Guidelines for Research on Publicly Available Data

Nature & Science of Sleep
EiC: Ahmed BaHammam

Publicly-available data have inherent limitations. Due to the inherent nature of the secondary analysis, the available data are not gathered to answer a specific research question or to test a specific hypothesis. As a result, some significant variables frequently are not available for analysis. Similar to this, not all population groupings or geographic areas of interest may have their data collected.

Another issue is that, in order to preserve respondents' privacy, publicly accessible datasets frequently remove information that could be crucial to the analysis at hand, such as an address, the names of the main sample units, and the race, ethnicity, and particular age of respondents. When the omitted variables are important factors to adjust for in the secondary analysis, this can lead to residual confounding.

The fact that the researchers evaluating the data are frequently different people from those who were part of the data collection process is another significant constraint of the analysis of current data. As a result, they are probably unaware of details or flaws unique to the study that may affect how certain variables in the dataset are interpreted. In addition, users may overlook crucial information if it is not prominently shown in the documents since there is sometimes an overwhelming amount of material (especially for complicated, large-scale surveys carried out by government bodies).

Therefore, we proposed the following guidelines to ensure the quality of the analysis of publicly available data.

A. Guidelines related to the dataset:

Used data should meet the substantive and procedural values identified as relevant to big data in health and research; with their definitions proposed by Xafis and colleagues have proposed an ethical decision-making framework (see Tables 1 and 2) (Ref 1).

B. Guidelines related to the analyses:

- It should be specified if the data analysis follows a “question-driven” approach or a “data-driven” approach.
  If a “question-driven” approach, is there a priori hypothesis or a question?

- State the specific variables to be considered

- What kinds of analysis will be carried out? (In the research question-driven technique, this is decided upon before the researchers examine the dataset's actual data; in the data-driven approach, this is decided upon following their examination of the dataset.)

- List the dataset's strengths and limitations. A thorough description of the population being studied, a sample plan and strategy, a time period for data collection, assessment tools, response rates, and quality control procedures are all required.

- All survey instruments, codebooks, instruction manuals, and any other documentation offered to database users must be obtained by authors and thoroughly studied. These documents should provide enough details to allow researchers to evaluate the data's internal and external validity and to decide whether or not there are enough cases in the dataset to produce accurate estimates about the topic(s) of interest.

- Researchers must create operational definitions for the exposure variable(s), outcome variable(s), covariates, and confounding factors that will be taken into account in the analysis before beginning the investigation.

- Running frequency tables and cross-tabulations of each variable that will be used in the main analysis is the first step in the study. This gives details on the coding style used for each variable and the profile of missing data for each variable.

- Skip patterns: How were missing values for some variables managed?

  For example, a survey module's first set of questions about hypnotic-used-related issues usually asks interviewees if they have ever consumed hypnotics. If the response is no, it is reasonable to infer that the interviewee has no such issues, and no further questions about associated issues are asked.
  These kinds of missing values (which show that a particular condition is not important for the respondent) need to be separated from missing values for which the data is actually missing prior to performing the entire analysis (which indicates that the status of the individual related to the variable is unknown). In order to make an informed decision about coding these variables, researchers should be aware of these skips.

- In order to properly manage missing values and, if necessary, modify the distribution of the variables to match the assumptions of the statistical model to be used in the intended study, the researcher should recode the original variables. All syntax for recording variables (and for the analysis itself) should be defined. The recoded variables should be kept in a new dataset. NEVER change the original dataset in any manner.

- It is crucial to verify the accuracy of the identification variable(s) when utilizing data from longitudinal surveys or combining data from several datasets to ensure that the data from various time periods or datasets are appropriately matched.

- Many population-based studies use multi-stage sampling techniques to expand the sample, especially those aimed at determining the prevalence of relatively uncommon illnesses. To alter the analysis of interest in this situation, the dataset often includes

design variables for each case, such as sample weight, stratum, and primary sampling unit (such as the prevalence of a condition, odds ratios, mean differences, etc.). In order to provide less biased results, researchers should take into account the design variables utilized in the original study and use these variables correctly in their analyses.

Table 1 Substantive values relevant to big data contexts

Substantive value	Definition
Harm minimisation	*Harm minimisation* involves reducing the possibility of real or perceived harms (physical, economic, psychological, emotional, or reputational) to persons.
Integrity	*Integrity* refers to a feature or property of those acting in accordance with personal and/or accepted scientific and professional values and commitments.
Justice	*Justice* consists in treating individuals and groups fairly and with respect. This includes the fair distribution of benefits and burdens of data activities (collection, storage, use, linkage, and sharing) and attention to issues of equity.
Liberty/autonomy	*Liberty* and *autonomy* are very closely related concepts. For the purpose of this document, we define *liberty* as the state of not being coerced by physical, legal, or social pressure into action by some outside influence. *Autonomy* is defined as the capacity of a person or group to be self-determining.
Privacy[1]	For the purposes of this Framework, *privacy* refers to controlling access to information about persons. *Privacy* is valuable because the ability to control access to information about persons promotes certain core interests that we have as individuals and groups. These are wide-ranging but include identity interests and the promotion of human autonomous decision-making, as well as freedom from potential harms such as dis- crimination and stigmatisation that may arise from our data being disclosed. This control may be exercised directly by individuals to whom the data pertains, or by designated persons, such as data custodians whose decisions aim to promote those core individual and group interests.
Proportionality	*Proportionality* is a consideration in decision-making that requires that the means are necessary and appropriate in relation to the end that is being pursued, and being cognisant of the competing interests at hand.
Public benefit	*Public benefit* is the overall good that society as a whole receives from a given project. This includes consideration of effects on wellbeing, distribution, societal cohesion, human rights, and other sources of value to society. It may not be possible to measure these factors by the same standards, so some judgement and critical analysis will be required in determining what is publicly beneficial.
Solidarity	*Solidarity* is the commitment among persons with recognised morally relevant sameness or similarity to sharing costs and benefits for the good of a group, community, nation, or global population.
Stewardship	*Stewardship* reflects a relationship with things, such as data, to promote twin objectives of taking care of the object of attention as well as seeking actively to promote its value and utility. It involves guiding others with prudence and care across one or more endeavours—without which there is risk of impairment or harm—and with a view to collective betterment.

[1] Confidentiality should be considered alongside any privacy consideration, where relevant. The obligation to protect and promote the non-disclosure of information imparted in a relationship of trust lies at the core of the concept of confidentiality

Table 2 Procedural values relevant to big data contexts

Accountability	*Accountability* refers to the ability to scrutinise judgements, decisions and actions, and for decision-makers to be held responsible for their consequences.
Consistency	In the absence of relevant differences between two or more situations, *consistency* requires that the same standards be applied across them. While *consistency* in decision-making may be regarded as valuable in its own right, adherence to a practice of consistency may help actors to secure other values, such as fairness and trustworthiness.
Engagement	*Engagement* is the meaningful involvement of stakeholders in the design and conduct of the data activities. *Engagement* goes beyond the dissemination of information and requires that data activities have been influenced in some way by the views of stakeholders.
Reasonableness	*Reasonableness* means appealing to reasons and values that are widely recognised as relevant and fair.
Reflexivity	*Reflexivity* refers to the process of reflecting on and responding to the limitations and uncertainties embedded in knowledge, information, evidence, and data. This includes being alert to competing and conflicting personal, professional, and organisational interests and to the management of associated biases. Reflexive institutions revise or create new policies and systems that change institutional processes and prompt further reflection and response.
Transparency	*Transparency* is openness to public scrutiny of decision-making, processes, and actions. *Transparency* helps to demonstrate respect for persons and contributes to trustworthiness.
Trustworthiness	*Trustworthiness* is the property of being worthy of trust. It is a value that applies not only to individuals, organisations, governments, and institutions, but also to data, evidence, and systems. It can manifest procedurally as being transparent and truthful, reliable and consistent, or dependable.

References:

1) Asian Bioethics Review (2019) 11:227–254. https://doi.org/10.1007/s41649-019-00099-x
2) Shanghai Arch Psychiatry. 2014; 26(6): 371-375. doi: http://dx.doi.org/10.11919/j.issn.1002-0829.214171