**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** | SynthEval: 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/thomvolker/OSF2023**](https://github.com/thomvolker/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, open science 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 <SynthEval> 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, by allowing for a thorough evaluation of 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 collected data cannot be shared openly, synthetic data can be a safe, privacy-preserving alternative that can be freely disseminated.1,2 The synthetic set is useful to verify the original data manipulation and analysis script, enabling reviewers/readers to run the original code and assess whether this works as intended. Although the corresponding results are never identical to the original, they will be close if the synthetic data is generated properly. Additionally, the synthetic set can provide insight into the robustness of results, by allowing others to assess how different choices in data manipulation or analysis affect the conclusions. Apart from its role in review/monitoring processes, synthetic data is useful for open 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 rarely 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 help to steer refinements in the right direction. However, 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.3,4 To this end, we develop **<SynthEval>**, an R-package for a thorough, systematic and informative evaluation of synthetic data. Regardless of the software used to generate the synthetic data, <SynthEval> takes the original and synthetic data as input, and evaluates how well the synthetic data reflects the original. The output contains similarity measures and indicates which variables or relationships between variables are inadequately reproduced in the synthetic data, using feature importance measures to explain discrepancies. Moreover, dimensionality reduction techniques are implemented to visually inspect the synthetic data quality, highlighting the dimensions with important inconsistencies between real and synthetic data. Finally, a Shiny app is created to enhance the user-friendliness and cross-platform usability of <SynthEval>.  <SynthEval> contributes to open science practices and privacy-protection of research respondents by improving the quality of synthetic data, in situations where collected data cannot be released. Moreover, <SynthEval> will be implemented in the open-source software R, with open and community-driven development on GitHub, allowing contributors to improve the functionality and methodology of synthetic data evaluation.  Starting with our first SMART-objective (the remaining objectives follow in the next section):  Specific: An R-package with functionality to systematically, but flexibly, evaluate the quality of synthetic data. Functionality is implemented to (1) discriminate between observed and synthetic data (support vector machines, random forests); (2) quantitatively assess discrepancies (feature importance measures); (3) visually inspect severe discrepancies (through dimension reduction techniques and corresponding visualization routines). |

**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: 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 will be measured through CRAN downloads, GitHub forks and stars, development contributions by other scientists and scientific referencing (long-term).  Achievable: We define manageable deliverables (D) that build up to milestones (M):   * D1.1: Open repository and website aimed at open development * D1.2: <SynthEval> with quantitative evaluation techniques for a single synthetic data set in R * D1.3: Extend <SynthEval> with dimension reduction tools for visual evaluation of one synthetic data set in R * D1.4: Increase accessibility through tutorials, documentation and ‘Getting started’-webpage * D1.5: Build package and submit to CRAN [M1] * D2.1: Extend <SynthEval> to allow for multiple synthetic data sets (fostering inferences from synthetic data) * D2.2: Develop and document a Shiny app to enhance user-friendliness and cross-software usability * D2.3 Update CRAN-package with D2.1 and D2.2 [M2]   Relevant: High-quality synthetic data is a solution to many data accessibility issues, and can therefore be paramount in a transition to a FAIR and open research and publishing climate. Having proper means of evaluating the utility and validity of synthetic data generation procedures is key in the success of this transition. The proposed package is timely: we need proper evaluation means now. The impact of our proposal is likely to accelerate novel data dissemination efforts and developments in the analysis of private data. The proposed research can fortify the position of Utrecht University as a leading academic institution in open science.  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 is a complex endeavour, especially because current evaluation tools give little guidance on how to improve the synthesis models. The <SynthEval> package will facilitate researchers with generating realistic and valid synthetic data 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, and fosters the dissemination of expensively collected research data. Our project highlights the virtues of data synthesis to colleagues, and invites broad collaboration on this topic. |  |

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

|  |
| --- |
| The total budget requested is €15.000, which would facilitate  75h (€2100) – Thom Volker  75h (€3675) – Erik-Jan van Kesteren  75h (€4800) – Gerko Vink  125h (€3500) – Student-Assistant  €925 – Conference travel budget to present the package at useR!/Joint Statistical Meetings  TV and EJvK serve as programming lead and core developers (deliverables D1.2, D1.3, D1.5, D2.1-D.23), GV leads the project and maintains the package and documentation (deliverables D1.1, D1.5, D2.3), and the SA will contribute to improving the package accessibility and writing unit-tests under supervision of the other contributors (D1.4, D2.2). Note that we pursue a team-effort rather than a clear separation of tasks: all contributors collaborate on all deliverables. Hence, all will contribute to documentation, interact with developers and users on open discussion boards and issue fora on GitHub, and accelerate developments in their network. |

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

**Literature**

1. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019 Oct 25;366(6464):447-53.
2. Zettler I, Schild C, Lilleholt L, Kroencke L, Utesch T, Moshagen M, Böhm R, Back MD, Geukes K. 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. 2022 Jan;13(1):299-310.
3. Drechsler J. Challenges in Measuring Utility for Fully Synthetic Data. InInternational Conference on Privacy in Statistical Databases 2022 (pp. 220-233). Springer, Cham.
4. Karr AF, Kohnen CN, Oganian A, Reiter JP, Sanil AP. A framework for evaluating the utility of data altered to protect confidentiality. The American Statistician. 2006 Aug 1;60(3):224-32.