123)	(25; 109)	(22; 106)	(20; 101)	(83; 95)	(73; 84)	(70; 81)	(67; 77)
Occup	0	0	0	0	147	132	126	120
	(0; 359)	(0; 308)	(0; 291)	(0; 276)	(126; 167)	(113; 150)	(108; 143)	(103; 137)

Number of valid observation days = 1389; Number of persons = 224; Mean wear time during valid observation days = 858 min (= 14 h 18 min); MVPA: moderate-to-vigorous physical activity; IQR: interquartile range; CI: confidence intervals; Leisu: Leisure-time; Trans: transport time; Occup: occupational time.

Supplemental material for Article 2: Using GPS, GIS, and accelerometer data to predict transportation modes

Supplemental Digital Content table 1: Overview of 170 predictors used in the random forest models

GPS variables (53 variables)
Number of GPS observations
Number of valid GPS observations
Percentage of GPS observations to theoretical number of observations (1 per 5 seconds)
Percentage of valid GPS observations to theoretical number of observations
Speed per epoch: 7 summary statistics[a]
Height per epoch: 7 summary statistics[a]
Positional dilution of precision: 7 summary statistics[a]
Vertical dilution of precision: 7 summary statistics[a]
Horizontal dilution of precision: 7 summary statistics[a]
Number of satelites used: 7 summary statistics[a]
Number of satelites in view: 7 summary statistics[a]
Accelerometer (48 variables with the regular filter and 48 with a low frequency filter)
Number of steps per epoch: 7 summary statistics[a]
Acceleration on X-axis per epoch: 7 summary statistics[a]
Acceleration on Y-axis per epoch: 7 summary statistics[a]
Acceleration on Z-axis per epoch: 7 summary statistics[a]
Kilo calories used per epoch: 7 summary statistics[a]
Vector Magnitude per epoch: 7 summary statistics[a]
Total number of steps
Total sedentary epochs
Total epochs with moderate to vigorous physical activity
Total number of kilocalories used
Percentage of epochs sitting
Percentage of epochs with moderate to vigorous physical activity
Combining GPS with GIS data (8 variables)
Straight line distance
Shortest street network distance considering street network restrictions for cars
Shortest street network distance not considering street network restrictions for cars
MapMatched distance
Speed given these 4 distances
Other trip information (5 variables)
Departure in Paris (or not)
Arrival in Paris (or not)
Duration of trip
Trip made during a weekend or not
Time of day of departure (morning, afternoon, evening, night)
Participants information (8 variables)
Residence in Paris (or not)
Gender
Age
In possession of a driving license
In possession of a car
In possession of a motorbike
In possession of a bike
In possession of a public transport pass

[a]7 summary statistics: Minimum, maximum, 10th and 90th quantile, median, mean and standard deviation

Supplemental material for Article 3: Associations of socioeconomic status with transport-related physical activity: combining a household travel survey and accelerometer data using random forests

Supplemental material 1: Overview of the demographic characteristics of the background population (people between 35 and 83 years old in Ile-de-France), the EGT sample and the RECORD GPS sample

	I-d-F [a] (%)	EGT [b] (%)	RECORD (%)
Gender			
Female	52	53	37
Male	48	47	63
Age [c]			
35-44 years	30	32	16
45-59 years	39	37	36
60-74 years	24	25	41
75-83 years	7	6	7
Location of residence			
Inner city (Paris)	19	14	26
First crown of counties around Paris	37	36	41
Second crown of counties around Paris	44	51	30
Population / sample size	5,887,647	21,332	236

[a] I-d-F: 2012 Census data from Ile-de-France, the French capital region; [b] EGT: Enquete globale transport; [c] The data for the age groups 35-44 and 75-83 were not available in the population statistics. The percentages for these categories are based on the assumption that the distribution within the broader category is uniform; [d] The categorization of urbanicity is based on an official administrative subdivision of the Ile-de-France region.

```
library(data.table)
library(randomForest)
library(mice)

#######################################################
# A Construct prediction model MVPA based on RECORD data
#######################################################
path <- "~/.../data/"
rec <- data.table(read.csv(paste0(path, "1. RECORD.csv")))

# 1. Impute missing values in RECORD dataset
# imputations are based on a Random Forest multiple imputation (1
iteration)

# 1.1 order variables in number of missing values.
# this will help the efficiency of the imputation process
seq <- dimnames(md.pattern(rec[,4:ncol(rec), with=FALSE]))[[2]]
seq <- seq[-length(seq)]
seq <- c(c("trip_code", "depcom_res", "dciris_res"), seq)
rec <- rec[, seq, with=FALSE]

# 1.2 Use mice() with the maximum number of iterations maxit set to
zero.
# This is a fast way to create the mids object called ini
# containing the default settings.
# (Van Buuren S, Groothuis-Oudshoorn K.
# mice: Multivariate Imputation by Chained Equations in R.
# Journal of Statistical Software. 2011;45(3):1-67.)
rec.ini <- copy(rec)

rec.ini[,':=' (trip_code='1', depcom_res = 1, dciris_res= 1)]
ini <- mice(rec.ini, max=0, meth='rf')

meth <- ini$meth
pred <- ini$pred
vis <- ini$vis

# 1.3 use these setting
# Method (meth, here random forest), predictors per variable imputed
(pred) and
# visiting sequence (vis)
mi.rf <- mice(rec, m=1, maxit = 5, pred=pred, meth=meth, vis=vis)

# 1.4 creating a dataset with all missings imputed
rec.nomiss <- complete(mi.rf)

# 2 MVPA prediction model on RECORD data

form.mv <- formula(mvpa_ep1m ~ mode_trans1 + duration_mn +
                     time_of_day + day_trip + rush_hour +
                     age + homme + dist_ld + speed_ld + rvnu +
                     emploi_sim + nivetude_sim +
                     dist_train_dep + dist_metro_dep + dist_tram_dep
+ dist_bus_dep +
```

```
                    dist_train_arr + dist_metro_arr + dist_tram_arr
+ dist_bus_arr +
                    dist_train_res + dist_metro_res + dist_tram_res
+ dist_bus_res +
                    dist_pt_res + dist_pt_dep + dist_pt_arr +
                    educ_res + educ_dep + educ_arr +
                    intersec_res + intersec_dep + intersec_arr +
                    dest_res + dest_dep + dest_arr +
                    park_res + park_dep + park_arr +
                    pdens_res + pdens_dep + pdens_arr +
                    res_cour + dep_cour + arr_cour +
                    pos.voiture + pos.moto + pos.TC +
pos.motorized)

fit.mvp <- randomForest(form.mv, data=rec.nomiss, ntree = 1000)


# list of 15 most important variables

a <- data.frame(importance(fit.mvp))
a$Variables <- rownames(a); rownames(a) <- NULL
a[order(a$IncNodePurity, decreasing=TRUE),c('Variables',
'IncNodePurity')][1:15,]

#         Variables IncNodePurity
# 1     mode_trans1    158005.805
# 2     duration_mn    113574.763
# 8         dist_ld     54074.314
# 9        speed_ld     48475.942
# 18 dist_metro_arr     12746.681
# 19  dist_tram_arr     12178.505
# 4        day_trip     10863.437
# 39       park_arr      9102.616
# 14 dist_metro_dep      8925.460
# 38       park_dep      8772.124
# 13 dist_train_dep      8584.357
# 41      pdens_dep      8534.186
# 33   intersec_arr      8323.625
# 32   intersec_dep      8078.804
# 15  dist_tram_dep      8004.453

# to visualize importance of variables
varImpPlot(fit.mvp)

#########################################################
# B Prediction of MVPA for EGT trips
#########################################################
path <- "~/.../data/"
egt <- data.table(read.csv(paste(path, "2. EGT.csv", sep="")))

# 1 Imputation of missing values in EGT datasets
#   the imputation will enable MVPA predictions for all trips
#   The imputations are based on predictive mean matching models
```

```
# 1.1 Ordering the variables on the amount of missing values
#     while making sure id variables won't be used in the imputation
seq <- dimnames(md.pattern(egt[,5:ncol(egt), with=FALSE]))[[2]]
seq <- seq[-length(seq)]
seq <- c(c("trip_code", "resc", "depcom_res", "dciris_res"), seq)
egt <- egt[, seq, with=FALSE]
egt.ini <- copy(egt)
egt.ini[,':=' (trip_code='1', resc = '1', depcom_res = 1,
dciris_res= 1)]

# 1.2 Use mice() with the maximum number of iterations maxit set to
zero.
#     This is a fast way to create the mids object called ini
#     containing the default settings.

ini <- mice(egt.ini, max=0)
meth <- ini$meth
pred <- ini$pred # since in egt.ini the id variables are constant,
the pred is already==0
vis <- ini$vis

# 1.3 Actual imputation of EGT dataset
mi.data <- mice(egt, m=1, maxit = 5, pred=pred, meth=meth, vis=vis)
egt.nomiss <- complete(mi.data)


# 2. Prediction of MVPA for each EGT trip
#   Using the new egt.nomiss dataset and
#   MVPA-prediction model on RECORD data
#   Note: the original not-imputed EGT dataset is used after this
step
#   egt.nomiss is only used for these predictions
egt[, pred.mvpa := predict(fit.mvp, egt.nomiss)]


# 3. Some variables for the regression analysis
# 3.1 Variable Weekend-Weekday
tmp <- data.table(day_trip=c('1. Monday', '2. Tuesday', '3.
Wednesday', '4. Thursday', '5. Friday', '6. Saturday', '7. Sunday'),
                  weekday=as.factor(c(rep('1. weekday', 5), rep('2.
weekend',2))))
egt <- merge(egt, tmp, by='day_trip', all.x=T)

# 3.2 Creating an id variable for the person
a <- unlist(strsplit(as.character(egt$trip_code), '_'))
men <- a[seq(1,length(a), 3)]
per <- a[seq(2,length(a), 3)]
egt$person <- paste(men, per, sep='_')

#########################################################
# C Construction of day-level EGT dataset
#########################################################
# 1. Construction of day level variables
# 1.1 MVPA per day, minutes in transport per day and number of trips
```

```
setkey(egt, person)
egt[,V1 := 1]

var1 <- c('pred.mvpa',  'duration_mn', 'V1') #'pred.mf.mvpa',
var2 <- c('mvpa.day',  'min.day', 'nb_trips') #'mvpa.mf.day',
egt[, var2 := lapply(.SD, sum, na.rm=TRUE), by=person, .SDcols=var1,
with=FALSE]


# 1.2 Day-level variables per type of transportation mode
# 1.2.1 MVPA (so how much each person profits of each transportation
mode in terms of MVPA)

setkey(egt, person, mode_trans1)
mvpa_by_mt <- egt[, sum(pred.mvpa),by=list(person,mode_trans1)]

setkey(mvpa_by_mt, person, mode_trans1)
out <- mvpa_by_mt[CJ(unique(person), unique(mode_trans1))][,
as.list(V1), by=person]

setnames(out, paste('V', 1:5, sep=''),
         paste("MVPA_", c('NA', "walking", "biking", "PM", "PT"),
sep=''))
var <- paste("MVPA_", c('NA', "walking", "biking", "PM", "PT"),
sep='')
replacena <- function(var){var <- replace(var, is.na(var), 0)}
out[,var := lapply(.SD, replacena),.SDcols=var, with=FALSE]

egt <- merge(egt, out, by='person')

# 1.2.2 Minutes in transport

setkey(egt, person, mode_trans1)
min_by_mt <- egt[, sum(duration_mn),by=list(person,mode_trans1)]

setkey(min_by_mt, person, mode_trans1)
out <- min_by_mt[CJ(unique(person), unique(mode_trans1))][,
as.list(V1), by=person]

setnames(out, paste('V', 1:5, sep=''),
         paste("MIN_", c('NA', "walking", "biking", "PM", "PT"),
sep=''))
var <- paste("MIN_", c('NA', "walking", "biking", "PM", "PT"),
sep='')
replacena <- function(var){var <- replace(var, is.na(var), 0)}
out[,var := lapply(.SD, replacena),.SDcols=var, with=FALSE]

egt <- merge(egt, out, by='person')

# 1.2.3 Number of trips

setkey(egt, person, mode_trans1)
nb_by_mt <- egt[, sum(V1), by=list(person,mode_trans1)]
```

```
setkey(nb_by_mt, person, mode_trans1)
out <- nb_by_mt[CJ(unique(person), unique(mode_trans1))][,
as.list(V1), by=person]

setnames(out, paste('V', 1:5, sep=''),
         paste("nb_", c('NA', "walking", "biking", "PM", "PT"),
sep=''))
var <- paste("nb_", c('NA', "walking", "biking", "PM", "PT"),
sep='')
replacena <- function(var){var <- replace(var, is.na(var), 0)}
out[,var := lapply(.SD, replacena),.SDcols=var, with=FALSE]
egt[,V1 := NULL]
egt <- merge(egt, out, by='person')

# 2. Add people with no trips
#   2066 people were not in the trip-dataset,
#   because they reported no trips at all during the day of
observation

egtnt <- data.table(read.csv(paste(path, '2. EGT no trips.csv',
sep="")))

# 2.1 Create the variables in egtnt that were created before in EGT
dataset
varnt <- names(egtnt)[which(names(egtnt) %in% names(egt))]
egtnt2 <- egtnt[,varnt, with=FALSE]

# 2.2 Set these variables to 0
#   (e.g. no transport-related MVPA observed for these people)
egtnt2[, var2:= 0, with=FALSE]
egtnt2[, paste("MVPA_", c('NA', "walking", "biking", "PM", "PT"),
sep=''):= 0, with=FALSE]
egtnt2[, paste("MIN_", c('NA', "walking", "biking", "PM", "PT"),
sep=''):= 0, with=FALSE]
egtnt2[, paste("nb_", c('NA', "walking", "biking", "PM", "PT"),
sep=''):= 0, with=FALSE]
egtnt2[, c('min.day', 'nb_trips'):= 0, with=FALSE]

# 2.3 Merge EGT dataset with EGT no trips dataset
egt <- rbindlist(list(egt, egtnt2), use.names=TRUE, fill=TRUE)


# 3. aggregate to day level
egt.day <- unique(setkey(egt, person), by='person')


#######################################################
# D recode some variables for the analysis
#######################################################

# 1 Round mvpa variable to the minute.
# This is necessary for count regression
egt[, mvpa.day.int := round(mvpa.day)]
```

```
# 2 Centralize variables for easier interpretable interaction
effects
# and divide variables by 1000 to get an interpretable scale (e.g.
km)

egt[, rvnu.1000 := (rvnu - mean(rvnu, na.rm=TRUE))/1000]
egt[, age.10 := ( age - mean(age, na.rm=TRUE))/10]
egt[, intersec_res.1000 := (intersec_res - mean(intersec_res,
na.rm=TRUE))/1000]

egt[dist_pt_res>1000, dist_pt_res := 1000]
egt[, dist_pt_res.1000 := (dist_pt_res - mean(dist_pt_res,
na.rm=TRUE))/1000]

egt[, educ_res.m := (educ_res  - mean(educ_res, na.rm=TRUE ))]
egt[, dest_res.1000 := (dest_res - mean(dest_res, na.rm=TRUE))/1000]



#########################################################
# E Multiple imputation of EGT day-level dataset
#########################################################
# 1 : imputation of missing values to have a MVPA prediction for all
trips

# 1.1 ordering the variables on the amount of missing values
# while making sure id variables won't be used in the imputation
seq <- dimnames(md.pattern(egt[,6:ncol(egt), with=FALSE]))[[2]]
seq <- seq[-length(seq)]
seq <- c(c('person', 'resc', 'depcom_res', 'dciris_res', 'over'),
seq)
egt <- egt[, seq, with=FALSE]

# 1.2 Use mice() with the maximum number of iterations maxit set to
zero.
# This is a fast way to create the mids object called ini
# containing the default settings.
egt.ini <- copy(egt)

egt.ini[,':=' (person='1', resc = '1',  depcom_res = 1, dciris_res=
1, over = 1)]
ini <- mice(egt.ini, max=0, meth='rf')

meth <- ini$meth
meth[c("person", "resc", "depcom_res", "dciris_res",
       "mvpa.day.int", "age.10", "homme",  "res_cour",
"pos.motorized",
       "weekday",  "min.day",  "MIN_walking",
       "MIN_biking", "MIN_PM", "MIN_PT", "nb_trips",
       "nb_walking", "nb_biking",  "nb_PM",  "nb_PT", "over",
       'intersec_res.1000',  'pdens_res')] <- ""
pred <- ini$pred # since in egt.ini the id variables are constant,
the pred is already==0
vis <- ini$vis
```

```
# 1.3 actual imputation of EGT dataset
# Method is Random Forest, 5 imputations, 100 trees per imputation
mi.rf <- mice(egt, m=5, pred=pred, meth=meth, ntree=100, vis=vis)



#########################################################
# F Negative binomial regression on multiple imputation dataset
#########################################################

library(MASS)
# Fit the model for each of the 5 data sets
fit.nb <- with(mi.rf, glm.nb(mvpa.day.int ~
                               (nivetude_sim  + rvnu.1000)^2 +
                               (rvnu.1000 + pos.motorized)^2 +
                               (rvnu.1000 + dist_pt_res.1000)^2  +
                               educ_res.m + dest_res.1000
                             + age.10 + homme + emploi_sim ))

# Pool the results for the 5 data sets
pnb <- pool(fit.nb)

#########################################################
# G Plots of interaction effects
#########################################################
# 1. Use 'typical' values for variables
#   These values are used for the plots
#   where the variables are not of interest
#   e.g. mean distance to a public transport station will used
#   for the effect plot 'education*income'

# 1.1 typical values for factors:
#   proportions in all categories but the reference category
#   This reflects the use of the first level as the baseline level.
#   Effect Displays in R for Generalised Linear Models (John Fox);
#   journal of statistical software, Vol. 8, Issue 15, Jul 2003

m <- mi.rf$m # number of imputations
typical <- function(var, ref.level){
  Q <- U <- rep(NA, m)
  for (i in 1:m) {
    var1 <- complete(mi.rf, i)[,var]
    var2 <- ifelse(var1 == ref.level, 1, 0)
    Q[i] <- mean(var2)
    U[i] <- var(var2) / nrow(complete(mi.rf, i))  # (standard error
of estimate)^2
  }
  a <- pool.scalar(Q, U, n = nrow(nhanes), k = 1)$qbar
  a
}

typ.etud2 <- typical('nivetude_sim', '2. bac - bacp2')
typ.etud3 <- typical('nivetude_sim', '3. bacp3 et plus')
```

```
typ.moto <- typical('pos.motorized', '1')

typ.empl2 <- typical('emploi_sim', '2. chomage')
typ.empl3 <- typical('emploi_sim', '3. retrait')
typ.empl4 <- typical('emploi_sim', '4. autre')

typ.homm <- typical('homme', '1. male')

# 1.2 'Typical' values for continuous variables: means
mean.pool <- function(var){
  Q <- U <- rep(NA, m)
  for (i in 1:m) {
    var1 <- complete(mi.rf, i)[,var]
    Q[i] <- mean(var1)
    U[i] <- var(var1) / nrow(complete(mi.rf, i))  # (standard error
of estimate)^2
  }
  a <- pool.scalar(Q, U, n = nrow(nhanes), k = 1)$qbar
  a
}
typ.dist <- mean.pool('dist_pt_res.1000')
typ.edre <- mean.pool('educ_res.m')
typ.dest <- mean.pool('dest_res.1000')
typ.ag10 <- mean.pool('age.10')

# 2. Creation of a new dataset
#   This dataset will be used to construct the plot.
#   This part of the script is inspired by:
#   Atkins DC, Gallop RJ. Rethinking how family researchers model
infrequent
#   outcomes: a tutorial on count regression and zero-inflated
models.
#   J Fam Psychol 2007;21:726-35.
#   The variables of interest have values over their full range
#   The other variables have a 'typical' value (see above)


newdata <- expand.grid(
  intercept = 1,
  nivetude_sim = c(0,1,2),
  rvnu.1000 =  seq(from=-2.5, to=3.5, by=0.01),
  pos.motorized = typ.moto,
  dist_pt_res.1000 = typ.dist,
  educ_res.m = typ.edre,
  dest_res.1000 = typ.dest,
  emploi_sim2 = typ.empl2,
  emploi_sim3 = typ.empl3,
  emploi_sim4 = typ.empl4,
  age.10 = typ.ag10,
  homme = typ.homm
)
newdata$nivetude_sim2 <- ifelse(newdata$nivetude_sim == 1, 1, 0)
newdata$nivetude_sim3 <- ifelse(newdata$nivetude_sim == 2, 1, 0)
```

```
# 3. Prediction values for the new dataset,
#    based on the negative binomial model
pred <- function(data){

  data$rvnu.etud2 <- data$rvnu.1000*data$nivetude_sim2
  data$rvnu.etud3 <- data$rvnu.1000*data$nivetude_sim3
  data$rvnu.motor <- data$rvnu.1000*as.numeric(as.character(data
$pos.motorized))
  data$rvnu.di.pt <- data$rvnu.1000*data$dist_pt_res.1000
  data$rvnu <- data$rvnu.1000*1000+3481.662 #set rvnu back to
original scale for plotting purposes
  data$nivetude_sim <- 0
  data$nivetude_sim[which(data$nivetude_sim2 == 1)] <- 1
  data$nivetude_sim[which(data$nivetude_sim3 == 1)] <- 2

  data2 <- data[, c("intercept", "nivetude_sim2",
"nivetude_sim3",
                          "rvnu.1000", "pos.motorized",
"dist_pt_res.1000",
                          "educ_res.m", "dest_res.1000",
                          "emploi_sim2", "emploi_sim3",
"emploi_sim4",
                          "age.10", "homme",
                          "rvnu.etud2", "rvnu.etud3", "rvnu.motor",
"rvnu.di.pt")]
  l <- t(data2)
  # below: Coefficient and variance-covariance matrix are used
  #        to predict point estimates and confidence bands
  predict.data <- data.frame(matrix(c(pnb$qbar %*% l,
                                      pnb$qbar %*% l - 1.96 *
sqrt(diag(t(l) %*% pnb$ubar %*% l)),
                                      pnb$qbar %*% l + 1.96 *
sqrt(diag(t(l) %*% pnb$ubar %*% l))),
                                    ncol=3, dimnames=list(NULL,
c("Estimate", "LL.95", "UL.95"))))
  data[c("Estimate","LL.95","UL.95")] <-
predict.data[c("Estimate","LL.95","UL.95")]
  data$Estimate <- exp(data$Estimate)
  data$LL.95 <- exp(data$LL.95)
  data$UL.95 <- exp(data$UL.95)
  data
}

plotdata1 <- pred(newdata)

# 4. Create plotting function
#   This function enables interaction plots for a continuous
variable
#   and a continuous or categorical variable.
#   For the latter, 2 or 3 values can be chosen
plot.int <- function(data, var, value1, value2, value3=NA){
  plot(Estimate ~ rvnu, data=data, type="n",
       ylim=c(min(data$LL.95)-1,max(data$UL.95))+0.5,
```

```
      xlab = "Household income",
      ylab= "Minutes T-MVPA", cex.lab=1,cex.axis=0.75) #
  # plot interval slope group 1
  with(subset(data, data[,which(names(data) == var)] == value1), {
    lines(rvnu, LL.95, lty=1)
    lines(rvnu, UL.95, lty=1)
    lines(x = rvnu, y = Estimate, lty=1, lwd = 1)
  })
  # plot interval slope group 2
  with(subset(data, data[,which(names(data) == var)] == value2), {
    lines(rvnu, LL.95, lty=3)
    lines(rvnu, UL.95, lty=3)
    lines(x = rvnu, y = Estimate, lty=3, lwd = 1)
  })
  # plot slope group 3
  if(!is.na(waarde3)) {
    with(subset(data, data[,which(names(data) == var)] == value3), {
      lines(rvnu, LL.95, lty=5)
      lines(rvnu, UL.95, lty=5)
      lines(x = rvnu, y = Estimate, lty=5, lwd = 1)
    })
  }
}

# 5. Create JPEG file and apply plotting function
fig <- "C:/.../graph/"
jpeg(paste(fig, 'plot int education level - income.jpg', sep=''),
width = 8.5, height = 8.5, units = "cm",  res = 500, quality = 150)
plot.int(plotdata1, 'nivetude_sim', 0, 1, 2)
legend(1000, 16.15, c("No secondary","Secondary - low
tertiary","Higher tertiary"),
       lty=c(1,3,5), cex=0.45)
dev.off()
```

Supplemental material for Article 4: Simulating the impact of transport mode shifts on transport-related physical activity

Description of the RECORD GPS Study

September 19, 2016

The participants in the RECORD Study (Residential Environment and CORonary heart Disease) were recruited during preventive health checkups in 20072008, and were born in 19281978.[1,2] Every participant residing in 112 pre-selected municipalities of the Ile-de-France Paris region at baseline presenting at the IPC Medical Center for a health checkup was invited to enter the RECORD Study.[3,4] The selected municipalities of the Ile-de-France region included a broad range of municipalities in terms of household income and urbanicity degree.

In the second wave of the study (2011-2012),[5–8] 410 RECORD participants were invited to enter the RECORD GPS Study.[2,9] Of these, 236 accepted to participate. Participants wore a BT-Q1000XT GPS (QStarz) and a GT3X+ accelerometer (The Actigraph) on the right hip with a dedicated elastic belt, for the recruitment day and 7 additional days, all day long from wake up to bedtime. The participants had to fill out a travel diary by reporting their activity places over the 7-8 days, each time with arrival and departure times.

The GPS data were collected every 5 seconds. After linear interpolation of the missing data, the GPS data were analyzed with an algorithm (ArcGIS Python script) that identified all of the activity locations of the participants (any activity at a stationary location) from the accumulation of GPS points over 7 days.[10] Based on these outputs of the algorithm, the Mobility Web Mapping application was then used to visualize the activity patterns on a map per participant per day. The Mobility Web Mapping application was designed by the University of Montreal. The application was used to survey the participants on the activity performed at each visited location and on the modes used in each trip. The survey operator could report activity locations and trips undetected by the algorithm and could modify/remove detected visits to locations that were inaccurate or incorrect. This procedure resulted in the identification of 7138 trips for 229 participants. Written informed consent was obtained from all participants. The RECORD GPS Study was approved by the French Data Protection Authority.

References

[1] Brondeel R, Weill A, Thomas F, Chaix B. Use of healthcare services in the residence and workplace neighbourhood: the effect of spatial accessibility

to healthcare services. Health Place. 2014;30:127–133.

[2] Chaix B, Kestens Y, Duncan S, Merrien C, Thierry B, Pannier B, et al. Active transportation and public transportation use to achieve physical activity recommendations? A combined GPS, accelerometer, and mobility survey study. Int J Behav Nutr Phys Act. 2014;11(1):124.

[3] Chaix B, Bean K, Daniel M, Zenk SN, Kestens Y, Charreire H, et al. Associations of Supermarket Characteristics with Weight Status and Body Fat: A Multilevel Analysis of Individuals within Supermarkets (RECORD Study). PLoS ONE. 2012;7(4):e32908.

[4] Van Hulst A, Thomas F, Barnett T, Kestens Y, Gauvin L, Pannier B, et al. A typology of neighborhoods and blood pressure in the RECORD Cohort Study. J Hypertens. 2012;30:1336–1346.

[5] Chaix B, Kestens Y, Bean K, Leal C, Karusisi N, Meghiref K, et al. Cohort profile: Residential and non-residential environments, individual activity spaces and cardiovascular risk factors and diseases-The RECORD cohort study. Int J Epidemiol. 2012;41(5):1283–1292.

[6] Chaix B, Kestens Y, Perchoux C, Karusisi N, Merlo J, Labadi K. An Interactive Mapping Tool to Assess Individual Mobility Patterns in Neighborhood Studies. Am J Prev Med. 2012;43(4):440–450.

[7] Leal C, Bean K, Thomas F, Chaix B. Multicollinearity in Associations Between Multiple Environmental Features and Body Weight and Abdominal Fat: Using Matching Techniques to Assess Whether the Associations are Separable. Am J Epidemiol. 2012;175(11):1152–1162.

[8] Perchoux C, Kestens Y, Thomas F, Van Hulst A, Thierry B, Chaix B. Assessing patterns of spatial behavior in health studies: Their socio-demographic determinants and associations with transportation modes (the RECORD Cohort Study). Soc Sci Med. 2014;p. 64–73.

[9] Chaix B, Kestens Y, Duncan DT, Brondeel R, Méline J, El Aarbaoui T, et al. A GPS-based methodology to analyze environment–health associations at the trip level: case-crossover analyses of built environment effects on walking. Am J Epidemiol. (In press);.

[10] Thierry B, Chaix B, Kestens Y. Detecting activity locations from raw GPS data: a novel kernel-based algorithm. Int J Heal Geogr. 2013;12(14).

Supplementary material 2: Overview of the variables used in the data integration

Integrated (predicted) variable
Minutes of transport-related moderate-to-vigorous physical activity (T-MVPA) per trip [a]
Personal variables
Household income [b]
Personal education level [b]
Age [b]
Gender [b]
Work situation (employed, unemployed, retired, other) [b]
A motorized vehicle available in the household [b]
A car available in the household [b]
A motorbike available in the household [b]
In possession of a public transport pass [b]
Spatial access to public transport at the residence
Street network distance to nearest public transport station from residence [c]
Street network distance to nearest train station [c]
Street network distance to nearest metro station [c]
Street network distance to nearest tram station [c]
Street network distance to nearest bus station [c]
Other residential neighborhood characteristics
Educational level in the residential neighborhood [d]
Number of destinations in the residential neighborhood [d]
Number of intersections in the area [d]
Area size of parks in the area [d]
Population density in the area [d]
Address located in Paris, or in the other counties adjacent to Paris, or in the other counties non-adjacent to Paris
Trip characteristics
Transportation mode [e]
Duration of the trip in minutes [e]
Time of the day at departure [e]
Day of the week at departure [e]
Rush hour or not at departure: from 8am to 11am and from 4pm to 7pm [e]
Straight-line distance from departure address to arrival address [e]
Speed based on duration and straight-line distance [e]
Trip departure and arrival location characteristics (2 separate sets of variables)
Distance to nearest train station [c]
Distance to nearest metro station [c]
Distance to nearest tram station [c]
Distance to nearest bus station [c]
Distance to nearest public transport station [c]
Educational level in the area [d]
Number of intersections in the area [d]
Number of destinations in the area [d]
Area size of parks in the area [d]
Population density in the area [d]
Address located in the city center or not (i.e., in Paris as opposed the other parts of Ile-de-France Region)

[a] Accelerometry information in RECORD or predicted time in EGT; [b] RECORD and EGT questionnaires; [c]Shortest street network distance determined with ArcGIS from the residence or from the departure/arrival of each trip geocoded at the center of a 100 m square; [d] The area around the residence or departure or arrival point of each trip was defined with ArcGIS as a 1 km buffer following the street network, and information was aggregated at the level of this buffer; [e]Information from the mobility survey in RECORD and in EGT.

```
################################################################
# Supplemetary material S3 for the article:
# Brondeel R, Kestens Y, Chaix B. The impact of transport interventions
# on transport-related physical activity:  a simulation study based on
# sensor data and random forests'. (2016)
# The results of these data analyses illustrated the impact of
# successfull transport interventions on transport-related MVPA (T-MVPA)
#
# These scripts are not the complete scripts used for this article,
# but only illustrative of the different steps in the
# data integration and simulation process. In this scripts,
# we use the example of the 'promotion of walking'-scenarios.
# Questions can be directed to Ruben.Brondeel@gmail.com
################################################################

#######################
# Two libraries were loaded besides the basic packages in
# R version 3.3.0 (2016-05-03) -- "Supposedly Educational"
# Copyright (C) 2016 The R Foundation for Statistical Computing
# References can be found below, or obtained by running
# citation('randomForest'), citation('data.table') and
# citation('mice') in your R-console

library(randomForest)
library(data.table)
library(mice)

#######################
# Read in EGT data and convert to data tables (instead of data frames)

path <- "~/.../data/"
rec <- data.table(read.csv(paste0(path, "1. RECORD.csv")))
egt <- data.table(read.csv(paste0(path, "1. EGT.csv")))

#######################
# A Random Forest prediction models
#   These models will later be used for the data integration
#   and the simulations.
#   (for more details on the data integration see
#   Brondeel R, Pannier B, Chaix B. Associations of socioeconomic
#   status with transport-related physical activity: combining a
#   household travel survey and accelerometer data using random forests.
#   J Transp Health. In press. http://dx.doi.org/10.1016/j.jth.2016.06.002)

# 1. Transportation mode prediction model in EGT dataset
#    This model is later used to calculate the probability of the
#    transportation modes for each trip

form.tm <- formula(mode_trans1 ~
                   time_of_day + day_trip + rush_hour + dist_ld +
```

```
                age + homme + rvnu + emploi_sim + nivetude_sim +
                dist_train_dep + dist_metro_dep +
                dist_tram_dep + dist_bus_dep +
                dist_train_arr + dist_metro_arr +
                dist_tram_arr + dist_bus_arr +
                dist_train_res + dist_metro_res +
                dist_tram_res + dist_bus_res +
                dist_pt_res + dist_pt_dep + dist_pt_arr +
                educ_res + educ_dep + educ_arr +
                intersec_res + intersec_dep + intersec_arr +
                dest_res + dest_dep + dest_arr +
                park_res + park_dep + park_arr +
                pdens_res + pdens_dep + pdens_arr +
                res_cour + dep_cour + arr_cour)

fit_tm <- randomForest(form.tm, data = egt, ntree = 1000)

# 2. Duration prediction model in EGT dataset
#    This later used in the simulation process to update
#    the duration of the trips after the mode is changed
#    Due to the large number of observations and
#    the continuous outcome variable, growing 1000 trees was not
#    possible due to calculation power. But the model was stable at
#    100 trees in terms of predictability. So , we decided to grow 150 trees

form_dur <- formula(duration_mn ~
                mode_trans1 +
                time_of_day + day_trip + rush_hour + dist_ld +
                age + homme + rvnu + emploi_sim + nivetude_sim +
                dist_train_dep + dist_metro_dep +
                dist_tram_dep + dist_bus_dep +
                dist_train_arr + dist_metro_arr +
                dist_tram_arr + dist_bus_arr +
                dist_train_res + dist_metro_res +
                dist_tram_res + dist_bus_res +
                dist_pt_res + dist_pt_dep + dist_pt_arr +
                educ_res + educ_dep + educ_arr +
                intersec_res + intersec_dep + intersec_arr +
                dest_res + dest_dep + dest_arr +
                park_res + park_dep + park_arr +
                pdens_res + pdens_dep + pdens_arr +
                res_cour + dep_cour + arr_cour )

fit_dur <- randomForest(form_dur, data = egt, ntree = 150)

# 3. T-MVPA prediction model in RECORD data set
#    This model is used for the data integration step and to update
#    the predicted T-MVPA during the simulation process after changing
#    the transportation mode and duration of a trip.
```

```
form_mv <- formula(mvpa_ep1m ~
                   duration_mn + speed_ld +
                   mode_trans1 +
                   time_of_day + day_trip + rush_hour + dist_ld +
                   age + homme + rvnu + emploi_sim + nivetude_sim +
                   dist_train_dep + dist_metro_dep +
                   dist_tram_dep + dist_bus_dep +
                   dist_train_arr + dist_metro_arr +
                   dist_tram_arr + dist_bus_arr +
                   dist_train_res + dist_metro_res +
                   dist_tram_res + dist_bus_res +
                   dist_pt_res + dist_pt_dep + dist_pt_arr +
                   educ_res + educ_dep + educ_arr +
                   intersec_res + intersec_dep + intersec_arr +
                   dest_res + dest_dep + dest_arr +
                   park_res + park_dep + park_arr +
                   pdens_res + pdens_dep + pdens_arr +
                   res_cour + dep_cour + arr_cour )

fit_mvp <- randomForest(form_mv, data = rec, ntree = 1000)


#######################
# B Prediction of MVPA for EGT trips (data integration step)
#   This will result in a predictive T-MVPA value for the observed trips
#   and is based on the above fitted random forest model.

# 1 Imputation of missing values in EGT dataset
#   The imputation will enable MVPA predictions for all trips
#   The imputution process are based on predictive mean matching models

# 1.1 Ordering the variables on the amount of missing values
#     while making sure id variables won't be used in the imputation

seq <- dimnames(md.pattern(egt[,5:ncol(egt), with = FALSE]))[[2]]
seq <- seq[-length(seq)]
seq <- c(c("trip_code", "resc", "depcom_res", "dciris_res"), seq)
egt <- egt[, seq, with = FALSE]
egt.ini <- copy(egt)
egt.ini[,':=' (trip_code = '1', resc = '1', depcom_res = 1, dciris_res = 1)]

# 1.2 Use mice() with the maximum number of iterations maxit set to zero.
#     This is a fast way to create the mids object called ini
#     containing the default settings.

ini <- mice(egt.ini, max = 0)
meth <- ini$meth
pred <- ini$pred
vis <- ini$vis
```

```
# 1.3 Actual imputation of the EGT dataset, only 1 dataframe retained

mi_data <- mice(egt, m = 1, maxit = 5, pred = pred, meth = meth, vis = vis)
egt.nomiss <- complete(mi_data)


# 2. Prediction of MVPA for each EGT trip
#    Using the new egt.nomiss dataset and
#    MVPA-prediction model on RECORD data
#    Note: the original not-imputed EGT dataset is used after this step
#    egt.nomiss is only used for these predictions

egt[, pred_mvpa := predict(fit_mvp, egt.nomiss)]


######################
# C Simulation step 1: Selection of trips
#    Here we give the example for the walking trips.
#    The procedure is similar for the other transportation mdoes
#    We created 100 selection variables (0,1) per scenario
#    by raising or lowering the probabilities,
#    in order to get the search mean probability of selection.
#    The selection variables will be written to separate files per scenario.

# 1. Probability estimates (= votes) from the random forest fitted
#    above added to the dataset

egt <- data.table(fit_tm$votes, egt)
setnames(egt, c('1. walking',  '2. bicycle', '3. motorized', '4. public'),
        c('walking', 'biking', 'motorized', 'public'))

# 2 create 100 empty variables to save the selection variables
#     these variables will serve later to select
#     the trips that will be changed into walking trips

egt.wal <- copy(egt)
egt.wal[, paste0('V', 1:100) := as.integer(NA)]

# 3. setting the probability of a change to 0 for walking trips

egt.wal['1. walking', walking := 0]

# 4.  The following steps calculate the shift in the probability,
#     so that the mean probability corresponds to the respective scenario

# 4.1 The function inv.logit back transforms logit values in probabilities
#     This function is used within the shift_search function
inv.logit <- function(logit.val, adjust) {
  small_shift <- 1 - 2*small_shift
  (small_shift*(1+exp(logit.val))+(exp(logit.val)-1))/
```

```
      (2*small_shift*(1+exp(f)))
}

# 4.2 The function shift_search returns the shift (in logit scale), so that
#     the mean proportion of non-walking trips to be performed by
#     walking trips is equal to the proportion we want to change
#     in the respective scenario.

shift_search <- function(shift, scenario_prop, logit_prob, ratio_mode) {
  logit.w1 <- logit_prob - shift
  proportion_change <- ratio_mode * scenario_prop
  (proportion_change - mean(zapsmall(inv.logit(logit.w1, adjust = .000025))))^2
}

# 4.3 The probability of the non-walking trips is transformed into
#     logit values

setkey(egt.wal, mode_trans1)
logit_prob_walking <- logit(egt.wal[!'1. walking', walking], adjust = .000025)

# 4.4 This is the ratio between walking and non-walking mode.
#     We want to know the proportion of trips within the non-walking modes
#     that is equivalent to the XX% of the trips within the walking modes.
#     so in the shift-function, we need this ratio

ratio_walking_nonwalking <- nrow(egt.wal['1. walking']) /
  nrow(egt[!'1. walking'])

# 4.5 The function optimize returns the shift value for which the
#     proportion change is equal to the mean of the shifted probabilities.

logit_shift_s10 <- - optimize(f = shift_search, interval = c(-4,4),
                          scenario_prop = 0.1,
                          logit_prop = logit_prob_walking,
                          ratio_mode = ratio_walking_nonwalking)$minimum
logit_shift_s20 <- - optimize(f = shift_search, interval = c(-4,4),
                          scenario_prop = 0.2,
                          logit_prop = logit_prob_walking,
                          ratio_mode = ratio_walking_nonwalking)$minimum
logit_shift_s30 <- - optimize(f = shift_search, interval = c(-4,4),
                          scenario_prop = 0.3,
                          logit_prop = logit_prob_walking,
                          ratio_mode = ratio_walking_nonwalking)$minimum


# 5. using the shifts to make 100 selection variables for each scenario

# 5.1 The sample_01 function takes randomly 1 or 0, based on the probabilities
#     for 0 and 1
```

```
sample_01 <- function(prob_change){
  sample(c(0,1), size = 1, prob = prob_change_01)
}

# 5.2 The sample_weighted function iterates of a number of simulations
#     (here n = 100)
#     The return is n selection variables with 0/1 values
#     The selection is based on the probability variable and the
#     logit shift calculated above

sample_weighted <- function(iteration, prob_change_var, logit_shift){
  logit_prob <- logit(prob_change_var, adjust = .000025)
  logit_shifted <- logit_prob + shift
  prob_shifted <- zapsmall(inv.logit(logit_shifted, adjust = 0.000025))
  prob_shifted <- replace(prob_shifted, which(prob.w1 < 0), 0)
  prob_shifted <- replace(prob_shifted, which(prob.w1 > 1), 1)
  prob_0_shifted <- 1 - prob_shifted
  df <- data.frame(cbind(prob_0_shifted, prob_shifted))
  apply(df, 1, sample_01)
}

# 5.3 The sel_write function applies the sample_weighted function on the dataset
#     and then writes the result to a .Rdata file

sel_write <- function(logit_shift, sim_label, df_original, weight_var, nsim){
  sel_vars <- paste0('V', 1:nsim)
  df <- copy(df_original)

  df[, sel_vars := lapply(1:nsim, sample_weighted,
                          prob_change_var = weight_var,
                          shift = logit_shift), with = FALSE]
  df[mode_trans1 == '1. walking', var.out := 0, with = FALSE]

  path <- "~/.../selection variables/"
  filename <- paste0(path, "selection walking ", sim, ".RData")
  save(list = c('sel'), file = filename)
}

sel_write(logit_shift = logit_shift_s10, sim_label = 'sim10',
          df_original = egt.wal, weight_var = egt.wal[['walking']],
          nsim = 100)
sel_write(logit_shift = logit_shift_s10, sim_label = 'sim20',
          df_original = egt.wal, weight_var = egt.wal[['walking']],
          nsim = 100)
sel_write(logit_shift = logit_shift_s10, sim_label = 'sim30',
          df_original = egt.wal, weight_var = egt.wal[['walking']],
          nsim = 100)


#######################
```

```
# D. Simulation step 2: actual simulations, i.e. the values of
#    transportation mode, duration and T-MVPA (in this order) are changed.
#    For the transportation mode, we used the selection variables created
#    above. For duration of the trip and T-MVPA for the trip, we used the
#    random forest models grown in the step A. The simulation procedure
#    mimics the result on T-MVPA after a predefined change in transportation
#    modes

# 1. The simulated_promo function changes the observed transporation mode
#    into the promoted mode, than changes duration, speed and T-MVPA

simulated_promo <- function(set, promoted_mode, nsim){
  ########################
  set[, pred_mvpa := predict(fit_mvp, set)]
  set[, mode_trans_orig := mode_trans1]
  set[, duration_mn_orig := duration_mn]
  set[, speed_ld_orig := speed_ld]

  ############################
  var <- c(paste0('V', 1:nsim))
  pre.var <- c(paste0('pre', 1:nsim))
  mod.var <- c(paste('mode', 1:nsim))

  # set up simulation function
  for(i in 1:nsim){
    set[, pre.var[i] := pred_mvpa]

    setkeyv(set, var[i])
    set[J(1), duration_mn := NA]
    set[J(1), speed_ld := NA]
    set[J(1), mode_trans1 := promoted_mode]
    set[    , mod.var[i] := mode_trans1]

    set[J(1), duration_mn := as.integer(predict(fit_dur, set[J(1)]))]
    set[J(1), speed_ld := dist_ld/(duration_mn/60)]
    set[J(1), pre.var[i] := predict(fit_mvp, set[J(1)])]

    set[J(1), duration_mn := duration_mn_orig]
    set[J(1), speed_ld := speed_ld_orig]
    set[J(1), mode_trans1 := mode_trans_orig]
    if((i %% 10) == 0) print(i)
  }

  set
}

# 2. The following for loop reads in the previously made datasets with the
#    selection variables, then applies the simulated_promo function
#    and finally writes the output in new .Rdata files
```

```
for(sim in c(10,20,30)){
  path <- "~/.../selection variables/"
  load(file = paste0(pathin, "selection walking sim", sim, ".RData"))

  sel <- simulated_promo(set = sel, promoted_mode = '1. walking', nsim = 100)

  pathout <- "~/.../simulations/"
  save(list = c('sel'), file = paste0(pathout, "simulations walking ",
                                      sim, ".RData"))
}


######################
# E Aggregate at day level
#   T-mvpa was measured at trip-level, but reported at day-level, i.e. we
#   reported the accumulated T-mvpa during the day.

# 1. The aggr_day function aggregates the data set to the day-level
#    by calculating the mean T-MVPA per day and per transportation mode per day
#    and this for each of the 100 simulation within a scenario

aggr_day <- function(set_orig, nsim, setnt){

  pre_var <- c(paste0('pre', 1:nsim))
  day_var <- c(paste0('mvpa_day', 1:nsim))

  # Total MVPA per day, and total number of trips per day
  setkey(set, person)
  set[, nb_ind := 1]

  var1 <- c(pre_var, 'nb_ind')
  var2 <- c(day_var, 'nb_trips')
  set[, var2 := lapply(.SD, sum, na.rm = TRUE), by = person, .SDcols = var1, with =
FALSE]


  # The following for-loop within the function works
  # but it could be coded more efficiently as 1 or several functions.
  # It was first developed outside this function.
  # It calculates the mvpa per type of transport mode, per person per simulation

  mod_var <- c(paste0('mode', 1:nsim))
  for(i in 1:nsim){
    set[, 'mode_sim' := set[,mod_var[i], with = FALSE]]
    set[, 'pred_sim' := set[,pre_var[i], with = FALSE]]
    setkey(set, person, mode_sim)

    # summary datasets
    mvpa_by_mt <- set[, sum(pred_sim), by = list(person,mode_sim)]
    setkey(mvpa_by_mt, person, mode_sim)
```

```
   out <- mvpa_by_mt[CJ(unique(person), unique(mode_sim))][, as.list(V1), by =
person]

   setnames(out, paste0('V', 1:4),
         paste0("MVPA_", c("walking", "biking", "PM", "PT"), "_sim", i))

   set <- merge(set, out, by = 'person')

   # number of trips

   setkey(set, person, mode_sim)
   nb_by_mt <- set[, sum(nb.ind), by = list(person, mode_sim)]

   setkey(nb_by_mt, person, mode_sim)
   out <- nb_by_mt[CJ(unique(person), unique(mode_sim))][, as.list(V1), by =
person]

   setnames(out, paste0('V', 1:4),
         paste0("nb_", c("walking", "biking", "PM", "PT"), "_sim", i))

   set <- merge(set, out, by = 'person')

   # accumulated distance of trips changed

   setkey(set, person, mode_sim)
   dist_by_mt <- set[, sum(dist_ld), by = list(person,mode_sim)]

   setkey(dist_by_mt, person, mode_sim)
   out <- dist_by_mt[CJ(unique(person), unique(mode_sim))][, as.list(V1), by =
person]

   setnames(out, paste0('V', 1:4),
         paste0("dist_", c("walking", "biking", "PM", "PT"), "_sim", i))

   set <- merge(set, out, by = 'person')
 }

 # We retain one observation per person
 set <- unique(setkey(set, person), by = 'person')

 # Merge set and set egtnt (set of people with no trips)
 varnt <- names(egtnt)[which(names(setnt) %in% names(set))]
 setnt2 <- setnt[, varnt, with = FALSE]
 set <- rbindlist(list(set, setnt2), use.names = TRUE, fill = TRUE)

 # All NA's are actually 0's
 replacena <- function(var){var <- replace(var, is.na(var), 0)}
 a <- rep(1:nsim,4) ; b <- a[order(a)]

 var <- c(day_var, 'nb_trips',
```

```
        paste0(rep(paste0("MVPA_", c("walking", "biking", "PM", "PT"), "_sim"), nsim),
b),
        paste0(rep(paste0("nb_",   c("walking", "biking", "PM", "PT"), "_sim"), nsim),
b),
        paste0(rep(paste0("dist_", c("walking", "biking", "PM", "PT"), "_sim"), nsim),
b))
  set[,var := lapply(.SD, replacena),.SDcols = var, with = FALSE]

  # Delete some variables that are not needed in the result
  var <- c(paste0('mode', 1:nsim),
        paste0('V', 1:nsim),
        paste0('pre', 1:nsim),
        c("mode_trans1", "dest_arr", "pdens_arr", "nivetude_sim", "intersec_arr",
"arr_cour",
          "dist_metro_arr", "dist_train_arr", "dist_tram_arr", "dist_bus_arr",
          "dist_pt_arr", "dest_dep", "pdens_dep", "intersec_dep", "dep_cour",
"dist_metro_dep",
          "dist_train_dep", "dist_tram_dep", "dist_bus_dep", "dist_pt_dep",
          "dist_ld", "duration_mn", "speed_ld", "park_arr", "park_res", "park_dep",
          "educ_arr", "educ_dep", "educ_res", "max.alt.prob", "nb.alt", "alt1",
          "alt2", "alt3", "alt.fin", "mode_sim", "pred_sim"))

  set[, (var) := NULL]
  set
}

# 2. This code reads in a dataset with people include in the EGT dataset
#    that did not perform any trip. Therefore, they were not included
#    in the EGT dataset at trip-level. But they need to be in the
#    aggregated dataset

path <- "~/.../aggregate to day/"
load(file = paste0(path, "people without trips.RData")) # contains dataset egtnt

# 3. This for-loop reads in the simulation results, aggregates the datasets and
#    writes the new dataset in new .Rdata files

for(sim in seq(10, 20, 30)){
  # selection file for transportation mode
  pathin = "~/.../simulations/"
  load(paste0(pathin, "simulations walking", sim, ".RData"))

  sel <- aggr.day(set_orig = sel, nsim = 100, setnt = egtnt)

  pathout <- "~/.../aggregate to day/"
  save(list = c('sel'),
       file = paste0(pathout, "aggregated walking promo", sim, ".RData"))
}
```

```
#######################
# F Multiple imputation of EGT day-level dataset
#   Some variables had missing values in the EGT-dataset
#   Therefore, we created a multiple imputation dataset with 5 datasets.
#   This dataset is created for the original EGT dataset, and then linked
#   to the simulated datasets. So, all simulated datasets and the original
#   dataset the imputations of the independent variables are equal.
#   Note: In this particular study, only the imputations for educational level
#   are important since the other variables are not used in the reported results.

# 1. Ordering the variables on the amount of missing values
#   while making sure id variables won't be used in the imputation

seq <- dimnames(md.pattern(egt[,6:ncol(egt), with = FALSE]))[[2]]
seq <- seq[-length(seq)]
seq <- c(c('person', 'resc', 'depcom_res', 'dciris_res', 'over'), seq)
egt <- egt[, seq, with = FALSE]

# 2. Use mice() with the maximum number of iterations maxit set to zero.
#   This is a fast way to create the mids object called ini
#   containing the default settings.

egt.ini <- copy(egt)

egt.ini[,':=' (person = '1', resc = '1',  depcom_res = 1,
               dciris_res = 1, over = 1)]
ini <- mice(egt.ini, max = 0, meth = 'rf')

meth <- ini$meth
meth[c("person", "resc", "depcom_res", "dciris_res",
       "mvpa_day_int", "age_10", "homme",  "res_cour",  "pos_motorized",
       "weekday",  "min_day",  "MIN_walking",
       "MIN_biking", "MIN_PM", "MIN_PT", "nb_trips",
       "nb_walking", "nb_biking",  "nb_PM",  "nb_PT", "over",
       'intersec_res_1000',  'pdens_res')] <- ""
pred <- ini$pred
vis <- ini$vis

# 3. Actual imputation of EGT dataset
#   Method is Random Forest, 5 imputations, 100 trees per imputation

mi_rf <- mice(egt, m = 5, pred = pred, meth = meth, ntree = 100, vis = vis)

#######################
# G Linking the simulation datasets to the multiple imputation dataset

# 1. This step defines the names of the variables to be calculated and
#    added to the multiple imputation dataset

mode <-  c("car", "walking", "biking", "public")
```

```
var_sim <- paste0('mvpa_day',1:nsim)

var_mvpa_wal <- paste0('MVPA_walking_sim',1:nsim)
var_mvpa_bik <- paste0('MVPA_biking_sim',1:nsim)
var_mvpa_car <- paste0('MVPA_PM_sim',1:nsim)
var_mvpa_pub <- paste0('MVPA_PT_sim',1:nsim)

var_nb_wal <- paste0('nb_walking_sim',1:nsim)
var_nb_bik <- paste0('nb_biking_sim',1:nsim)
var_nb_car <- paste0('nb_PM_sim',1:nsim)
var_nb_pub <- paste0('nb_PT_sim',1:nsim)

var_dis_wal <- paste0('dist_walking_sim',1:nsim)
var_dis_bik <- paste0('dist_biking_sim',1:nsim)
var_dis_car <- paste0('dist_PM_sim',1:nsim)
var_dis_pub <- paste0('dist_PT_sim',1:nsim)

# 2. This for loop reads in all datasets at day-level
#    (Note: day-level = person-level, since there is one day per perons)
#    Then, the mean MVPA total and per day are calculated for the
#    simulation dataset and added to the multiple imputation dataset.


for(i in 1:4){
  if(mode[i] == 'car'){simulations = c('05', '10', '15', '20', '25', '30')}
  if(mode[i] %in% c('walking', 'public')){simulations = seq(10,60,10)}
  if(mode[i] == 'biking'){simulations = c('050', '100', '150',
                                          '200', '250', '300')}

  for(sim in simulations){
    # selection file for transportation mode
    pathin = "~/.../aggregate to day/"
    load(paste0(pathin, i, "aggregated walking promo ", sim, ".RData"))

    sel <- sel[order(as.character(sel$person)),]
    mi_rf$data[,paste0('mvpa_day_', mode[i],'_', sim)] <-
      sel[, rowMeans(.SD), .SDcols = var_sim]

    mi_rf$data[,paste0('mvpa_wal_day_', mode[i],'_', sim)] <-
      sel[, rowMeans(.SD), .SDcols = var_mvpa_wal]
    mi_rf$data[,paste0('mvpa_bik_day_', mode[i],'_', sim)] <-
      sel[, rowMeans(.SD), .SDcols = var_mvpa_bik]
    mi_rf$data[,paste0('mvpa_car_day_', mode[i],'_', sim)] <-
      sel[, rowMeans(.SD), .SDcols = var_mvpa_car]
    mi_rf$data[,paste0('mvpa_pub_day_', mode[i],'_', sim)] <-
      sel[, rowMeans(.SD), .SDcols = var_mvpa_pub]

    print(paste(mode[i], sim))
  }
}
```

```
################################################
# H. Results
#    To calculate the results, the means had to be calculated per
#    multiple imputed dataset and then pooled. Function mean_pool
#    calculated the overall means, mean_pool_g calculated
#    the means per category of a variable, here educational level.

mean.pool <- function(mi_data, var, id, gvar = NULL, group = NULL,
                      change = NULL){
  m <- mi_data$m
  Q <- U <- rep(NA, m)
  for (i in 1:m) {
    set <- complete(mi_data, i)
    if(!is.null(change)){
      set <- set[which(set[,change] == 1),]
    }
    if(is.null(group)){
      var1 <- set[,var]
    } else {
      var1 <- set[which(set[,gvar] == group),var]
    }
    Q[i] <- mean(var1)
    U[i] <- var(var1) / nrow(complete(mi_data, i))
    }
  me <- round(pool.scalar(Q, U, n = nrow(set), k = 1)$qbar,2)
  se <- round(pool.scalar(Q, U, n = nrow(set), k = 1)$ubar,2)
  data.table(sim = id, TMVPA = me)
}

mean.pool.g <- function(mi_data, var, gvar, id, change = NULL){
  m <- mi_data$m
  set <- complete(mi_data)
  lev <- levels(set[,gvar])
  lev2 <- gsub(" ", "_", lev)
  lev2 <- gsub("-", "", lev2)
  lev2 <- gsub("__", "_", lev2)
  lev2 <- gsub("[.]", "", lev2)
  lev2 <- paste0('Educ', 1:3)
  res <- data.table(level = lev, mean = rep(as.numeric(NA), length(lev)))
  for(i in 1:length(lev)){
    pres <- mean.pool(mi_data, var, id, gvar, lev[i], change)
    res[level == lev[i], mean := pres[1,TMVPA]]
  }
  res2 <- data.table(t(res[,mean]))
  setnames(res2, paste0('V', 1:3), lev2)
  #res2[, sim := id]
  res2
}
```

```
mean.pool(mi_rf, 'mvpa_day', 'original')
mean.pool.g(mi_rf, 'mvpa_day', 'nivetude_sim', 'original')



#############
# References to R-packages used
#   A. Liaw and M. Wiener (2002). Classification and Regression by
#   randomForest. R News 2(3), 18--22.
#
#   M Dowle, A Srinivasan, T Short, S Lianoglou with contributions
#   from R Saporta and E Antonyan (2015). data.table: Extension of
#   Data.frame. R package version 1.9.6.
#   https://CRAN.R-project.org/package=data.table

#   Stef van Buuren, Karin Groothuis-Oudshoorn (2011). mice: Multivariate
#   Imputation by Chained Equations in R. Journal of Statistical Software,
#   45(3), 1-67. http://www.jstatsoft.org/v45/i03/
```