**Application form for Fostering Open Science Practice Fund**

**Closing date: 18 January 2023**

The Open Science Fund is an opportunity for **Utrecht University** and **University Medical Centre Utrecht** employees to access small grants with which they can apply Open Science principles into their research. This funding amounts to € 10.000 (minimum) - € 15.000 (maximum) per application.

**Contact and information**

If you are considering an application and you would like to discuss this with a member of the Open Science Programme team, please send a mail to [openscience@uu.nl](mailto:openscience@uu.nl) or contact [Judith de Haan](https://www.uu.nl/medewerkers/JJdeHaan), programme director.

More information, such as selection criteria, who can apply and the selection process, can be found on the [fund website](https://www.uu.nl/onderzoek/open-science/fostering-open-science-fund).

|  |  |  |  |
| --- | --- | --- | --- |
| **Names** | Thom Volker, Erik-Jan van Kesteren and Gerko vink | | |
| **Position/role** | Phd Candidate / Assistant Professor / Associate Professor | | |
| **Department** | Methodology and Statistics | | |
| **Faculty** | Social and Behavioural Sciences | | |
| **Email address** | t.b.volker@uu.nl | | |
| **Telephone number** | 0634866352 | | |
| **Title of proposed project** | Synthemice: An R package for evaluating and improving the quality of synthetic data sets | | |
| **Project start date** | March 2023 | **Project end date** | March 2024 |
| **WBS number** | SA.130402.101 | | |

**THIS PROPOSAL HAS AN OPEN DEVELOPMENT REPOSITORY:** [**https://github.com/gerkovink/OSF2023**](https://github.com/gerkovink/OSF2023)

**Please provide a summary of your project (max. 100 words):**

*(to describe the project on our website)*

|  |
| --- |
| Synthetic data rapidly gains popularity, being used in reviewing, model building pipelines, and as intermediate step for researchers who are applying for access to real, but restricted, research data. Yet, generating synthetic data is a cumbersome process, that heavily relies on ad hoc and unsystematic measures of data utility. We will improve and standardize this process, by developing an easy-to-use software package <synthemice> to systematically assess the quality of synthetic data. The software highlights the aspects of the synthetic data that do not resemble the real data, and provides concrete guidance on how to improve the synthesis models. |

**Please outline the proposed project, including the *purpose* of Open Science Practice, the specific** [***topic***](https://www.uu.nl/en/research/open-science/topics) ***it addresses*, the *approach* being taken and the *links* to research’ (max. 500 words):**

|  |
| --- |
| Open data is paramount to open science practices, as it allows for a thorough evaluation of the analyses and results reported in research papers, including choices made in data manipulation. Yet, increasing concerns for privacy and confidentiality restrict the open dissemination of research data. When the actual data cannot be shared openly, synthetic data can be a safe, privacy-preserving alternative that can be freely disseminated.[1,2] A synthetic data set can be used to verify the original data manipulation and analysis script, because it allows to re-run the original code on a data set that is similar to the original. As such, a synthetic set enables reviewers/readers to run each line of the original code to assess whether the script works as intended. A synthetic data set will never provide results identical to the original, but, if properly generated, will provide results that are close to this. Moreover, a synthetic set can be used to get an indication of how different choices in data manipulation or the analysis would affect the conclusions. Apart from its role in a review/monitoring process, synthetic data can be used for educational purposes, data exploration, model testing, data processing pipelines and exploratory research. These virtues of synthetic data are available if, and only if, the synthetic data is of sufficient quality.  Yet, high quality synthetic data is usually not generated automagically by clicking a button, but requires multiple cycles of refinements regarding the generating models. This process would greatly benefit from good measures of data utility that can help to steer the refinements in the right direction. Yet, evaluation of synthetic data often happens on an ad hoc basis, because good implementations of data utility measures are lacking, while those that exist hardly provide information on which aspects of the original data are inadequately reflected in the synthesis models. To this end, we develop **<synthemice>**, an R-package that allows for a thorough, systematic and informative evaluation of synthetic data. In its essence, <synthemice> takes the original and one or multiple synthetic data set as input, and evaluates to what extent the synthetic data reflects the original data, regardless of which software is used to create the synthetic data. The output not only contains measures of similarity, but also indicates which variables, or relationships between variables, are not adequately reproduced in the synthetic data. To this end, feature importance measures will guide quantitative assessment of the quality of the synthetic data. Moreover, dimensionality reduction techniques will be used to foster a visual inspection of the synthetic data quality, because they allow to map the, potentially many, dimensions of the real and synthetic data on a lower dimensional space that highlights the dimensions on which there are discrepancies between real and synthetic data. Ultimately, a Shiny app will be created to enhance the user-friendliness of <synthemice> and allow for cross-platform usability.  <synthemice> will contribute to open sciences practices through its purpose to improve the quality of synthetic data when collected data cannot be disseminated, protecting the privacy of respondents. Moreover, it’s development will be open and community-driven, and implemented in the open-source software R, allowing contributors to add functionality and improve the methodology of synthetic data evaluation.  Starting with the first of the SMART-objectives, with the remaining objectives left to the next section:  Specific: An R-package with functionality to systematically, but flexibly, evaluate the quality of synthetic data. Functionality will be implemented to (1) differentiate between the observed and synthetic data (support vector machines, random forests); (2) obtain a quantitative assessment of discrepancies (feature importance); (3) allow for a visual assessment of the most severe discrepancies (dimension reduction techniques and corresponding plotting functions). |

**How will you evaluate the progress, outcomes and impact of your project? How will these results be shared? (max. 300 words)**

|  |
| --- |
| We state the remaining SMART-objectives:   * Measurable: Upon finishing, there will be an open-source software package with public development repository, R-package, development instructions for contributors and a tutorial paper/vignette. The project will operate under a GNU GPL-3 license, which prevents closed source distribution. Impact can be measured through CRAN downloads, GitHub forks and stars, development contributions by other scientists and scientific referencing (long-term). * Achievable: We work with manageable deliverables (D) that build up to milestones (M):   + D1.1: Open repository and website aimed at open development   + D1.2: <synthemice> with quantitative evaluation of a single synthetic data set in R   + D1.3: Extend <synthemice> with dimension reduction tools for visual evaluation of one synthetic data set in R   + D1.4: Writing a package vignette   + D1.5: Building a package and submitting to CRAN [M1]   + D2.1: Extend <synthemice> to allow for multiple synthetic data sets (fostering inferences from synthetic data)   + D2.2: Develop a Shiny app to enhance user-friendliness and cross-software usability   + D2.3: Extending the package vignette   + D2.4 Update the CRAN package with developed functionality [M2]   + D3: Extend functionality to allow for more evaluation methods and increase flexibility for the user (based on user requests; long term). * Relevant: Synthetic data is increasingly being used in practice, with researchers sharing it with their paper and analysis code, data collectors sharing it when the collected data cannot be disseminated and national statistical institutes sharing it to researchers who are in the process of obtaining access to real data. Yet, generating and evaluating the quality of synthetic data is cumbersome, and smoothing this process can foster a wide dissemination of the practice of sharing synthetic data. This accelerates novel research by allowing for the dissemination of useful research data without harming respondents’ privacy, or enhances transparency about the data analysis pipeline, improving the reproducibility and evaluation hereof of published research. * Timebound: Start and end dates are clear. Open Science Festival or Open Science Community Utrecht present opportunities to share intermediate results and get feedback (around milestone 1; September 2023). |

**Please describe the potential for learning and/or development for researchers (max. 150 words):**

|  |  |
| --- | --- |
| Synthetic data has many virtues when collected data cannot be openly disseminated. Yet, although its popularity increases, the provision of synthetic data in the absence of the real data is the exception rather than the rule. For many researchers, generating a synthetic data set will be a complex endeavour, especially because current evaluation tools give little guidance on how to improve the synthesis models. The <synthemice> package will facilitate researchers with generating a realistic and highly useful synthetic data set by guiding the synthesis process. As such, researchers can strengthen their open science practices by disseminating a useful surrogate data set to the public, even when access to the real data should be restricted. Sharing such synthetic data openly can contribute to the review process, or fosters the dissemination of expensively collected research data. Our project will highlight the virtues of synthetic data to colleagues, and provides an invitation to collaboration on this topic. |  |

**Please detail the amount of funding applied for and justify the costs requested:**

|  |
| --- |
| **TO DO**  The total budget requested is €15.000, which would facilitate  100h – Thom Volker  100h – Erik-Jan van Kesteren  100h – Gerko Vink  ??? – SA? Dan waarschijnlijk minder geld voor ons.  TV & EJvK: Programming lead and core developer.  GV: Project lead, package and website maintainer.  All three will contribute to documentation, and reach out to potential other developers and users to contribute/provide feedback. |

**Please send the completed application form to** [**openscience@uu.nl**](mailto:openscience@uu.nl) **by 18 January 2023.**

**Literature**

[1] Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. "Dissecting racial bias in an algorithm used to manage the health of populations." *Science* 366, no. 6464 (2019): 447-453.

[2] Zettler, Ingo, Christoph Schild, Lau Lilleholt, Lara Kroencke, Till Utesch, Morten Moshagen, Robert Böhm, Mitja D. Back, and Katharina Geukes. "The role of personality in COVID-19-related perceptions, evaluations, and behaviors: Findings across five samples, nine traits, and 17 criteria." *Social Psychological and Personality Science* 13, no. 1 (2022): 299-310.