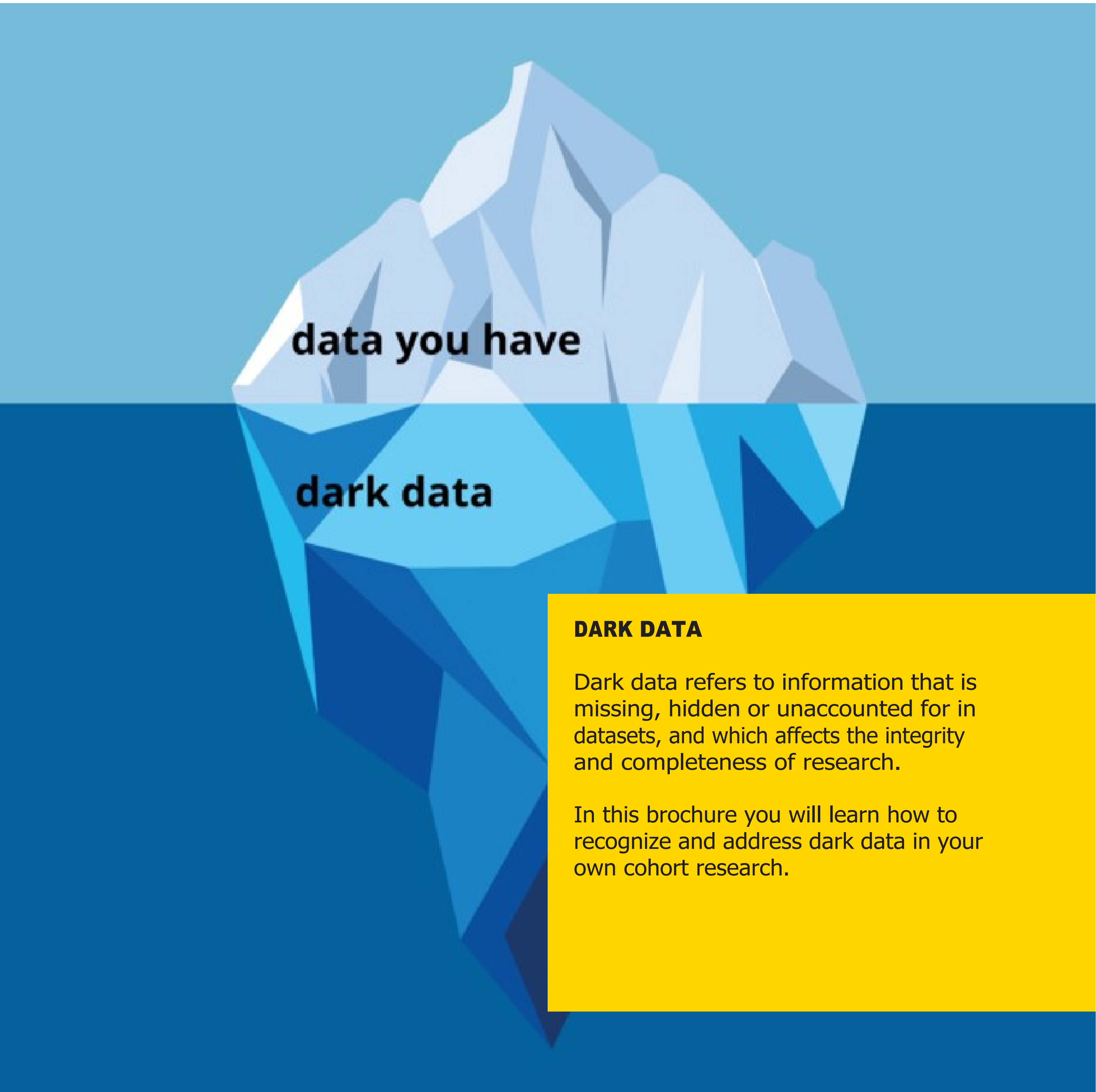
Utrecht University

DYNAMICS OF YOUTH

Uncovering dark data in cohort studies



DARK DATA

Dark data refers to information that is missing, hidden or unaccounted for in datasets, and which affects the integrity and completeness of research.

In this brochure you will learn how to recognize and address dark data in your own cohort research.

DARK DATA IN COHORT STUDIES

What is dark data?

Dark data is information that is missing, hidden or unaccounted for in datasets, which can lead to gaps or inaccuracies in research findings. As David J. Hand writes in his book about this topic (2020), *Dark data is what we don't know that may impact our conclusions, sometimes with catastrophic results.*

Dark data is analogous to dark matter in physics – just as dark matter is an invisible force that affects the universe's structure, dark data is unseen information that can heavily influence research outcomes. Like dark matter, this unobserved data has a 'gravitational pull' on findings. It can lead researchers toward conclusions that might be misleading if the missing factors aren't accounted for.

We – DoY Thriving & Healthy Youth, UMC Child Health, and UU Methods & Statistics - examined dark data patterns in the PROactive cohort study, a longitudinal study of children and adolescents with chronic illnesses at the Wilhelmina Children's Hospital.

In this brochure, we want to share the lessons we've learned on how to recognize and address dark data in your own (cohort) research.

Types of dark data and how they occur

We've learned that dark data arise in different shapes and forms and for different reasons. David Hand describes 15 types of dark data in his book. Some that are especially relevant to cohort studies are:

- **Known missing data (DD-Type 1)**: This category includes data that researchers know are missing, for example, due to equipment failure, data entry errors, or loss-to-follow-up. In longitudinal cohort studies, participants often drop out due to relocation, personal circumstances, or health changes.

- **Unknown missing data (DD-Type 2)**: This refers to data that researchers aren't aware are missing, such as systemic biases that unintentionally exclude participants. For example, one such influence was the COVID-19 pandemic. During the beginning of data collection for the PROactive cohort study, the pandemic unexpectedly impacted participants' health and well-being in ways that were not initially captured.

- **Choosing just some cases (DD-Type 3):** This type of dark data emerges when only a subset of participants is selected. For example, when examining the effects of extracurricular activities (e.g., participating in a sports club) on the well-being of chronically ill children, the sample may unintentionally favour children from higher socioeconomic backgrounds. These families are often better equipped to support their children's participation in such activities, and this can skew the findings toward higher SES families.

- **Self-selection (DD-Type 4)**: Self-selection bias occurs when certain participants choose to drop out or do not respond to surveys. In a cohort of vulnerable children, it's possible that those who experience higher stress or more severe conditions are more likely to drop out. This self-selection can result in a dataset that underrepresents the experiences of the most affected participants. In addition, it might not always be transparent that there is an underlying, systematic reason for non-response, and researchers might mistakenly attribute missing responses to random dropout.

DARK DATA IN COHORT STUDIES

- **Data which might have been (DD-Type 6):** This dark data type includes missing counterfactuals data that could have been recorded if circumstances were different. For example, in studies affected by the pandemic, like the PROactive study, there is no data reflecting the 'normal' trajectory of children's well-being without pandemic-related disruptions. This absence creates a gap in understanding what participants' experiences might have been under typical conditions.

- **Changes over time (DD-Type 7):** Over the duration of a cohort study, the composition of participants can shift. In the PROactive cohort study, for instance, the age and gender distributions in different illness groups evolved over time. Such shifts can influence the representativeness of findings and the comparability of different study phases.

- **Measurement error and uncertainty (DD-Type 10)**: Cohort studies often introduce new questionnaire versions in different waves to meet harmonization standards or align with updated guidelines. In the PROactive cohort study, such changes included modifications to answer formats. Although these adjustments improve cross-study comparability, they can also introduce measurement error and add uncertainty to longitudinal comparisons.

- **Intentionally darkened data (DD-Type 13)**: This type of dark data occurs when certain information is intentionally withheld, often to protect participant privacy or for strategic reasons. For instance, in cohort studies involving sensitive health information, certain participant details might be withheld to protect confidentiality. While this ensures privacy, it can also limit researchers' ability to analyse the full range of factors influencing outcomes.

Implications for cohort studies

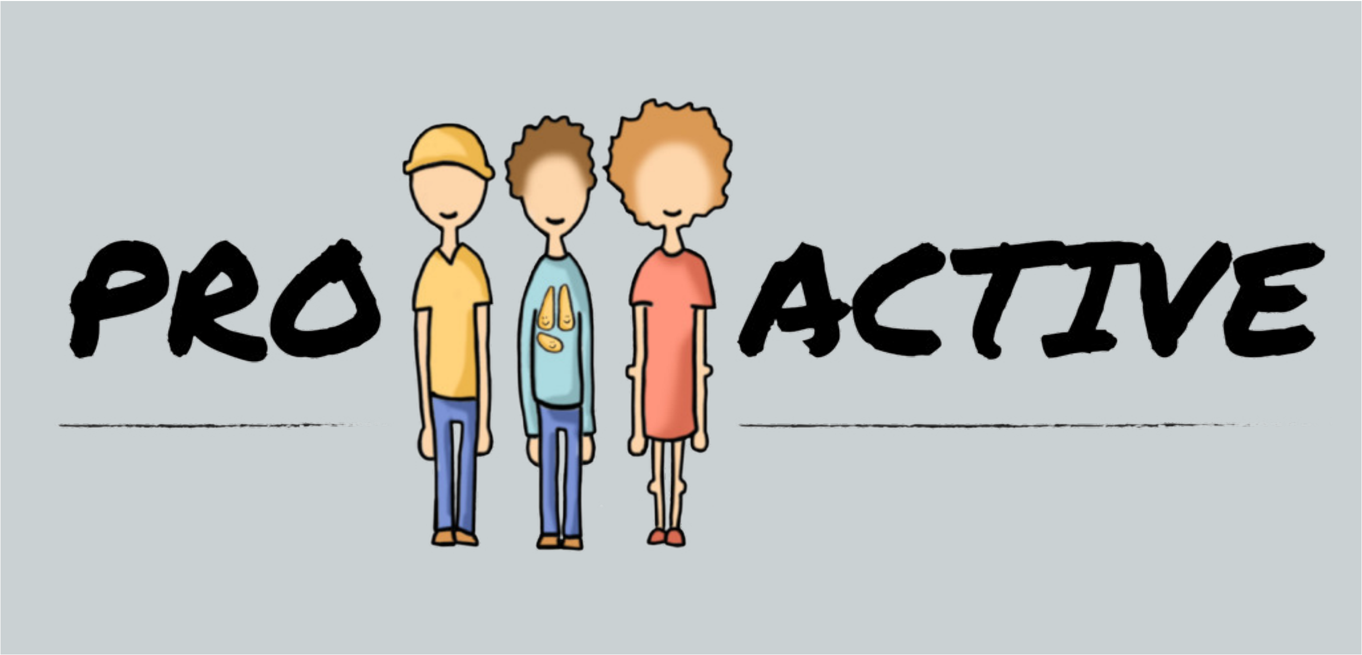
Dark data can lead to:

- **Bias and misleading results**: Known and unknown missing data can bias study findings, leading to incorrect conclusions about interventions or health outcomes.

- **Loss of generalizability**: Selection biases, self-selection, and choosing just some cases can make it difficult to generalize study findings to the broader population.

- **Difficulty in longitudinal comparisons**: Changes over time and inconsistencies in survey tools create challenges in comparing data across different phases of a cohort study.

- **Measurement issues**: Errors in measurement tools or survey changes can add uncertainty to the validity of results.



DARK DATA IN COHORT STUDIES

Strategies for researchers

- **Document and analyse missing data:** Try to identify missing data explicitly and determine if it's 'known' or 'unknown'. Understanding which type of dark data is present helps determine how it might influence results.

- **Avoid excluding incomplete responses:** Whenever possible, use all available data. Excluding participants with incomplete data can lead to biases, particularly if the missing data is non-random (such as when participants from lower SES backgrounds are more likely to have incomplete data).

- **Imputation with caution:** Imputation, or the process of estimating missing values, should be approached with caution. For instance, imputing sensitive variables like SES or health outcomes requires careful consideration of context and potential confounding variables.

- **Track changes**: Documenting changes in measurement tools (e.g., updated questionnaires) and cohort composition (e.g., demographics) can help adjust for differences and ensure continuity.

- **Combine data sources:** Merging data from multiple sources can help fill in gaps. For example, linking survey responses with medical records or administrative data can provide a more complete picture and reduce the effects of missing data from any one source.

Bayesian methods: Bayesian methods allow researchers to include prior knowledge about the data. This probabilistic approach allows for more nuanced predictions and to quantify uncertainty around these estimates. As new data becomes available, Bayesian methods also allow researchers to update their models and adjust initial assumptions.

Colophon

Authors:

- Anne Margit Reitsema, UU DoY Thriving & Healthy Youth: a.m.reitsema@uu.nl
- Anne Hoefnagels, Child Health UMC: j.w.hoefnagels@umcutrecht.nl
- Gerko Vink, UU Methodology & Statistics: g.vink@uu.nl

Literature:

- Nap-van der Vlist, M. M., Hoefnagels, J. W., Dalmeijer, G. W., Moopen, N., van der Ent, C. K., Swart, J. F., ... & Nijhof, S. L. (2022). The PROactive cohort study: rationale, design, and study procedures. *European Journal of Epidemiology*, *37*(9), 993-1002. https://doi.org/10.1007/s10654-022-00889-y

- Hand, D. J. (2020). *Dark data: Why what you don't know matters*. Princeton University Press

MORE INFORMATION

Research theme Dynamics of Youth - Thriving & Healthy Youth
UMC Utrecht WKZ Child Health
https://doy-community.sites.uu.nl/

This project was funded by the DoY-Child Health Booster Grant 2023