PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Identifying and sharing data for secondary data analysis of physical activity, sedentary behaviour and their determinants across the life course in Europe: general principles and an example from DEDIPAC
AUTHORS	Lakerveld, Jeroen; Loyen, A; Ling, Fiona; DeCraemer, Marieke; van der Ploeg, Hidde; O'Gorman, Donal; Carlin, Angela; Caprinica, Laura; Kalter, Joeri; Oppert, Jean-Michel; Chastin, Sebastian; Cardon, Greet; Brug, Johannes; MacDonncha, Ciaran

VERSION 1 – REVIEW

REVIEWER	Calum Mattocks University of Cambridge, UK.
REVIEW RETURNED	02-Jun-2017

GENERAL COMMENTS	This is a useful and interesting paper and serves as both a methods paper for the DEDIPAC study and something of a “how to” guide for others interested in similar data pooling projects. The concept that the manuscript describes is both laudable and timely. I do have a few comments and issues, however. The idea of a Europe-wide pooled dataset is a good one however could the authors comment on potential for generalisability beyond Europe? Clearly the issue of PA and sedentary behaviour is a worldwide problem and while some determinants may impact differently beyond (and within) Europe, it would be good to see the authors discuss the broader implications of their research. At least one of the datasets included in this study, according to the DEDIPAC website, is itself pooled from a number of studies (ICAD http://www.mrc-epid.cam.ac.uk/research/studies/icad/). Did the authors consult the publications or the authors themselves from ICAD (or similar studies) to make use of the learning from similar projects? For example, some of the challenges the authors faced could have been foreseen, although I accept that this may have been done but not reported here. Did the authors have access to the raw or processed data and, if raw, what standards were used to process it? This could be important where, for example, different accelerometry cut-points were used to estimate minutes of moderate to vigorous PA or time spent sedentary. Strengths and Limitations – the strengths part reads more like a list of what the authors did. It could be rewritten to make punchier. For example, a stronger case could be made to support the use of FAIR principles and this goes for the main text as well.

	The authors state what these principles are but don't really state strongly enough why they are a good thing. Minor issues Page 2, line 44 typo - "Strengths" Page 4, line 46 should read "Participants had to be aged 6 years or older" Page 5, line 47 should read "Implicit in these requirements are pathways…"

REVIEWER	Adilson Marques Faculdade de Motricidade Humana, Universidade de Lisboa, Portugal
REVIEW RETURNED	15-Jul-2017

GENERAL COMMENTS	The manuscript aimed to describe the inventory and development of a comprehensive European dataset compendium, and the process towards cross European secondary data analyses of pooled data on physical activity, sedentary behaviour and their correlates across the life course. It is an original article, that address a very interesting and very important topic, because the utilization of the available data for secondary analysis increase the potential utility of the data. I like to read the manuscript, but I have some recommendations that are expressed in my comments. Page 2, line 23. what does it mean DEDIPAC. The meaning of the abbreviation should be explained before the use of the abbreviation. Page 3, line 5. The reference number 1 should be deleted. There is no need to use it, because the others are enough to support the idea presented in the sentence. Page 3, lines 9-10. Provide a reference to support the sentence: "Physical activity and sedentary behaviours are influenced by a wide range of individual-level and contextual factors". e.g. Bauman AE, Reis RS, Sallis JF, Wells JC, Loos RJ, Martin BW. Correlates of physical activity: why are some people physically active and others not? Lancet. 2012;380(9838):258-71. Marques A, Martins J, Peralta M, Catunda R, Nunes LS. European adults' physical activity socio-demographic correlates: a cross-sectional study from the European Social Survey. PeerJ. 2016;4:e2066. Page 3, lines 43-47. I think this sentence can be deleted. Although the information is true, it does have an added value for the argumentation. Page 4, lines 25-31. There are some topics they need clarification. What does it mean "easy to find". On the other hand, if data are available and interoperable, naturally the data are reusable. So, reusable should not be a condition, because it is what will happen if all other premises are true.

REVIEWER	Tim Olds University of South Australia Australia
REVIEW RETURNED	17-Jul-2017

GENERAL COMMENTS	This is an interesting and useful paper which addresses the "afterlife" of data — the issue of his best to cumulate and re-use the masses of data from various studies. It is a topic too rarely addressed. Grants too often cover the conduct of studies, but not the cleaning, storage and accessibility of data afterwards. Major comments (1) The authors point out the difficulty of retrospective data harmonisation. I wonder if they might like to add something about prospective harmonisation, i.e. recommending particular instruments, protocols and operationalisations. In this connection, they might like to comment on the relative importance of reliability and validity, how well the instrument/item captures the construct, and the critical mass of existing data using the same instrument/item. (2) I'm not sure that expert consensus is the best way of deciding whether similar questions or methods are sufficiently close to be considered "the same". Sometimes a study is required using competing forms of the same instrument. For example, did the authors class together questions which were otherwise identical except that one asked about "typical" days, and another about "average" days? What about typical/average vs yesterday? (3) Sometimes summary data, and not raw data, are available from datasets (e.g. the % of children who are overweight or obese). These too can be harmonised and cumulated for historical trend and geographical distribution analysis. Would there authors like to comment on this form of data? (4) One of the steps was proposing research questions. I'm not sure this isn't quite a separate issue. Surely researchers should devise their own research questions, and making the data available FAIRly is the important thing. However, this may indeed also suggest that sometimes data harmonisation should be up to the researchers using the data. For example, if I am interested in the relationship between educational level and physical activity, there are many ways I might choose to harmonise educational data (e.g. as tertiles across countries with different educational exposures; or in terms of the same absolute exposure, such as university education). A "pre-digested" variable might not be of much use. (5) Some of the benefits of data harmonisation which the authors don't mention are the development of new statistical techniques, such as compositional data analysis, which may not have been available at the time of data collection; and historical trend analysis. (6) For objective data, how was the use of different protocols, analytical paradigms, algorithms and measurement protocols handled? Were any of there objective data considered commensurable?

	(7) I wonder whether data legacy might not be encouraged as a requirement of granting agencies in future? Minor points (1) The word "data" is plural and requires the plural form of the verb. Please check this throughout. (2) The authors seem to have abandoned the hyphen: "cross-European", "meta-analyses" (3) Methods, line1: "complementary" not "complimentary" (4) I was unclear whether intervention trials were included or excluded. On p4 para5 the authors mention only cross-sctional and longitudinal studies, but Table 1 suggests interventions are also included. (5) No literature systematic review was undertaken. Or was there a systematic review of reviews? Might this not have uncovered more datasets? (6) p5 para3 line4: "Implicit in these requirements IS a pathway ..." (7) p6 para2:"The reference dataset could contain multiple versions that describe a single variable". I'm not sure what this means. Are you trying to say "multiple variables for a single construct"?

VERSION 1 – AUTHOR RESPONSE

Reviewer 1

Comment: This is a useful and interesting paper and serves as both a methods paper for the DEDIPAC study and something of a “how to” guide for others interested in similar data pooling projects. The concept that the manuscript describes is both laudable and timely. I do have a few comments and issues, however.

Reply: We thank the reviewer for these kind words.

Comment: The idea of a Europe-wide pooled dataset is a good one however could the authors comment on potential for generalisability beyond Europe? Clearly the issue of PA and sedentary behaviour is a worldwide problem and while some determinants may impact differently beyond (and within) Europe, it would be good to see the authors discuss the broader implications of their research.

Reply: We would like to clarify that the aim was not to develop one EU-wide pooled dataset – but indeed we limited the compendium of datasets -consisting of metadata- to European studies only. We agree that adding some discussion on the potential generalizability would be useful. We have now done so as follows (page 9, first paragraph):

“Whereas low levels of physical activity and too much sitting are causing health problems worldwide, the focus of this paper is on the European setting. Many of the issues encountered regarding data pooling are expected to be similar for other regions, but the potential for data pooling and secondary data analyses may also partly be different for studies conducted in and by researchers from other regions, for example because of differences in rules and regulations regarding data sharing.

Furthermore, the data pooling was conducted to further research into potential determinants of sedentary behaviour and physical activity, and such determinants themselves may be (partly) different between regions of the world, e.g. because of differences in affluence, infrastructure and cultural differences."

Comment: At least one of the datasets included in this study, according to the DEDIPAC website, is itself pooled from a number of studies (ICAD http://www.mrc-epid.cam.ac.uk/research/studies/icad/). Did the authors consult the publications or the authors themselves from ICAD (or similar studies) to make use of the learning from similar projects? For example, some of the challenges the authors faced could have been foreseen, although I accept that this may have been done but not reported here.

Reply: We have indeed made use of the learning of past data pooling initiatives as much as possible. Mostly through publications of such initiatives, but also based on experiences verbally shared within the DEDIPAC consortium members by DEDIPAC members who were or are involved in pooling or harmonising activities. Initiatives that provided such input include POLARIS but also ICAD. In fact, a stated action of the DEDIPAC methodology within the steps of the data harmonisation proposals was "To learn from the experience of existing data pooling/harmonisation initiatives (e.g. ICAD, POLARIS)." And we have been in contact with ICAD throughout DEDIPAC. We have now specifically referred to ICAD where we note examples of existing pooling initiatives and platforms.

Comment: Did the authors have access to the raw or processed data and, if raw, what standards were used to process it? This could be important where, for example, different accelerometry cut-points were used to estimate minutes of moderate to vigorous PA or time spent sedentary.

Reply: We recognize the importance of choosing consistent cut-points in behavioural outcomes derived from objective measures when harmonising these data. The various exemplar projects that engaged in the actual pooling and harmonising used raw data as much as possible – although only a small minority of variables under study were accelerometer- based. Because of the latter we did not specifically go into detail on the standards used, but addressed it in the paragraph where we discuss the harmonisation process. That paragraph now reads as follows (page 6, last paragraph):

"If the studies measured and reported the same construct in the same way (e.g., self-reported total physical activity based on the IPAQ-short and reported in minutes per day), these variables were linked to the same variable in the reference dataset. However, if there was a difference in terms of concepts (e.g. total physical activity vs. physical activity in leisure time), measurement (e.g., self-reported vs. objective measurements), or reporting (e.g., minutes per day vs. meeting recommendations) these variables were linked to different variables in the reference dataset. Thus, the reference dataset could contain multiple versions that describe a single variable. In a later step, the multiple versions of a variable could be integrated into a unified measurement scheme. This unified measurement scheme describes a set of variables that have been measured in a similar fashion across a number of datasets, thus having potential for harmonisation. Within DEDIPAC this step involved the identification of specific variables across target datasets which could/would be harmonised and required detailed examination of procedures/methods used by each dataset. Through an iterative review process, by a consensus committee, an agreed approach to the actual harmonisation of the variables was found. The selection process required a balance between uniformity (e.g., exact same question wording and data collection procedures, or using the same data processing standards and cut-points) and the acceptance of a certain level of heterogeneity across datasets (i.e., slightly different wording or procedures)."

Comment: Strengths and Limitations – the strengths part reads more like a list of what the authors did. It could be rewritten to make punchier. For example, a stronger case could be made to support the use of FAIR principles and this goes for the main text as well. The authors state what these principles are but don't really state strongly enough why they are a good thing.

Reply: We redrafted the bullet points as suggested (starting page 2, last paragraph):

• We applied the FAIR principles to provide guidance in the discovery and reuse of data for further investigation
• We have identified more than one hundred potentially relevant European datasets through our outlined search approaches
• It was possible to retrieve the metadata from these datasets on the types of variables, age groups under study, study design, measurement instruments used, time frame, etc.
• Limited potential for reuse has been noted and this highlights the immediate need to manage future data collection within Europe using the FAIR principles
• More consistent data collection methodologies among the scientific community should be promoted as the variation in assessment methods and operationalization of outcome variables across current European studies hampered data harmonisation

In addition we have now built in a stronger case for the use of FAIR principles in our conclusions paragraph (page 11 , first paragraph):

"It is recognized that researchers should be facilitated, funded, requested, or required (e.g. by funding agencies) to share their data and comply to the FAIR principles [16,34]. While recognising the importance of utilising existing data it is equally, if not more important, to highlight the absence of data that would be needed to investigate in detail the determinants of these behaviours. Not complying to the FAIR principles will limit the reusability of relevant data. In turn, The lack of suitable data will severely limit the ability of research to understand and predict these behaviours and inform policy."

Comment:
Minor issues

Page 2, line 44 typo - "Strengths"
Page 4, line 46 should read "Participants had to be aged 6 years or older"
Page 5, line 47 should read "Implicit in these requirements are pathways…"

Reply: We have changed the above-mentioned sentences as per reviewer suggestions, and thank the reviewer for pointing these out.

Reviewer: 2
Comment: The manuscript aimed to describe the inventory and development of a comprehensive European dataset compendium, and the process towards cross European secondary data analyses of pooled data on physical activity, sedentary behaviour and their correlates across the life course. It is an original article, that address a very interesting and very important topic, because the utilization of the available data for secondary analysis increase the potential utility of the data. I like to read the manuscript, but I have some recommendations that are expressed in my comments.

Reply: We thank the reviewer for these positive words.

Comment: Page 2, line 23. what does it mean DEDIPAC. The meaning of the abbreviation should be explained before the use of the abbreviation.

Reply: We apologize for this and have changed this as follows (page 2, Methods section abstract):

“METHODS: A five-step methodology was followed by the European Determinants of Diet and Physical Activity (DEDIPAC) Knowledge Hub, covering the 1) identification of relevant datasets across Europe, 2) development of a dataset compendium including details on the design, study population, measures, and level of accessibility of data from each study, 3) definition of key topics and approaches for secondary analyses, 4) process of gaining access to datasets, and 5) pooling and harmonisation of the data and the development of a data harmonisation platform.”

Comment: Page 3, line 5. The reference number 1 should be deleted. There is no need to use it, because the others are enough to support the idea presented in the sentence.

Reply: We agree and now only refer to the relevant evidence at the end of the sentence.

Comment: Page 3, lines 9-10. Provide a reference to support the sentence: "Physical activity and sedentary behaviours are influenced by a wide range of individual-level and contextual factors".
e.g. Bauman AE, Reis RS, Sallis JF, Wells JC, Loos RJ, Martin BW. Correlates of physical activity: why are some people physically active and others not? Lancet. 2012;380(9838):258-71.
Marques A, Martins J, Peralta M, Catunda R, Nunes LS. European adults' physical activity socio-demographic correlates: a cross-sectional study from the European Social Survey. PeerJ. 2016;4:e2066.

Reply: We have now provided references as suggested, including the two mentioned above.

Comment: Page 3, lines 43-47. I think this sentence can be deleted. Although the information is true, it does have an added value for the argumentation.

Reply: We understand the reviewers point here, but are keen to keep that sentence, as it illustrates the heterogeneity in levels of physical activity and sedentary behaviour and their determinants across Europe.

Comment: Page 4, lines 25-31. There are some topics they need clarification. What does it mean "easy to find".

Reply: We have now clarified and further specified what we mean with ‘easy to find’, as follows (page 4, fourth paragraph):

“The FAIR principles suggest that each data resource, associated metadata and complementary files should be registered or indexed in a searchable resource, so that they can be located (‘Findable);(…)”

Comment: On the other hand, if data are available and interoperable, naturally the data are reusable. So, reusable should not be a condition, because it is what will happen if all other premises are true.

Reply: Reusability indeed naturally infers that the data should be available. We actually indicated this in the method section: (…) “Reusable’, i.e., made available.” However, we would like to argue that reusability is a principle distinct from interoperability – which refers to the meta-data (and not to the data itself).

Reviewer: 3
Comment: This is an interesting and useful paper which addresses the "afterlife" of data — the issue of his best to cumulate and re-use the masses of data from various studies. It is a topic too rarely addressed. Grants too often cover the conduct of studies, but not the cleaning, storage and accessibility of data afterwards.

Reply: We very much appreciate these kind words.

Major comments
Comment 1: The authors point out the difficulty of retrospective data harmonisation. I wonder if they might like to add something about prospective harmonisation, i.e. recommending particular instruments, protocols and operationalisations. In this connection, they might like to comment on the relative importance of reliability and validity, how well the instrument/item captures the construct, and the critical mass of existing data using the same instrument/item.

Reply: We agree and have now included a paragraph on prospective harmonisation (page 10, last paragraph):

“This may be prevented in prospective harmonisation efforts – where particular instruments, protocols and operationalisations are specified and aligned beforehand. Such prospective harmonisation is especially required in surveillance systems to monitor regional variations and temporal trends of health behaviours and health outcomes, although this is not often done.[35, 36] In fact an important goal of DEDIPAC was to work towards harmonization of measurement instruments to better enable and promote prospective data harmonization.[19]”

Comment 2: I'm not sure that expert consensus is the best way of deciding whether similar questions or methods are sufficiently close to be considered "the same". Sometimes a study is required using competing forms of the same instrument. For example, did the authors class together questions which were otherwise identical except that one asked about "typical" days, and another about "average" days? What about typical/average vs yesterday?

Reply: The reviewer raises a valid point here with which we agree. We have involved the original researchers as much as possible in the expert consensus process. Consensus should, where possible, also be informed by additional data analysis of how targeted variables for harmonisation relate to each other. Additional analysis of the nature of the variables may provide a better and more objective quantification of how similar questionnaire items actually are and such procedures have been utilised within DEDIPAC Such an approach should be an integral part of “Expert Consensus” – we have included the following in the revised text (page 6, last paragraph):

“As an aspect of expert consensus and where possible an examination of the frequency and descriptive statistics of similar variables informed the potential for harmonisation. Sharing solutions for data harmonisation across research question groups, regular telephone conferences and shared guidance documents were important aspects of this step.”

Comment 3: Sometimes summary data, and not raw data, are available from datasets (e.g. the % of children who are overweight or obese). These too can be harmonised and cumulated for historical trend and geographical distribution analysis. Would there authors like to comment on this form of data?

Reply: Also here we agree with the reviewer. In the exemplar projects the previously mentioned experts had to handle a range of types of variables, ranging from raw accelerometry data to summary data. We have added this nuance to the result section (page 8, last paragraph):

“Variables to be harmonised ranged from raw accelerometry data to summary data (eg proportion of the population above a defined threshold).”

Comment 4: One of the steps was proposing research questions. I'm not sure this isn't quite a separate issue. Surely researchers should devise their own research questions, and making the data available FAIRly is the important thing. However, this may indeed also suggest that sometimes data harmonisation should be up to the researchers using the data. For example, if I am interested in the relationship between educational level and physical activity, there are many ways I might choose to harmonise educational data (e.g. as tertiles across countries with different educational exposures; or in terms of the same absolute exposure, such as university education). A "pre-digested" variable might not be of much use.

Reply: We like to note that the exemplar research questions were addressed -and harmonisation was done- locally by the research groups that posed the question. As such, the harmonisation decisions were made with the specific research question in mind, by (or in very close contact with) the researcher who posed the research question.

Comment 5: Some of the benefits of data harmonisation which the authors don't mention are the development of new statistical techniques, such as compositional data analysis, which may not have been available at the time of data collection; and historical trend analysis.

Reply: We agree that this may be beneficial and have now added the following to the text (page 3, last paragraph):

“increasing the power allows for mediation and moderation analyses and subsequent stratified (subgroup) analyses. Also the re-analysis of previously collected data using contemporary statistical techniques is possible. Such retrospective pooling often requires a data harmonization process in which similar variables across multiple datasets are made compatible.[23]”

Comment 6: For objective data, how was the use of different protocols, analytical paradigms, algorithms and measurement protocols handled? Were any of there objective data considered commensurable?

Reply: Objective data were handled starting from raw data, if available. If these were not available the researchers had to examine the specific data processing standards and cut-points that were used, and come towards solutions for data harmonisation through the described consensus process.

Comment 7: I wonder whether data legacy might not be encouraged as a requirement of granting agencies in future?

Reply: In Europe this is increasingly added as a requirement of funders – especially from granting agencies that provide ‘public’ money. We have suggested a role for funders in the conclusion paragraph (page 11, first paragraph):

“It is recognized that researchers should be facilitated, funded, requested, or required (e.g. by funding agencies) to share their data and comply to the FAIR principles.”

Minor points
Comment 1: The word "data" is plural and requires the plural form of the verb. Please check this throughout.

Reply: This has been changed where appropriate, and we thank the reviewer for pointing this out

Comment 2: The authors seem to have abandoned the hyphen: "cross-European", "meta-analyses"

Reply: We have checked and have now put the hyphen back again.

Comment 3: Methods, line1: "complementary" not "complimentary"

Reply: This has been corrected.

Comment 4: I was unclear whether intervention trials were included or excluded. On p4 para5 the authors mention only cross-sectional and longitudinal studies, but Table 1 suggests interventions are also included.

Reply: Intervention trials were also included. This has now been clarified (page .., paragraph…):
“A relevant European dataset was defined as a dataset collected during an on-going or recently (≤10 years) completed observational or intervention study focusing on physical activity and/or sedentary behaviour and its potential individual and/or contextual correlates.”

Comment 5: No literature systematic review was undertaken. Or was there a systematic review of reviews? Might this not have uncovered more datasets?

Reply: Systematically reviewing the literature of the life span determinants of physical activity and sedentary behaviour was an aspect of DEDIPAC – this process was also utilised to identify relevant existing datasets. This is implicitly noted in the third aspect of our dataset search strategy. The text has been altered to emphasise this contribution (page 5, first paragraph):

“(…) iii) activities of the DEDIPAC Knowledge Hub and expert consultation.”

Comment 6: p5 para3 line4: "Implicit in these requirements IS a pathway ..."

Reply: This has been corrected

Comment 7: p6 para2:"The reference dataset could contain multiple versions that describe a single variable". I'm not sure what this means. Are you trying to say "multiple variables for a single construct"?

Reply: The reviewer is right and we now kindly use the suggested wording.

VERSION 2 – REVIEW

REVIEWER	Calum Mattocks University of Cambridge, UK
REVIEW RETURNED	01-Sep-2017

GENERAL COMMENTS	Th authors have done a good job of addressing my comments and concerns and I thank them for their efforts.

REVIEWER	Adilson Marques Faculdade de Motricidade Humana, Universidade de Lisboa, Portugal
REVIEW RETURNED	17-Aug-2017

GENERAL COMMENTS	I am pleased with authors revision. Although it is not important, I still recommend deleting the sentence "Although comparable, objectively measured data are currently lacking, there are indications that adults in northern European countries engage in more sitting time than in countries in the south of Europe,[20] and that some southern European countries generally appear to be among the less physically active countries.[22,24]" This information does have an added value for the argumentation.

REVIEWER	Prof Tim Olds University of South Australia, Australia
REVIEW RETURNED	22-Aug-2017

GENERAL COMMENTS	The authors have clearly and honestly addressed my concerns.