Institution: Brunel University London		
Unit of Assessment: 11 – Computer Science and Informatics		
Title of case study: A Synthetic Data Generator Service for Improving Public Health and Enabling Growth in the Health-Tech Sector		
Period when the underpinning research was undertaken: 2018-2020		
Details of staff conducting the underpinning research from the submitting unit:		
Name(s): Allan Tucker	**Role(s) (e.g. job title):** Reader	**Period(s) employed by submitting HEI:** 2005-present
Period when the claimed impact occurred: 2019-2020		
Is this case study continued from a case study submitted in 2014? N		

1. Summary of the impact (indicative maximum 100 words)

Brunel has worked with the Medicine Health Regulatory Authority (MHRA) to create and successfully introduce a new synthetic data generation service enabling the release of valuable NHS primary care data for the first time. This opens a new “machine learning” sector in the market for developing mobile health apps that is expected to save the NHS GBP2,000,000,000. In addition, it ensures patient privacy concerns with respect to General Data Protection Regulation are addressed. This work has led the MHRA to update its regulation for AI technology and to initiate a new synthetic data generator service, which enables health innovators to develop and validate state-of-the-art health apps using NHS data that would otherwise be unavailable. Two new full-scale synthetic datasets have been released in August 2020 based on cardiovascular disease and Covid-19, funded by NHSx. These will lead to the development of new diagnostic tools by enabling health app-innovation through machine learning. Already, Sensyne Health have been using the service to develop and validate new mobile apps for diabetes, a disease which affects 3,500,000 people in the UK, and which currently costs the NHS GBP9,800,000,000 per annum. Apps such as these will empower patients to take control of their own health through improved monitoring and reporting, and therefore enable the delivery of more personalised clinical decision-making.

2. Underpinning research (indicative maximum 500 words)

Probabilistic models such as Bayesian networks have proved to be extremely valuable in modelling complex data in a transparent way. They can be used to model many types of data including categorical data, numerical data, and temporal data and have been particularly popular in environmental and medical applications **(REF 1, REF 2)**. However, these models are limited by the scalability of truly “huge” datasets as learning a Bayesian network from data is NP-Hard. In a recent collaboration with Oxford University, Tucker developed efficient algorithms for the scalable learning of Bayesian networks from huge datasets **(REF 3)**. These algorithms exploited a new metric for scoring models, taking advantage of the availability of closed-form estimators for local distributions with few parents within a network. Tucker showed that using predictive instead of in-sample goodness-of-fit scores improves speed and accuracy of network reconstruction. These developments were included in the *bnlearn* package in 2019 and are now published.

A collaboration with the Medicine and Health Regulator Authority (MHRA), funded by a pioneer grant from INNOVATE UK, resulted in the extension of the scalable Bayesian network learning algorithm to build high-fidelity synthetic datasets. This approach integrated resampling, probabilistic graphical modelling, latent variable identification, and outlier analysis within a Bayesian network framework to capture structurally missing data when inferring models from millions of primary care patient records **(REF 4)**. Tucker demonstrated that datasets generated using this method include much of the richness and value of the original primary care data by exploring multivariate distributions, correlation structure and sensitivity analyses. What is more, many of the privacy concerns of making real patient data publicly available are mitigated as demonstrated through simulating the difficulty in matching patients to synthetic datapoints. This research has enabled the MHRA to release synthetic data to the health tech sector, facilitating innovation in the sector through the use of the datasets for the development and validation of new mobile health apps. In turn this has led to the MHRA updating its regulation of new software for diagnosing, monitoring disease and identifying risk factors **(REF 5, REF 6)**. Sensyne Health are already using the techniques for developing and validating a health app for modelling gestational diabetes.

3. References to the research (indicative maximum of six references)

REF 1: Uusitalo, L., Tomczak, MT., Müller-Karulis, B., Putnis, I., Trifonova, N., Tucker, A. (2018) 'Hidden variables in a Dynamic Bayesian Network identify ecosystem level change'. *Ecological Informatics*, 45 (May 2018). pp. 9 - 15. ISSN: 1574-9541

REF 2: Ceccon, S., Garway-Heath, D., Crabb, D., Tucker, A. (2014) 'Exploring early glaucoma and the visual field test: Classification and clustering using Bayesian networks'. *IEEE Transactions on Biomedical and Health Informatics*, 18 (3). pp. 1008 - 1014. ISSN: 2168-2194

REF 3: Scutari, M., Vitolo, C.,Tucker, A. “Learning Bayesian Networks from Big Data with Greedy Search”. *Statistics and Computing*. pp. 1 - 15. (2018) ISSN: 0960-3174

REF 4: Tucker, A., Wang, Z., Rotalinti, Y., Myles, P. “Generating High-Fidelity Synthetic Patient Data for Assessing Machine Learning Healthcare Software”, *Nature Digital Medicine*, *npj Digit. Med.* **3,** 147 (2020)

REF 5: Wang, Z., Myles, P., Tucker, A. “Generating and Evaluating Synthetic UK Primary Care Data: Preserving Data Utility & Patient Privacy.” *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)* (2019): 126-131 DOI:10.1109/CBMS.2019.00036

REF 6: Wang, Z, Myles, P, Tucker, A. Generating and evaluating cross-sectional synthetic electronic healthcare data: Preserving data utility and patient privacy. *Computational Intelligence*. 2021; 1– 33. https://doi.org/10.1111/coin.12427

4. Details of the impact (indicative maximum 750 words)

The optimised scalable methods for learning probabilistic models are being used as part of the CRAN *bnlearn* package **(S1, S2)**. The *bnlearn* package is used worldwide and has had 113,028 downloads since the optimised code was made available **(S3)**. In particular, the optimised algorithms in this package are being used by the MHRA for modelling complex disease comorbidities from huge primary care datasets (Wang et al. 2019) to generate synthetic data **(S4, S5)**.

The collaboration with the MHRA has used the scalable Bayesian network package to develop a synthetic data generation framework that enables the generation of high-fidelity synthetic datasets that can be used for benchmarking machine learning algorithms for regulatory purposes (Tucker 2020). The collaboration was highly successful, far exceeding the promised deliverables in the original project scope. Based on this work, the MHRA put in a business case for scaling up production of synthetic data by the MHRA’s specialist division, the Clinical Practice Research Datalink (CPRD), and awarded additional funds from NHSx to build on the previous work **(S6)**.

1) Changing National Policy – As a result of the success of the scalable synthetic data generator, the CPRD has revised its guidance for reviewing machine learning research applications submitted as part of data access requests based on the empirical evidence from this project **(S4, S5)**. In addition, the CPRD is planning to launch a synthetic data generation service, with 12 staff already devoted to working on the project **(S4)**. This new service will involve government analysts adopting innovative methodological approaches based on the work. The MHRA's intention is that these research informed decisions, which include further work with Brunel, will ultimately lead to direct benefits for the health and wellbeing of people by bringing medical device products to markets more quickly. In addition, commerce and the economy will benefit through making the UK an attractive environment for supporting the development of software medical devices. This will be supported by better public policy with improved regulatory pathways that support innovation **(S5)**. This is the first time such data has been made available and the project is currently shortlisted for a Civil Service award. The work has led to the MHRA releasing two synthetic datasets for the first time: a proof-of-concept synthetic cardiovascular risk dataset has been made available for access by the wider research community. In addition, a synthetic dataset has been generated to facilitate Covid-19 research, which was separately funded by NHSx and can also be accessed via CPRD **(S7)**. The work has been the subject of a press release issued jointly by the MHRA, the Department for Health and Social Care (DHSC) and the Department for Business, Energy and Industrial Strategy **(S8)**. Already, Sensyne Health (https://www.sensynehealth.com) are developing an app for modelling diabetes using the service. The app aims to predict the need for intervention based upon blood glucose level monitoring and the development and validation of this app is made possible through using the synthetic data service.

2) Expanding the UK Economy – HealthTech is now the largest employer in the broader Life Sciences sector, employing 131,800 people in 4,060 companies, with a combined turnover of GBP25,600,000,000 (according to the association of British HealthTech Industries - https://www.abhi.org.uk). A report by the market research firm Mordor Intelligence suggests that the global mobile health market will reach GBP46,370,000,000 in 2021. This market is expected to save the UK NHS GBP2,000,000,000 per year (IQVIA Institute **S9**). According to Deloitte, the biggest barriers to this sector in the UK are cultural and regulatory **(S10)**. One stumbling block to expansion of this sector is the sharing of primary care data that is limited by data protection laws. The new synthetic datasets enable a rapid expansion of health-based apps that utilise historical primary care data, minimising risk of patient identification. This, in turn, will open up a new sector in machine-learning based health apps, which is currently not available, giving the UK a distinct competitive advantage by generating a new economy with the associated high-skill jobs, and offering considerable savings to the NHS.

3) Improving Public Health – The public will see the benefit of a whole new breed of apps that can learn from synthetic historical patient data, saving the NHS substantial costs through better monitoring and diagnosing. For example, diabetes costs the UK NHS GBP9,800,000,000 (diabetes.org.uk). The Sensyne app enables patients to reduce the need for appointments and to give them greater control over their care. The synthetic data generator has enabled Sensyne Health to generate new insights, develop new algorithms, and enable experiment repeatability whilst abiding by GDPR regulation. In addition, some clinically significant populations are under-represented in real-world data. The synthetic data generation methods enable them to produce arbitrarily large samples of such patients for analysis and algorithm development. **(S11)**. The Covid-19 synthetic data holds extremely valuable information about GP visits and has the potential to unlock some of the early signals of the arrival of Covid-19 into the UK based upon recorded symptoms, geographic location and demographic information. The potential of releasing the synthetic data on cardiovascular disease (CVD) is huge. CVD affects 7,600 000 people in the UK and accounts for 27% of all annual deaths – costing the economy GBP19,000,000,000. The opportunity for the health tech sector to develop and validate new apps that can predict and monitor cardiovascular risk using historical data with state-of-the-art machine learning technology, will benefit patients, clinicians and the economy.

5. Sources to corroborate the impact (indicative maximum of 10 references)

S1: Marco Scutari bnlearn package: https://www.bnlearn.com/research/statco19/

S2: Scalable bnlearn package: https://www.bnlearn.com/documentation/man/network.scores.html

S3: CRAN downloads statistics as of Nov 2019: https://cran.r-project.org/web/packages/dlstats/vignettes/dlstats.html

S4: CPRD letter of support (attached)

S5: MHRA letter of support (attached)

S6: NHS X Press Release: https://www.nhsx.nhs.uk/documents/8/NHSX_AI_report.pdf

S7: CPRD datalink: https://cprd.com/content/synthetic-data

S8: MHRA Press Release: https://www.gov.uk/government/news/new-synthetic-datasets-to-assist-covid-19-and-cardiovascular-research

S9: IQVIA Institute: https://www.basw.co.uk/system/files/resources/basw_55937-1_0.pdf

S10: Deloitte: https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/life-sciences-health-care/deloitte-uk-connected-health.pdf

S11: Sensyne Health letter of support (attached)