EDITORIAL

Case-cohort design in hematopoietic cell transplant studies

Series Editors– Note

Imagine you and your colleagues have done 1000 transplants in persons with acute myeloid leukaemia (AML) in 1st remission. 5 percent of the 20 percent of recipients relapsing posttransplant have an isolated central nervous system relapse. You are curious and want to know whether there is anything special about this 5 percent, specifically whether this risk corelates with any pretransplant clinical and laboratory co-variates. You have extensive clinical data and some typical laboratory data on all 1000 but you suspect the culprit is mutation topography. What to do? Fortunately you have bio-banked DNA from the 1000. If resources and monies are not limiting you can do targeted or next generation sequencing on all 1000 DNA samples and off you go. However, most of us lack unlimited resources and monies. How can you sensibly and efficiently tackle this research problem? The answer is a case-cohort design study. In the typescript which follows Profs. Cai and Kim explain how to accomplish this. If you follow their advice you may need only to analyze samples from < 300 recipients rather than 1000 to test your hypothesis. They explain how to design such a study and provide references to estimate sample size.

Sadly, their typescript will not tell you how to get funding for the study, whish poor devil who will have to write the protocol, worse, who will shepherd it though endless committees for approval and the like. Help on these issues is outside the scope of our statistics series. In this context we suggest advice from Woody Allen's article in the New Yorker: *The Kugelmass Episode* (April 24, 1977). When Prof. Kugelmass (English, City College) tells his analyst Dr. Mandel he has fallen in love with Emma Bovary who died of arsenic poisoning near Rouen, France 120 years earlier the analyst says: *After all, I'm an analyst, not a magician*. Kugelmass' reply: *Then perhaps what I need is a magician* and is off to Coney Island to find one. Good luck, the magician may still be there! (Note: This typescript is R-rated. It contains an equation.)

Robert Peter Gale, Imperial College London, and Mei-Jie Zhang, Medical College of Wisconsin and CIBMTR.

Bone Marrow Transplantation (2022) 57:1–5; https://doi.org/10.1038/s41409-021-01433-4

INTRODUCTION

Case-cohort study-design, 1st proposed by Prentice in 1986 is a commonly-used cost-effective outcome-dependent study-design embedded in large cohort studies [1–3]. This design is used to reduce costs or conserve resources when the rate of the outcome event of interest is low and/or resources to ascertain exposure data are limited. The case-cohort sample consists of a (stratified) random sample of the full cohort supplemented by cases who are not in the random sample. There are several advantages to the case-cohort design: (1) it reduces the cost/effort for collecting redundant data on non-cases; (2) the random sample can be used for monitoring study progress; (3) data collected through a case-cohort study-design can be used to study the prospective relationship between the exposure and the outcome; and (4) because the random sample is selected independent of the outcome of interest collected exposure data can be used to study other outcomes of future studies. The nested case-control study is an alternative study design to the case-cohort design. In a nested case-control study, controls are selected at each failure time, consequently there is no representative random sample from the full cohort and the data collected from one nested case-control study cannot be easily used to study other outcomes.

The case-cohort study design can be used in transplant research. For example, the Center for International Blood & Marrow Transplant Research (CIBMTR) has two levels of data collection: (1) Transplant Essential Data (TED); and (2) Comprehensive Report Form (CRF) data. Collecting CRF data takes more resources than TED data. Transplant centers designated as CRF centers collect CRF data on some but not all recipients at their center. CRF data include detailed information such as co-variates as pretransplant conditioning, acute graft-*versus*-host disease (GvHD) *etc*.

Consider a study correlating to identify pretransplant co-variates with risk of developing a central nervous system (CNS) cancer posttransplant [4]. Posttransplant CNS cancers are rare occurring in <1% of recipients. A case-cohort study can be an efficient way to interrogate this question. At a CRF center one could select a random recipient sample and select all recipients developing a CNS cancer. CRF data can then be collected on the selected random sample and on the few subjects with CNS cancer. CRF data could include co-variates such as age at radiotherapy, prior CNS radiation exposure to anti-cancer drugs crossing the blood brain barrier, GvHD, corticosteroid exposure and others.

Competing risks are common in transplant recipients studies including death from leukemia recurrence before developing a CNS cancer. It is important to analyze competing risks data from case-cohort studies properly. In this tutorial we briefly describe case-cohort study-design and data available from a case-cohort design. We also introduce the commonly used analytic method,

Received: 12 July 2021 Revised: 14 July 2021 Accepted: 3 August 2021
Published online: 16 August 2021

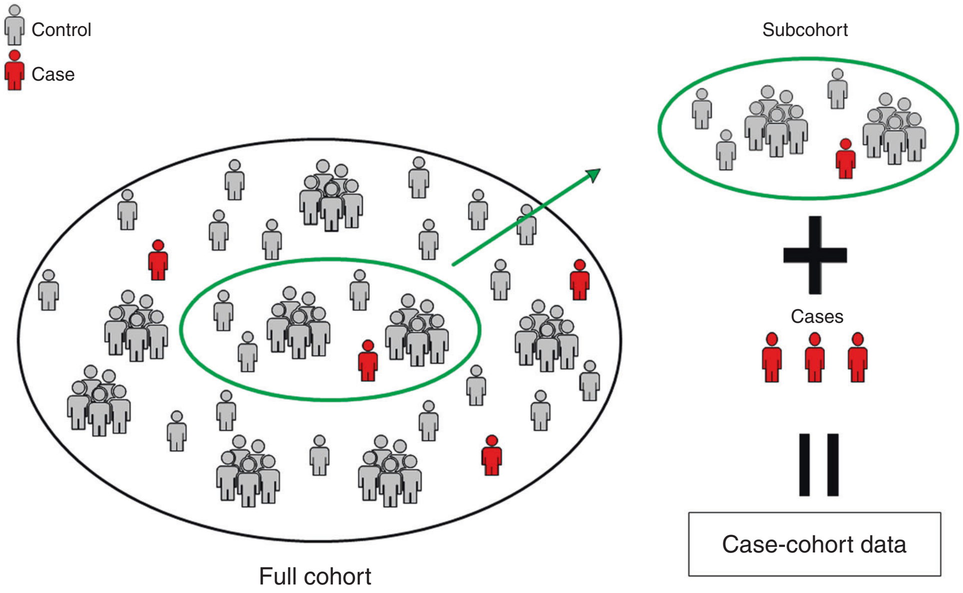

Fig. 1 Case-cohort design. illustration for subjects selection in the case-cohort design.

the cause-specific hazards model, and software for analyzing data from case-cohort studies with competing risks.

CASE-COHORT STUDY DESIGN AND THE DATA STRUCTURE

Let T_i and C_i be the potential failure and censoring times and μ_i (=1, 2, …, or K) denote the cause of failure for subject i (= 1,…, n). Without losing generality we denote the event of interest as 'cause 1' ($\mu_i = 1$) and refer to it as the 'cause of interest' or 'event of interest'. If there is only one cause of failure (i.e., K = 1) this reduces to the situation with a uni-variable survival outcome. Let X_i (i.e., minimum of T_i and C_i) and Δ_I (i.e., = 1 if T_i is observed before C_i and otherwise 0) denote the observed time and failure indicator. Let $Z_i(t)$ denote co-variates. For a case-cohort study, we sample a random sub-cohort of all subjects and all subjects with the event of interest regardless of whether they are in the selected subcohort. Figure 1 provides an illustration on the case-cohort sample. Co-variate information $Z_i(t)$ can be decomposed into two parts as $Z_i(t) = (Z_{iC}(t), Z_{iE}(t))$, where $Z_{iC}(t)$ are available on the entire cohort and $Z_{iE}(t)$ are co-variates only available for subjects in the case-cohort sample. For example, $Z_{iE}(t)$ can include the CRF data such as pretransplant radiation dose and $Z_{iC}(t)$ can include TED level data such as age at transplant and sex. Let ξ_i be an indicator for subject i being selected into the sub-cohort. The observable data is $\{X_i, \Delta_i, \Delta_i\mu_i, \xi_i, Z_{iC}(t), Z_{iE}(t)\}$ if subject i is in the case-cohort sample, and $\{X_i, \Delta_i, \Delta_i\mu_i, \xi_i, Z_{iC}(t)\}$ otherwise.

For example, suppose we are interested in assessing the impacts of mutations *ASXL1, EZH2, SRSF2, IDH1, IDH2*, and *TP53* on death [5]. Collecting these data from stored DNA samples is expensive. To reduce cost and preserve samples we can design a case-cohort study. Assume there are 1000 subjects in the full cohort, 20% die and we set the selection probability of the sub-cohort at 25%. The size of the case-cohort dataset is 400 subjects, 250 in the sub-cohort and 150 outside the sub-cohort. Overall, 200 subjects died and 200 are alive. In this scenario mutations data are collected on only these subjects whereas survival data and other co-variates such as age and sex are collected from all 1000 subjects in this study.

MODELS AND WEIGHTS FOR CASE-COHORT STUDIES

For competing risks data there are in general two commonly used models: (1) the cause-specific proportional hazards; and (2) sub-distribution hazards. The cause-specific hazards model is useful when one's interest is in studying disease etiology whereas the sub-distribution hazards model is of greater interest when the emphasis is on estimating actual risk and prognosis. Here we focus on cause-specific hazard model for case-cohort studies because of the availability of statistical software packages.

The hazard function in the cause-specific hazard model for cause *k* is given by:

$$\lambda_k(t|Z(t)) = \lambda_{0k}(t)exp(\beta_k Z(t)),$$

where $\lambda_{0k}(t)$ is an unspecified baseline hazard function and β_k is an unknown parameter of interest. The effects of risk factor for cause k outcome can be measured by the hazard ratio $exp(\beta_k)$. In the cause-specific hazard model one treats subjects who experienced competing risks as censored. When there is only one cause (i.e., $K = 1$) the cause-specific hazard model is reduced to the Cox proportional hazards model.

Because we lack extensive co-variate data outside the case-cohort sample the estimation method for the Cox proportional hazards model needs to be modified. The so-called weighed partial likelihood is widely-used for case-cohort design. The key to the weighted partial likelihood is to understand the weighting of subjects with the event of interest and sub-cohort subjects without the event of interest. Several weighting functions for case-cohort design are proposed [6–8]. In this tutorial, we focus on a time-independent weight function which uses the sub-cohort sampling probability, denoted by α. Specifically, weights for subjects with the event of interest is 1 because all subjects in the full cohort with the event of interest are included in the case-cohort sample i.e., cases in the case-cohort sample are all cases in the full cohort. In contrast, some subjects without the event of interest are not in the case-cohort sample. Consequently, sub-cohort subjects without the event of interest are weighted by 1/α. For example, suppose α is 25%. Then the weight for subjects in the sub-cohort who do not experience the event of interest is 1/0.25 = 4 indicating one subject in the sub-cohort without the event represents four subjects without the event in the full cohort. In practice sampling probability α is unknown and needs to be estimated.

To analyze case-cohort data using SAS (PHREG procedure), two steps are required. Step (1) create weights for each subject. Step (2) calculate the robust variance to account for case-cohort

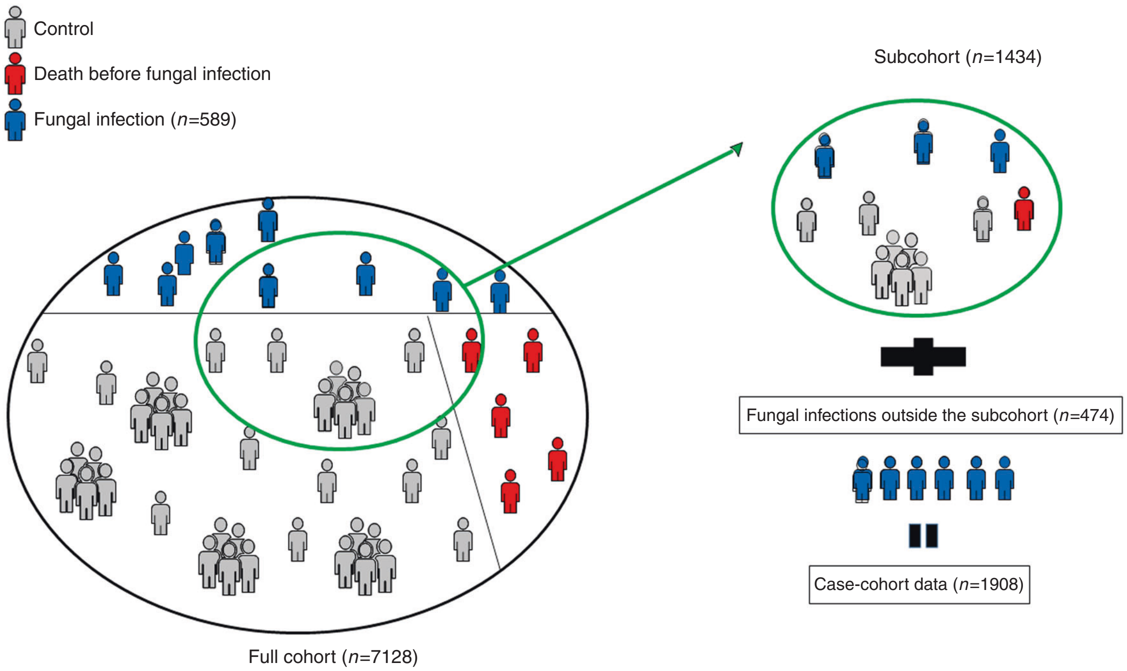
Control
Death before fungal infection
Fungal infection (*n*=589)
Subcohort (*n*=1434)
Fungal infections outside the subcohort (*n*=474)
Case-cohort data (*n*=1908)
Full cohort (*n*=7128)

Fig. 2 Case-cohort example. the case-cohort sample for the fungal infection.

data structure. The example SAS code is provided in the Supplementary material. In PHREG procedure, "COVS(AGGREGATE)" and "ID" statement options allow to calculate robust sandwich type of variance. The *R* statistical package provides similar capabilities. An example of *R* code is in the Supplementary material. We now show how to fit these cause-specific models using CIBMTR data.

EXAMPLE

Consider the transplant dataset reported by Ustun et al. (2018) of 7128 subjects receiving a 1st allograft for acute myeloid leukemia, acute lymphoblastic leukemia, or myelodysplastic syndrome from January, 2008 to December, 2012 [9]. The primary outcome of interest is a fungal infection in this data. 589 (8%) had a fungal infection by day 100 and 1059 (15%) died without a fungal infection before day 100. In a case-cohort study we create a case-cohort sample by randomly selecting 20% of subjects from the 7128 to 1434 subjects to form a sub-cohort. Next, we add everyone not in the sub-cohort who had a fungal infection before day 100 (Fig. 2). 115 of the 1434 randomly-selected subjects had a fungal infection before day 100, 163 died before day 100 without a fungal infection and 1156 had neither a fungal infection nor died before day 100. Next, we add 474 subjects (589–115) with a fungal infection before day 100 not in the randomly-selected sub-cohort bringing numbers of subjects in the case-cohort sample to 1908 (1434 + 474). In this case-cohort sample, 1319 (1434 − 115) did not have a fungal infection before day 100 and were weighted by 1/0.2 = 5 whereas 589 had a fungal infection before day 100 and were weighted by 1.

Co-variates of interest in this study were age at transplant, graft-type, GvHD prophylaxis, and year of transplant. We checked the proportional hazards assumption by testing whether the coefficient of log transformed time × each co-variate is equal to zero and all *p* values were >0.05.

Table 1. Frequencies of the cohorts.

	Full cohort (*n* = 7128)		**Sub-cohort (*n* = 1434)**	
Co-variates	**Frequency**	**Percent**	**Frequency**	**Percent**
Age at transplant (years)				
≤20	1423	20	297	21
21–40	1439	20	293	20
41–50	1125	16	205	14
51–60	1641	23	319	22
>60	1500	21	320	22
Graft-type				
Bone Marrow	1067	15	217	15
Blood	4227	59	845	59
Umbilical cord blood	1834	26	372	26
GvHD prophylaxis				
Other	163	2	28	2
T-cell depletion	148	2	27	2
FK506/CSA + MMF ± other	2359	33	481	34
FK506/CSA + MTX ± other	3577	50	729	51
FK506/CSA + other	655	9	130	9
FK506/CSA	226	3	39	3
Transplant year				
2008–2009	3831	54	766	53
2010–2012	3297	46	668	47

FK506 tacrolimus, *CSA* cyclosporine, *MMF* mycophenolate mofetil, *MTX* methotrexate.

Data of co-variate frequencies in the full and sub-cohorts displayed in Table 1 indicate reasonable comparability. Next, we fit the cause-specific hazard model using the case-cohort sample and fit the same model using the full cohort to compare results. Note the full cohort analysis is only possible because we generated the case-cohort sample from the full cohort. This full cohort analysis would not be possible in real

Table 2. Analyses using the cause-specific hazards model.

	Full cohort ($n = 7128$)			**Case-cohort ($n = 1908$)**		
Co-variates	**HR**	**95% CI**	***p* value**	**HR**	**95% CI**	***p* value**
Age at transplant (years)			0.003			0.030
≤20 (Reference)	1.00			1.00		
21–40	1.30	(1.01, 1.66)	0.039	1.21	(0.90, 1.64)	0.208
41–50	0.97	(0.73, 1.30)	0.857	1.05	(0.74, 1.49)	0.808
51–60	0.77	(0.58, 1.02)	0.072	0.76	(0.54, 1.07)	0.120
>60	1.01	(0.76, 1.33)	0.950	1.16	(0.82, 1.64)	0.392
Graft-type			<0.0001			0.003
Bone Marrow (Reference)	1.00			1.00		
Blood	0.90	(0.69, 1.18)	0.446	0.84	(0.62, 1.15)	0.275
Umbilical cord blood	1.57	(1.16, 2.12)	0.003	1.42	(1.00, 2.02)	0.053
GvHD prophylaxis			0.003			0.011
Other (Reference)	1.00			1.00		
T-cell depletion	0.61	(0.3, 1.25)	0.179	0.76	(0.32, 1.80)	0.527
FK506/CSA + MMF ± other	0.57	(0.35, 0.91)	0.018	0.62	(0.35, 1.09)	0.094
FK506/CSA + MTX ± other	0.45	(0.29, 0.71)	0.001	0.50	(0.29, 0.87)	0.013
FK506/CSA + other	0.65	(0.39, 1.09)	0.103	0.86	(0.46, 1.60)	0.635
FK506/CSA	0.43	(0.23, 0.84)	0.013	0.49	(0.23, 1.06)	0.071
Transplant year						
2008–2009 (Reference)	1.00			1.00		
2010–2012	0.85	(0.72, 1.00)	0.048	0.82	(0.67, 0.99)	0.043

HR Hazard Ratio, *CI* Confidence Interval, *FK506* tacrolimus, *CSA* cyclosporine, *MMF* mycophenolate mofetil, *MTX* methotrexate.

case-cohort studies. Table 2 shows hazard ratios, 95% confidence intervals and *p* values. Hazard ratios based on the case-cohort sample are very close to those based on the full cohort. The data indicate age at transplant, graft-type, GvHD prophylaxis, and year of transplant are significantly correlated with risk of a fungal infection before day 100 in the full and the case-cohort sample. As expected, the 95% confidence intervals for the case-cohort ($N = 1908$) are wider than those for the full cohort ($N = 7128$).

CONCLUSION/DISCUSSION

Case-cohort design is an efficient, cost effective statistical method when an event(s) of interest is rare and/or when obtaining co-variate data is difficult and/or expensive and has great potential in hematopoietic cell transplant research. We provide a brief review of the case-cohort design and show how to properly analyze case-cohort data when there are competing risks using statistical software packages. In this tutorial we considered only cause-specific hazards models for competing risks but one can easily apply these weighting scheme to sub-distribution hazards model such as the Fine-Gray model [10, 11].

In our example we selected the sub-cohort by simple random sampling but stratified sampling can also be used to ensure balance for important co-variates. Also, in the tutorial we only considered time-independent weights. Several methods have been proposed to improve efficiency for case-cohort studies using time-dependent weights and extra information such as auxiliary co-variate data whereby time-dependent weights are calculated among subjects at-risk at each time point [12, 13]. The case-cohort design can also be used to analyze multiple outcomes [14–18]. Lastly, there are sample size and power calculations. Sample size estimation is an important first step for designing a study and formulae for these are available [19, 20].

We hope readers will find this discussion useful and share it with their center statisticians. We expect increased use of the case-cohort method to tackle important questions in hematopoietic cell transplantation in the near future.

Jianwen Cai[1] and Soyoung Kim[2]✉

[1]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [2]Division of Biostatistics, Medical College of Wisconsin, Wauwatosa, WI, USA. ✉email: skim@mcw.edu

REFERENCES

1. Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. Biometrika. 1986;73:1–1.
2. Knuiman MW, Divitini ML, Olynyk JK, Cullen DJ, Bartholomew HC. Serum ferritin and cardiovascular disease: a 17-year follow-up study in Busselton, Western Australia. Am J Epidemiol. 2003;158:144–9.
3. Ballantyne CM, Hoogeveen RC, Bang H, Coresh J, Folsom AR, Heiss G, et al. Lipoprotein-associated phospholipase A2, high-sensitivity C-reactive protein, and risk for incident coronary heart disease in middle-aged men and women in the Atherosclerosis Risk in Communities (ARIC) study. Circulation. 2004;109:837–42.
4. Gabriel M, Shaw BE, Brazauskas R, Chen M, Margolis DA, Sengelov H, et al. Risk factors for subsequent central nervous system tumors in pediatric allogeneic hematopoietic cell transplant: a study from the Center for International Blood and Marrow Transplant Research (CIBMTR). Biol Blood Marrow Transpl. 2017;23:1320–6.
5. Gupta V, Kennedy JA, Capo-Chichi JM, Kim S, Hu ZH, Alyea EP, et al. Genetic factors rather than blast reduction determine outcomes of allogeneic HCT in BCR-ABL–negative MPN in blast phase. Blood Adv. 2020;4:5562–73.
6. Self SG, Prentice RL. Asymptotic distribution theory and efficiency results for case-cohort studies. Ann Statistics.1988;16:64–81.
7. Barlow WE. Robust variance estimation for the case-cohort design. Biometrics. 1994;50:1064–72.
8. Borgan O, Langholz B, Samuelsen SO, Goldstein L, Pogoda J. Exposure stratified case-cohort designs. Lifetime Data Anal. 2000;6:39–58.
9. Ustun C, Young JA, Papanicolaou GA, Kim S, Ahn KW, Chen M, et al. Bacterial blood stream infections (BSIs), particularly post-engraftment BSIs, are associated

with increased mortality after allogeneic hematopoietic cell transplantation. Bone Marrow Transpl. 2019;54:1254–65.

10. Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. J Am Stat Assoc. 1999;94:496–509.
11. Kim S, Xu Y, Zhang MJ, Ahn KW. Stratified proportional subdistribution hazards model with covariate-adjusted censoring weight for case-cohort studies. Scand J Stat. 2020;47:1222–42.
12. Kulich M, Lin DY. Improving the efficiency of relative-risk estimation in case-cohort studies. J Am Stat Assoc. 2004;99:832–44.
13. Samuelsen SO, Ånestad H, Skrondal A. Stratified case-cohort analysis of general cohort sampling designs. Scand J Stat. 2007;34:103–19.
14. Kang S, Cai J. Marginal hazards model for case-cohort studies with multiple disease outcomes. Biometrika. 2009;96:887–901.
15. Kim S, Cai J, Lu W. More efficient estimators for case-cohort studies. Biometrika. 2013;100:695–708.
16. Kim S, Cai J, Couper D. Improving the efficiency of estimation in the additive hazards model for stratified case–cohort design with multiple diseases. Stat Med. 2016;35:282–93.
17. Kim S, Zeng D, Cai J. Analysis of multiple survival events in generalized case-cohort designs. Biometrics. 2018;74:1250–60.
18. Langholz B, Thomas DC. Nested case-control and case-cohort methods of sampling from a cohort: a critical comparison. Am J Epidemiol. 1990;131:169–76.
19. Cai J, Zeng D. Sample size/power calculation for case–cohort studies. Biometrics. 2004;60:1015–24.
20. Cai J, Zeng D. Power calculation for case-cohort studies with non-rare events. Biometrics. 2007;63:1288–95.

ACKNOWLEDGEMENTS

The authors would like to thank Drs Robert Peter Gale and Mei-Jie Zhang for inviting us to contribute this paper. This work was partially supported by grants from the National Cancer Institute (U24CA076518, and P01CA142538) and National Institute of Environmental Health Science (P30ES010126).

AUTHOR CONTRIBUTIONS

Both JC and SK wrote the paper and SK analyzed the data example.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41409-021-01433-4.

Correspondence and requests for materials should be addressed to S.K.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.