Improving Data Transparency and Accessibility in the Research Community through the Construction of Accurately Simulated Time-to-Event Datasets

Aiden Smith (✉ ajs134@le.ac.uk)

University of Leicester

Paul Lambert

University of Leicester

Mark Rutherford

University of Leicester

Research Article

Keywords: Simulation, Survival, Data Accessibility, Flexible Parametric Survival Models, Reproducible Research, Time-to-event, Synthetic Data

Posted Date: December 20th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-1170348/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background

A lack of availability of data and statistical code being published alongside journal articles provides a significant barrier to open scientific discourse, and reproducibility of research. Information governance restrictions inhibit the active dissemination of individual level data to accompany published manuscripts. Realistic, accurate time-to-event synthetic data can aid in the acceleration of methodological developments in survival analysis and beyond by enabling researchers to access and test published methods using data similar to that which they were developed on.

Methods

This paper presents methods to accurately replicate the covariate patterns and survival times found in real-world datasets using simulation techniques, without compromising individual patient identifiability. We model the joint covariate distribution of the original data using covariate specific sequential conditional regression models, then fit a complex flexible parametric survival model from which to simulate survival times conditional on individual covariate patterns. We recreate the administrative censoring mechanism using the last observed follow-up date information from the initial dataset. Metrics for evaluating the accuracy of the synthetic data, and the non-identifiability of individuals from the original dataset, are presented.

Results

We successfully create a synthetic version of an example colon cancer dataset consisting of 9064 patients which aims to show good similarity to both covariate distributions and survival times from the original data, without containing any exact information from the original data, therefore allowing them to be published openly alongside research.

Conclusions

We evaluate the effectiveness of the simulation methods for constructing synthetic data, as well as providing evidence that it is almost impossible that a given patient from the original data could be identified from their individual unique date information. Simulated datasets using this methodology could be made available alongside published research without breaching data privacy protocols, and allow for data and code to be made available alongside methodological or applied manuscripts to greatly improve the transparency and accessibility of medical research.

Background

It is common to simulate data with known properties to assess the performance of current and newly developed statistical methods. This approach of simulating data is more complex when the data are a time-to-event nature; with careful thought required on appropriate distributional assumptions and the

impact of censoring/competing risks. Often a series of simplistic assumptions are made for the data generating mechanisms, but there have been suggested improvements to create more realistic and complex data, whilst still retaining a known functional form. In the survival analysis framework, this can pertain to simulate biologically plausible data with flexible distributions for both the event censoring times(1, 2). Many simulation studies use data generated informed by real-world dataset in order to provide realistic scenarios. A further use of simulation approaches is to generate data that closely resembles a given dataset, but with the intent of releasing the simulated data when there are information governance restrictions on releasing the real individual patient data. It is this latter use of simulation approaches that we focus on in this manuscript. We aim to simulate data which shows good similarity to the covariate patterns and survival times of an existing real-world dataset.

It is a common request from journals for authors to release code and data alongside their manuscripts, however this is not always feasible due to data access and privacy constraints (3), with studies in cancer research releasing data in only 11-25% of publications (4, 5, 6). Through simulation methods, we can construct broadly similar datasets which appear and behave in a very similar way as the original data, which can be publicly released as they contain no real-world patient information. Reproducibility of research is paramount, and by releasing code and data alongside journal articles, we allow the opportunity to be able to reproduce the analyses being conducted. Limited access to statistical code is potentially a barrier to open science, methods development and reproducible research (7, 8), and having more broadly releasable data may aid in encouraging this practice.

We first construct a dataset which replicates the covariate distributions and patterns found in the original data using sequential regression modelling, then use complex survival modelling conditional on those covariate values to construct the simulated survival times.

In this paper, we provide a step by step overview of the methods developed to simulate a synthetic dataset which accurately replicates the data structure and survival patterns found in a real-world dataset, as well as providing evidence for choices made regarding the simulation process such as survival model complexity and covariate model ordering. The effectiveness of the simulation method is appraised by comparing agreement of both all-cause and relative survival metrics estimated in both the original data and the synthetic replica data. Evidence to demonstrate the unlikeliness of individual replication is provided through the use of a distance measure for individual patient date and survival time information. We also discuss further applications of these methods, as well as potential extensions to make it more broadly utilizable by researchers wishing to disseminate their work with accompanying data and code.

Methods

Motivating Dataset:

The initial data requires a set of core covariates potentially included in the survival model. In this example we focus on cancer survival, including age, stage and year of diagnosis, a time-to-event variable and an

indicator of vital status at the end of follow-up. These covariates are common in cancer research due to their influence on survival (9, 10, 11). There is scope to include further patient level information such as deprivation, treatment information or tumor grade and anatomical subsite where applicable (12, 13), depending on what the data availability and the aims of the analysis. These papers (9–13), like the majority of papers, did not publish data alongside the manuscript as a result of data confidentiality issues with identifiable patient level information. A concern would also be to create data which is so closely correlated, with patients with unique covariate patterns that simulated patients would have covariate values almost identical to real data individuals, however survival models have typically low levels of explained variation (R^2 values) and therefore the chances of fitting a model with such high degrees of correlation is very low (14). To demonstrate the simulation methods here, we use an example Colon cancer dataset (15) with diagnoses from 1985-1994 of 9,064 patients available from: <https://pclambert.net/data/colon.dta> (and as a .csv file in the supplementary materials), to create a simulated dataset from which to estimate a range of all-cause and relative survival metrics.

Table 1
Overview of Simulation Process to Construct Replicated Time-to-Event Datasets

Simulation Methods:
1. Clean and format the original data which is going to be replicated.
2. Define administrative censoring date and create an individual exit variable before assigning data as time-to-event.
3. Assign new factor levels for variables with missing data groups, generate dummy variables, and generate any required non-linear and interaction terms for the survival model.
4. Fit increasingly complex models for individual covariates and store model estimates from which to recover simulated covariate distributions.
5. Fit an all-cause survival model, including between-covariate interactions and time-dependent effects.
6. Set a seed and number of observations for the replica data, and using the stored model estimates for each covariate model, sequentially generate covariate values in the replica data based on conditional values of earlier covariates in the sequence.
7. Recreate any non-linear effects and model interactions which were included in the original survival model.
8. Use post estimation predict option from the survival model to generate simulated survival times based on individual patient covariate patterns from the stored survival model estimates.
9. Format vital status variable and generate diagnosis date and exit date variables.
10. Re-format vital status variable using exit date and the administrative censoring date to reconstruct the original data censoring distribution.
11. Clean and label all simulated variables.
12. Assign the simulated data as time-to-event for use in future survival analysis.

It is important to recover the distributions of all covariates included in the survival model, here: stage at diagnosis, sex, age-at-diagnosis and anatomical subsite. Similarly, to reconstruct diagnosis date which feeds into our exit date variable and the derivation of our administrative censoring date, we must accurately replicate the distribution of year at diagnosis within the whole dataset. The simulated survival times are generated using parameters in the flexible parametric all-cause survival model and generated based on new individual simulated patient covariate patterns.

All simulation methods and statistical analyses here were implemented using Stata 17 (16) with code provided as supplementary material, but could also be applied in other statistical software such as R using equivalent modelling commands.

Table 1 outlines the 12 steps required to execute the simulation process for generating synthetic data, each of which will be expanded on from here on in.

Data Cleaning and Formatting (Step 1, Table 1):

We require certain covariates in our initial data in order to simulate an accurate replica. In this example our initial dataset contains: calendar year of diagnosis (1985-1994), stage at diagnosis (“Localized”, “Regional”, “Distant” and “Missing”), sex (Male or Female), anatomical subsite (“Coecum and Ascending”, “Transverse”, “Sigmoid and Descending” and “Other and NOS”), age at diagnosis (18-99), vital status at time of censoring (Dead or Alive), unique patient id number, cancer diagnosis date and exit date (diagnosis date + survival time in days, administratively censored). We include covariates such as anatomical subsite, which is not necessarily included in the final analysis model, but incorporates additional heterogeneity to increase survival variation above and beyond that which is derived solely from the covariates which are included in the analysis model (17). The initial distributions of these covariates are shown in the first column of Table 2.

Table 2
Comparison of Original Data and Simulated Data Covariate Distributions

	Original Data (%)	Simulated Data (%)	Absolute Difference (%)
<i>Stage at Diagnosis</i>			
Localized	3716 (40.91)	3740 (41.45)	0.54
Regional	1148 (12.64)	1156 (12.81)	0.17
Distant	2907 (32.00)	2839 (31.46)	0.54
Missing	1313 (14.45)	1288 (14.27)	0.18
<i>Anatomical Tumour Subsite</i>			
Coecum and Ascending	3239 (35.66)	3163 (35.05)	0.61
Transverse	1607 (17.69)	1635 (18.12)	0.43
Sigmoid and Descending	3660 (40.29)	3647 (40.42)	0.13
Other and NOS	578 (6.36)	578 (6.41)	0.05
<i>Age Group</i>			
< 45	379 (4.17)	365 (4.05)	0.12
45-60	1338 (14.73)	1527 (16.92)	2.19
60-75	3699 (40.72)	3832 (42.47)	1.75
> 75	3688 (40.38)	3299 (36.56)	3.82
<i>Sex</i>			
Male	3799 (41.82)	3731 (41.35)	0.47
Female	5285 (58.18)	5292 (58.65)	0.47
<i>Vital Status</i>			
Alive	3557 (39.16)	3609 (40.00)	0.84
Dead	5527 (60.84)	5414 (60.00)	0.84

Defining Date Information (Step 2, Table 1):

Following data cleaning, step 2 states we must define an administrative censoring date from the original data to be used in the final simulated dataset. This is defined as the last observed date of follow-up found in the original data, here this is 31/12/1995. Using this censoring date, we create an exit date variable to help define our data as time-to-event, constructed by adding the survival time in days to the diagnosis date for every individual in the dataset. If an individual's last date of follow-up exceeds the censoring date, is it reassigned to be equal to our administrative censoring date.

Data Preparation for the Survival Model (Step 3, Table 1):

Step 3 of the simulation process requires the construction of dummy variables for our factor variables which will be included in the survival model (stage, subsite, sex), following the reassignment of missing data as its own factor level. Here, we have missing data in stage at diagnosis. To replicate patterns of missingness, we assign missing values as an additional level in the covariates of interest and treat them as their own unique group. The necessity of this process depends on the intended use for the data. If the data is being used to assess imputation methods, it is paramount to have missing data simulated as accurately as possible, however if the data is solely for testing analysis methods it becomes less important.

Model Fitting for Covariate Pattern Recovery (Step 4, Table 1):

To draw simulated survival predictions from the fitted survival model, it is necessary to first accurately recapture the joint distribution of the covariates which inform the survival model. Step 4 involves creating individual models for each covariate to model the joint distribution using a series of conditional distributions which are sequentially more complex, with example code provided below. For the process of simulating covariate values for individual i for the M variables of interest, we fit M regression models, with each variable, m ($= 1..M$), being the outcome for one of the M models. The form of the m^{th} regression model is dependent on the variable type for m (categorical, continuous, etc.). We start with a base model with no other covariates to get the marginal distribution of our first covariate of interest (age in the example below). We then include any prior variables ($1, \dots, m - 1$) in the process in all subsequent models; effectively with the final M^{th} model containing all the other variables and corresponding interaction terms. We include both the main covariate effects and the interaction effects between the specified covariates. Factor variables are fitted using a multinomial logistic regression model. For continuous covariates we propose a linear regression model with spline terms for an inverse-normal-rank-based-transformation of the covariate of interest in order to approximate continuous covariates with non-normal distributions. The ranks of the original covariate are converted to a normal distribution (18) and then we flexibly model the relationship between the original covariate and the inverse rank-normal values using restricted cubic splines to allow for a non-normally shaped distribution to be captured. Increased flexibility could be incorporated with an increased number of degrees of freedom for the rank splines if an even more precise age distribution was required.

Continuous covariates could also be simulated by specifying skewness and kurtosis values to allow non-normal distributions or a transformation applied to approximate normality. The effectiveness of each method will vary depending on the data, but we have found the method of using the restricted cubic splines based on the ranks to be robust to a range of distributions for the original covariate (see *Supplementary Material 2*). The ordering of covariate models fitted can make some difference in producing survival predictions which are similar to the original data, and in other datasets the best-fitting order will depend on the relationships between covariates defined through the interaction terms.

Fitting the Survival Model (Step 5, Table 1):

The 5th step is to fit a survival model. The survival time predictions we later generate are derived from the parameters from a flexible parametric survival model (19), implemented using `stpm2` in Stata (20). These models are highly flexible and allow for non-proportional hazards. The use of restricted cubic spline functions allows for far more complex hazard functions than in standard parametric models. We model age using restricted cubic splines to allow for non-linear effects, and use Winsorizing for age, below the 2nd percentile and above the 98th percentile of the age distribution to aid with model fitting (21). Two-way interaction terms are generated and included in the survival model for the age splines and stage at diagnosis, anatomical subsite and sex. Age and stage at diagnosis are all included as time-dependent effects. The overall model is fitted with 5 degrees of freedom, and 3 degrees of freedom for each time-dependent effect.

An important issue is the level of modelling complexity required. We can simulate from a simple survival model or a very complex one, dependent on what we aim to gain from the data. A simple model with no time-dependent effects and no interactions will give broadly similar marginal all-cause survival estimates, however when estimating more complex metrics, or looking at comparisons within missing data groups, the complex model may provide estimates more similar to that which was present in the original data. It is necessary to simulate from a survival model which is at least as complex as the analysis model we intend to use in the simulated data. It may be beneficial to simulate from an even more complex model to ensure we do not create an artificially well-fitting dataset to the model we intend to fit for the analysis.

Recovering Covariate Distributions and Survival Model Terms (Step 6 & 7, Table 1):

Having modelled our covariate distributions and fitted the survival model from which to draw predictions, we now start with a blank dataset from which to recreate the data based on the stored parameter estimates from models fitted to the original data. Step 6 requires setting the simulation seed and number of observations to be generated, the covariate distributions can be recreated by sampling from the stored covariate models. The covariate models are restored in the same order the models were fitted to ensure all simulations have the necessary information. For factor variables direct probabilities define the proportions in each group. *Figure 1* demonstrates the effectiveness of using the inverse-normal-rank-based transformation with the age distribution being recovered accurately due to the increased flexibility away from a rigid normal distribution (see Supplementary Material 2 for further demonstration of the effectiveness of this method for varying distributions).

[IMAGE-C:\Workspace\ACDC\ImageHandler\cbThe age splines, and any interaction terms fitted in the original survival model also need to be recreated prior to drawing survival predictions from the stored model estimates, per step 7.

Generating Simulated Survival Times (Step 8, Table 1):

Step 8 generates survival times from the overall survival model fitted to the original data, using the inversion methods where a random uniform (0,1) variable is transformed to a survival time using Brent root finding methods. This is implemented using the `stpm2` post-estimation predict command, individual survival times dependent on patient covariate patterns are simulated. The survival times depend on the randomly generated covariate patterns and thus none of the original data is directly simulated from.

Formatting Vital Status and Date Information (Step 9, Table 1):

Step 9 codes vital status using the last observed exit time from the original data as a threshold for if an event (death) has been experienced for any given individual. Survival times are rounded up to the nearest full day.

To reconstruct diagnosis date, month and day of diagnosis are generated uniformly across the calendar year, reflecting the patterns of diagnosis found in the original data. For disease areas where diagnoses are seasonal, month of diagnosis could be treated the same as year of diagnosis by more precisely recreating the distribution of diagnoses across the year. However, cancer is not a disease in which diagnoses varies seasonally, so in this instance this is not a major concern.

Replicating Administrative Censoring in the Simulated Data (Step 10, Table 1):

Step 10 calculates survival time in days in order to define an exit date for all individual simulated patients, and then administrative censoring is implemented using a final date of follow-up taken from the original data (31/12/1995). Further vital status coding occurs using the generated exit times, with patients labelled as alive if their exit dates exceed the final date of follow-up. Here we use a combination of calendar year and the administrative censoring date to replicate the censoring distribution found in the original data. For this method of simulation this is sufficient to accurately capture the known distribution (see Table 2), rather than modelling censoring directly through a Weibull or exponential distribution (22). We simulate a model based time-to-event distribution, and artificially construct our time-to-censoring distribution, whereas other simulation methods require model based estimates of both distributions.

Cleaning and Preparing Simulated Data for Survival Analysis (Step 11 & 12, Table 1):

Step 11 involves basic data cleaning, such as replacing assigned missing levels as truly missing data for stage at diagnosis, labelling variables and variable levels to match the original data and removing any unnecessary variables which have been created during the simulation process. Finally, in step 12 we declare our data as time-to-event in preparation for survival analysis.

Following the construction of the synthetic data the recovery of the covariate distributions will be evaluated against the original dataset. Comparisons of all-cause survival across the length of follow-up, and 5-year relative survival will be made between the true data and the synthetic data. In population-

based cancer studies, it is typical to report relative survival estimates. Relative survival estimates are used to try to isolate the mortality that is directly associated with the diagnosis of cancer by removing the impact of deaths due to causes other than cancer. This is achieved by incorporating general population background mortality rates, making it possible to isolate the theoretical survival that is only attributable to the disease being investigated (23, 24). The accuracy of the synthetic data here is evaluated across a range of metrics.

The consistency of a marginal 10-year relative survival estimates is compared with the true data across 25 different simulated datasets, and an age-specific life expectancy measure is also compared between the true data and 10 simulated datasets.

Finally, a single patient with a rare covariate pattern and unique age at diagnosis will be simulated 1000000 times and the date and survival time information compared between the single real patient and 1000000 synthetic patients to demonstrate that no individual from the original data can be identified in the synthetic data based on their unique date of diagnosis, birth date and survival time in days.

Results

Covariate Distribution Recovery

It is vital to replicate covariate distributions found in the original data with a high degree of accuracy in order to ensure that survival predictions being made conditional on unique covariate patterns are reflected appropriately in the synthetic data.

For factor covariates simulated from multinomial logistic regression models (stage at diagnosis, anatomical tumour subsite, sex, year of diagnosis), the absolute difference in factor level proportions between the original and simulated datasets never exceeds 1% (see Table 2). Similarly, there is only discrepancy of 0.84 percentage points in the proportion of subjects recorded as being dead.

In this example, the simulated data has fewer by 3.8 percentage points for the 75+ age category. As demonstrated in *Figure 1*, this redistribution of individuals does not vastly affect the overall age distribution being reconstructed.

Demonstrating Survival Pattern Replication

The effectiveness of these simulation methods in replicating survival patterns found in the original data can be demonstrated by comparing various survival estimates for the original and simulated data. Marginal all-cause survival estimates stratified by population subgroups are compared, as well as a model based estimate of 5-year relative survival.

Figure 2: Marginal all-cause survival estimates stratified by population subgroups (Subsite: Transverse and Coecum; Age Group: Under 45 and Over 75; Stage: Localized, Regional and Distant; Sex: Male and Female) comparing true estimates from the original and replicated estimates in the simulated data.

2 demonstrates the replication of the marginal all-cause survival estimates stratified by anatomical subsite (Coecum and Transverse), age group (under 45s and over 75s), stage at diagnosis and sex. There is some slight deviation in agreement in the age group subgroups, however it is still close enough to be broadly comparable when using the simulated data.

Figure 3: 5-year relative survival estimates stratified by population subgroups (Subsite: Transverse and Coecum; Age Group: Under 45 and Over 75; Stage: Localized, Regional and Distant; Sex: Male and Female) comparing original and simulated data.

Marginal 10-year restricted mean survival time is estimated for varying covariate patterns across 25 different simulation seeds and the distribution of the metric examined to compare point estimates. Individual conditional life expectancy estimates are also compared across simulation seeds for males across a range of ages and stages at diagnosis.

Figure 4: Marginal 10-year relative survival predictions estimated across assorted covariate patterns (Top Left: Stage: Localized, Sex: Male, Age Group: 45-60, Subsite: Sigmoid. Top Right: Stage: Distant, Sex: Male, Age Group: 60-75, Subsite: Coecum. Bottom Left Stage: Localized, Sex: Female, Age Group: <45, Subsite: Sigmoid. Bottom Right: Stage: Distant, Sex: Female, Age Group: >75, Subsite: Coecum.) for a range of 25 simulation seeds to demonstrate simulation consistency. (True prediction from original data shown as red horizontal line) Figure 4 demonstrates the consistency of the simulated data at closely replicating survival estimates regardless of the simulation seed specified in the code. Across 25 different seeds, 10-year restricted mean survival time predictions have been estimated and compared with the corresponding prediction made in the original dataset.

Figure 5: Conditional Life Expectancy estimated across a range of ages (40-90) for males with either a localized or distant cancer stage at diagnosis comparing original and simulated data (10 different seeds). Figure 5 demonstrates the effectiveness of the simulated data at replicating a conditional life expectancy metric where extrapolation of survival times beyond the known follow-up in the data is necessary. This metric is predicted for individual males diagnosed with both localized and distant stage cancer across ages ranging from 40-90, with the survival patterns replicated accurately when compared to the same metric in the original data.

Simulation Model Specifications

Two key questions for constructing the simulated dataset pertained to the complexity of survival model required to accurately represent individual survival times conditional on patient covariate patterns, and the ordering of covariates in the sequential conditional models used to recreate the overall joint distribution of covariates within the original dataset. The simulated datasets here have been expanded to 1000000 observations in order to eliminate differences in survival patterns due to random variation.

Figure 6: Kaplan-Meier estimates with marginal predictions in the corresponding population subgroups (Subsite: Transverse and Sigmoid; Age Group: Under 45 and Over 75; Stage: Localized and Distant; Sex: Female), from two flexible parametric models of differing complexity compared with equivalent estimates

from the original data. Figure 6 shows that across all population subgroups it is beneficial to include added complexity, through interaction terms, to best capture survival patterns across all unique covariate patterns. The simple model includes only main effects for stage, subsite, sex and age, without consideration of interactions and time-varying effects. The complex model builds on the simple model by including interactions between stage and age, stage and sex, subsite and age, and finally sex and age. It also includes time-varying effects for age and stage. In this example, the covariate order which is most consistent with the true survival estimates is as follows: Age, Year of Diagnosis, Stage at Diagnosis, Sex, and Anatomical Subsite. This order, as demonstrated in Figure 7, provides more accurate all-cause survival estimate regeneration compared to ordering the covariates from most to least categorical levels and vice versa. Ordering the age model first provides the best results and allows the other covariate models to be adjusted for age which is highly prognostic. The similarity of these simulated covariate distributions in comparison to the original data is shown in Table 2. More detailed cross-tabulations of covariate distribution recovery can be seen in *Supplementary Material 1*.

Figure 7: Kaplan-Meier estimates with marginal predictions in the corresponding population subgroups (Subsite: Sigmoid; Age Group: Over 75; Stage: Distant; Sex: Female), where covariate distributions are modelled in different orders. Ascending Order: Sex, Stage, Subsite, Year of Diagnosis; Descending Order: Year of Diagnosis, Subsite, Stage, Sex; Chosen Order: Year of Diagnosis, Stage, Sex, Subsite).

Non-Replication of Exact Patient Information

A significant concern with simulating patient level information to replicate survival patterns from a real-world dataset is the issue of patient confidentiality. Despite every single data value for every single hypothetical patient being simulated, it may be necessary to convince data controllers that the synthetic data is still not identifiable. Using this simulation method, there are very small probabilities that any single uniquely measured covariate pattern and corresponding survival time and date information will be replicated for any single individual within the dataset. Here, we identify a single individual within the data with a rare covariate pattern (Female, 60-74 Age Group, Missing Stage at Diagnosis, 1990 Calendar Year at Diagnosis and "Other and NOS" Anatomical Subsite) and simulate 1 million observations to attempt to replicate their information using the same simulation methodology.

A distance measure (measuring distance in days) using birth date, diagnosis date and survival time in days is constructed to assess if the simulated observations are exact matches to the original individual. From 3 possible individuals out of the initial 9064 with the chosen covariate pattern, a single patient with a unique age at diagnosis of 65 was chosen. After simulating 1 million patients with the same rare covariate pattern using the same methods described in this manuscript to generate survival times and date information, 3071 had matching diagnosis dates, 225 had matching survival time in days, and 2450 had a matching birth date. Of the 1 million observations, there were no simulated individuals who had a distance score of zero days for all three covariates being measured. This indicates there is a very low chance of simulating an individual with exact times and dates as in the original data. Even if there were such an individual, it would be entirely due to chance.

Discussion

The simulation methods presented in this paper describe a way to accurately simulate time-to-event data through the use of flexible parametric survival models. By simulating covariate distributions which reflect those found in the original data, and using this to simulate survival times.

Alternative methods to de-identify data include suppression, generalization, date offsetting and noise injection (25). Simulating a dataset using the methods presented here does not risk re-identification of individuals from the original data, while still maintaining the overall structure and internal distributions.

Previous literature regarding simulation of survival data tends to focus on the creation of entirely new synthetic data under a range of set parameters (1, 2, 26, 27). Many simulation studies make use of either exponential or Weibull distributions (28, 29), however these are often not flexible enough to fully capture the shape of underlying hazard functions found in real-world clinical trial or population based data, where at least one turning point is observed in the hazard function (21). Making use of the flexibility in a Royston-Parmar model provides a good solution for replicating real-world survival data. We are aware of one other statistical package for directly generating synthetic versions of data, *synthpop* (30), however this uses a combination of parametric and classification and regression tree (CART) methods to accurately recreate covariate distributions, and is only available in R.

It is important to be able to accurately simulate the distributions of continuous variables, including the potential for non-normal distributions. We have shown in *Supplementary Material 2* that the inverse rank-based transformation is effective at capturing a range of distributions, including normal, bimodal, skewed and uniform. The distribution shape could be captured more closely with increased degrees of freedom for our splines, however the effectiveness of this method is also a function of the sample size of our initial data due to the innate nature of the ranking method being used.

A further consideration for the application of these methods is within smaller datasets such as clinical trials. Here we are demonstrating the efficacy of these methods in a dataset of approximately 9000 patients, and specific adaptations and considerations for smaller datasets will form part of future work.

A development of the example data we use here would be the inclusion of an increased number of covariates. To include additional covariates it is necessary to not only recover their distributions, but also include them in the survival model if we are interested in survival differences within that covariate. In doing so, we increase the complexity of the model significantly with each additional model parameter. It is important to balance increasing model complexity without overfitting the model (31). In real-world cancer registry data there is scope for much more detailed patient information, treatment information and molecular information, all of which could be incorporated using this methodology should it be required by the user. Any method which is effective at predicting survival could be used, for example data with high covariate to observation structure could benefit from lasso type modelling with shrinkage. Future work will extend into a competing risks framework.

Despite comparing relative survival estimates on the generated data, we originally fitted an all-cause model in the simulation procedure. We did so because of complications found when predicting from the relative survival model in certain settings, particularly when the relative survival curves for a specific covariate pattern reached a plateau. This led to unstable simulated values of the survival times because of the influence of the plateau on the survival time predictions, which are based on centiles of the survival distribution.

We handled factor variables with missing data by assigning the missing group as a separate individual level of the variable so that the patterns of missingness could be preserved and recovered in the simulation process. Through this paper we have shown that there is a negligible possibility for an individual from the original data to be identified within the simulated data, we have shown the all-cause and relative survival estimates are consistently accurate when compared to equivalent estimates made in the original data, and the covariate distributions are recovered accurately throughout.

Conclusions

The approach that we have outlined allows the possibility for realistically simulated individual patient data to be made available and thus protecting patient confidentiality. We have focused on the construction of data that are typical in population-based cancer survival research, but the ideas are transferrable to other time-to-event settings. Our proposal to closely capture the distribution of continuous, but non-normal covariates has great flexibility. The benefits of making data available with corresponding code to run the main analyses of a published paper are numerous; allowing reproducible research, offering greater transparency and enabling extensions by other researchers.

Declarations

Ethics Approval and Consent to Participate: Ethics approval was not required for this epidemiological simulation study.

Consent to Publish: Not applicable.

Availability of Data and Materials: The datasets generated and analyzed during this study are included in this published article and its supplementary information files.

Competing Interests: The authors declare that they have no competing interests.

Funding: This work was supported by Cancer Research UK (Grant Number C41379/A27583). The study funders had no role in the design of the study, the collection, analysis and interpretation of the data, the writing of the paper or the decision to submit the paper for publication. This study is supported by the National Institute for Health Research (NIHR) Applied Research Collaboration East Midlands (ARC EM). The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

Authors Contributions: Study concept and design: A.S., M.R. and P.L.; acquisition of data: A.S., M.R. and P.L.; statistical analysis: A.S.; interpretation of data: A.S., M.R. and P.L.; drafting of the paper: A.S.; critical revision of the paper for important intellectual content: M.R. and P.L.

Acknowledgements: Not applicable.

References

1. Crowther, M. and Lambert, P. (2012). Simulating complex survival data. *The Stata Journal*, 674-687.
2. Crowther, M. and Lambert, P. (2013). Simulating Biologically Plausible Complex Survival Data. *Statistics in Medicine*, 4118-4134.
3. Gonclaves, A., Ray, P., Soper, B., Stevens, J., Coyle, L. and Sales, A. (2020). Generation and evaluation of synthetic patient data. *BMC Medical Research Methodology*.
4. Prinz, F., Schaalnege, T. and Asadullah, K. (2011). Believe it or not: how much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery*.
5. Begley, C. and Ellis, L. (2012). Raise standards for preclinical cancer research. *Nature*, 531-533.
6. Mikayawa, T. (2020). No raw data, no science: another possible source of the reproducibility crisis. *Molecular Brain*.
7. Goldacre, B., Morton, C. and DeVito, N. (2019). Why researchers should share their analytical code. *British Medical Journal*.
8. DeBlanc, J., Kay, B. and Lehigh, J. (2020). Availability of Statistical Code from Studies Using Medicare Data in General Medical Journals. *JAMA International Medicine*.
9. Smith, A., Lambert, P. and Rutherford, M. (2020). Understanding the impact of sex and stage differences on melanoma cancer patient survival: a SEER-based study. *British Journal of Cancer*.
10. Rutherford, M., Andersson, T., Bjorkholm, M. and Lambert, P. (2019). Loss in life expectancy and gain in life years as measures of cancer impact. *Cancer Epidemiology*, 168-173.
11. Syriopoulou, E., Rutherford, M. and Lambert, P. (2020). Understanding disparities in cancer prognosis: An extension of mediation analysis to the relative survival framework. *Biometrical Journal*, 341-353.
12. Wang, T., Nelson, R., Bogardus, A. and Grannis, F. (2010). Which advanced stage non-small cell lung cancer patients attain long-term survival? *Cancer*, 1518-1525.
13. Syriopoulou, E., Morris, E., Finan, P., Lambert, P. and Rutherford, M. (2019). Understanding the impact of socioeconomic differences in colorectal cancer survival: potential gain in life-years. *British Journal of Cancer*, 1052-1058.

14. Stare, J., Pohar Perme, M. and Henderson, R. (2011). A measure of explained variance for event history data. *Biometrics*, 750-759.
15. Dickman, P., Sloggett, A., Hills, M. and Hakulinen, T. 2004. Regression models for relative survival. *Statistics in Medicine*, 51-64.
16. StataCorp. (2021). Stata Statistical Software: Release 17. College Station, TX: StataCorp LLC.
17. Royston, P. and Parmar, M. (2002). Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine*, 2175-2197.
18. Lambert, P. and Royston, P. (2009). Further development of flexible parametric models for survival analysis. *The Stata Journal*, 265-290.
19. Syriopoulou, E., Mozumder, S., Rutherford, M. and Lambert, P. (2018). Robustness of individual and marginal model-based estimates: A sensitivity analysis of flexible parametric models. *Cancer Epidemiology*, 17-24.
20. Pain, O., Dudbridge, F. and Ronald, A. (2017). Are your covariates under control? How normalization can re-introduce covariate effects. *European Journal of Human Genetics*, 1194-1201.
21. Royston, P. (2012). Tools to simulate realistic censored survival-time distributions. *The Stata Journal*, 639-654.
22. Pohar Perme, M., Stare, J. and Esteve, J. (2011). On estimation in relative survival. *Biometrics*, 113-120.
23. Kronin, K. and Feuer, E. (2000). Cumulative cause-specific mortality for cancer patients in the presence of other causes: a crude analogue of relative survival. *Statistics in Medicine*, 1729-1740.
24. Emam, K. and Hoptroff, R. (2019). The synthetic data paradigm for using and sharing data. *Data Analytics and Digital Technologies*.
25. Wan, F. (2020). Simulating survival data with predefined censoring rates under a mixture of non-informative right censoring schemes. *Communications in Statistics- Simulation and Computation*.
26. Bender, R., Augustin, T. and Blettner, M. (2005). Generating survival times to simulate cox proportional hazards models. *Statistics in Medicine*, 1713-1723.
27. Rashid, I., Marcheselli, L. and Federico, M. (2008). Estimating survival in newly diagnosed cancer patients: use of computer simulations to evaluate performances of different approaches in a wide range of scenarios. *Statistics in Medicine*, 2145-2158.

28. Belot, A., Abrahamowicz, M. and Giorgi, R. (2010). Flexible modelling of competing risks in survival analysis. *Statistics in Medicine*.
29. Nowok, B., Raab, G. and Dibben, C. (2016). synthpop: Bespoke creation of synthetic data in R. *Journal of Statistical Software*, 1-26.
30. Azizi, Z., Zheng, C., Mosquera, L., Pilote, L. and El Emam, K. (2021). Can synthetic data be a proxy for real clinical trial data? A validation study. *BMJ Open*.
31. Coolen, A., Barrett, J., Paga, P. and Perez-Vicente, C. (2017). Replica analysis of overfitting in regression models for time-to-event data. *Journal of Physics A: Mathematical and Theoretical*.

Figures

Figure 1

Comparison of original data age-at-diagnosis distribution with simulated data age-at-diagnosis distribution derived using the Inverse-Normal-Rank-Based transformation method.

Figure 2

Marginal all-cause survival estimates stratified by population subgroups (Subsite: Transverse and Coecum; Age Group: Under 45 and Over 75; Stage: Localized, Regional and Distant; Sex: Male and Female) comparing true estimates from the original and replicated estimates in the simulated data.

Figure 3

5-year relative survival estimates stratified by population subgroups (Subsite: Transverse and Coecum; Age Group: Under 45 and Over 75; Stage: Localized, Regional and Distant; Sex: Male and Female) comparing original and simulated data.

Figure 4

Marginal 10-year relative survival predictions estimated across assorted covariate patterns (Top Left: Stage: Localized, Sex: Male, Age Group: 45-60, Subsite: Sigmoid. Top Right: Stage: Distant, Sex: Male, Age Group: 60-75, Subsite: Coecum. Bottom Left Stage: Localized, Sex: Female, Age Group: <45, Subsite: Sigmoid. Bottom Right: Stage: Distant, Sex: Female, Age Group: >75, Subsite: Coecum.) for a range of 25 simulation seeds to demonstrate simulation consistency. (True prediction from original data shown as red horizontal line)

Figure 5

Conditional Life Expectancy estimated across a range of ages (40-90) for males with either a localized or distant cancer stage at diagnosis comparing original and simulated data (10 different seeds).

Figure 6

Kaplan-Meier estimates with marginal predictions in the corresponding population subgroups (Subsite: Transverse and Sigmoid; Age Group: Under 45 and Over 75; Stage: Localized and Distant; Sex: Female), from two flexible parametric models of differing complexity compared with equivalent estimates from the original data.

Figure 7

Kaplan-Meier estimates with marginal predictions in the corresponding population subgroups (Subsite: Sigmoid; Age Group: Over 75; Stage: Distant; Sex: Female), where covariate distributions are modelled in different orders. Ascending Order: Sex, Stage, Subsite, Year of Diagnosis; Descending Order: Year of Diagnosis, Subsite, Stage, Sex; Chosen Order: Year of Diagnosis, Stage, Sex, Subsite).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryMaterial1.docx](#)
- [SupplementaryMaterial2.docx](#)
- [syntheticdatagenerationfinalannotated.do](#)
- [generatingsyntheticdatastatacode.pdf](#)

- [colon.csv](#)