ASA Section on Lifetime Data Science

LiDS Newsletter

Volume 5, Number 1 – January 2020

ISSN 2473-5159

A section associated with the American Statistical Association

In Brief

* * * * * * * * * * * * *

`https://community.amstat.org/lids/home`
Published newsletters are archived under "Newsletters".

* * * * * * * * * * * * *

New Section Officers
Chair-elect: Dr. Mimi Y. Kim
Program chair-elect: Dr. Haoda Fu

2020 Election Candidates
Chair-elect, Treasurer, Program Chair-elect

LiDS Sessions at 2020 JSM
Two invited sessions; four topic-contributed sessions

2021 Conference on Lifetime Data Science
May 19–21, 2021, Raleigh, NC

Software Review
Small Sample Inferences with Kaplan–Meier Estimates with R Package **bpcp**

Regularized Cox Cure Rate Model with R package **intsurv**

* * * * * * * * * * * * *

Membership in the LiDS
Join us at `https://community.amstat.org/lids/about-lids/lids-join-us`

LiDS Officers

Chair:	Nicholas P. Jewell
Chair-Elect:	Mimi Y. Kim
Past Chair:	Jianwen Cai
Secretary 2019-2021:	Joan Hu
Treasurer 2016-2020:	Chiung-Yu Huang
Program Chair (2020):	Zhezhen Jin
Program Chair-Elect :	Haoda Fu
Past Program Chair:	Guoqing Diao
COS Representative:	Xiaonan Xue
Webmaster:	Ker-Ai Lee
Newsletter Editor:	Jun Yan

Chair's Message

Happy new year and welcome to our first newsletter of the year! It is a great honor for me to serve as Chair of LiDS for 2020 in our second year as an ASA Section. When you are updating your ASA membership this year, I encourage you to renew your section membership or join the section for the first time! Restating our objectives, the goal of LiDS is promotion and support for the development, application and appropriate use of statistical methods for the design and analysis of studies of life history processes. This challenge includes the development of new methods, the identification of new areas of application, and the fostering of interdisciplinary research across biomedical research, finance, economics, imaging, engineering, genomics, and genetics.

As our first year as an official ASA section, 2019 was a signal year for LiDS. It also saw the second LiDS conference on Lifetime Data Science: Foundations and Frontiers that was held at the University of Pittsburgh from May 29-31, 2019. The conference was again an enormous success, following on from the inaugural 2017 conference at the University of Connecticut. Great thanks are due to Richard Cook and Jianwen Cai, Co-Program Chairs, and their committee for a stimulating scientific program that included several short courses, a myriad of scientific sessions, and student paper and poster competitions. Special thanks also to Ying Ding and Yu Cheng, Co-Chairs of the Local Organizing Committee.

Encouraged by the strong demand for these events, the Executive Committee requested proposals for a 2021 LiDS conference. After review of submitted proposals, we have determined that the 2021 LiDS Conference will be held in North Carolina from May 19-21, 2021, most likely at the Marriott Raleigh City Center. Please mark your calendars with these dates and stay tuned for further details to come in the following months.

I want to take this opportunity to thank Richard Cook as he steps off the Executive Committee as Past-Chair. Richard has provided extraordinary leadership to LiDS for the past three years, particularly through the successful establishment of LiDS as an ASA Section. And he accomplished all this at the same time as contributing enormously to the 2019 conference in Pittsburgh as noted above.

Please let me also thank Jianwen Cai for her service this past year as Chair of the Section. Jianwen has been a powerful and effective leader and will be a very hard act to follow. I look forward to continue working with her this year as she remains on the Executive Committee as Past-Chair. Join me also in congratulating Mimi Kim, from the Albert Einstein College of Medicine, on her election as Chair-Elect. Mimi will serve as Chair in 2021, a period that will include the planned conference in North Carolina. I have known Mimi since she graduated with her Sc.D. from Harvard, and from her early work on HIV research. I am very much looking forward to collaborating with her again over these next two years.

We are also very appreciative that Chiang-Yu Huang has agreed to continue her role as Treasurer of the Section. Chiang-Yu has been doing a fantastic job, particularly as the Section has grown in membership and our activities have both become more intense and required coordination with the ASA. Finally, we also send our thanks to Yu Shen as she ends her three years of service as Program Chair

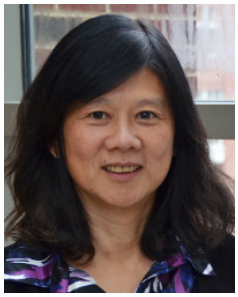
(including a first year as Program Chair Elect and as Past Program Chair in 2019). She has done a marvelous job and contributed richly to the Section's activities from which we all benefit. Zhezhen Jin now moves into the role of Program Chair while Guoqing Diao moves on to the position of Past Program Chair. The new Program Chair Elect is Haoda Fu–welcome! We all very much appreciate these volunteers for the enormous amount of work they have contributed, or are contributing, to the Section.

For additional news and information about LiDS, please visit the LiDS website at https://community.amstat.org/lids/home.

I again wish you a very happy new year for 2020! The ASA Section on LiDS has an exciting year ahead and we encourage you and your colleagues to join and get involved!

Nicholas P. Jewell, Chair 2020

Message from the Past Chair



It was my great pleasure and honor to have had the opportunity to work with the outstanding group of colleagues on the Executive Committee (EC) for the Lifetime Data Science (LiDS) section in 2019. I admire their dedication and positive can-do mentality. In particular, I greatly appreciate the excellent efforts and support of Nick Jewell as Chair-Elect, Joan Hu as Secretary, Chiung-Yu Huang as Treasurer, Yu Shen as Past Program Chair, Zhezhen Jin as the Program Chair Elect, Xiaonan Xue as Council on Sections Representative, and Jun Yan as Editor of the Newsletter.

I want to take this opportunity to thank Richard Cook for his outstanding leadership over the past three years. Richard continued the great work of the previous chairs and successfully obtained the approval of the section status for LiDS. Richard has also done a remarkable job orchestrating the highly successful LiDS conference in Pittsburgh. I have very much enjoyed working with you, Richard, and have benefited greatly from your enthusiastic and thoughtful leadership, which also helped me ease into my role as the Chair in 2019. Special thanks also go to Guoqing Diao for his excellent service as Program Chair. He ensured LiDS had a high profile at the 2019 Joint Statistical Meetings (JSM) through sponsorship of several sessions. He also chaired the Student Paper Award competition for the 2019 LiDS conference and is chairing the committee for the newly established student paper award for JSM 2020. Guoqing, your dedication and willingness to take things on are much appreciated! I also want to give great appreciation to our Web Master Ker-Ai Lee. Ker-Ai has done a wonderful job in transitioning and maintaining the LiDS website. Her speed in getting things done is just amazing! Lastly, congratulations and sincere thanks to Mimi Kim and Haoda Fu as 2020 Chair-Elect and 2020 Program Chair-Elect, respectively. I am also grateful to Chiung-Yu who has agreed to serve an additional year as Treasurer in 2020.

In May 2019, LiDS held its second conference at the University of Pittsburgh. The conference was another huge success, with 50 exciting invited sessions and over 230 registered participants. Hearty thanks to the local organizing committee (co-chaired by Ying Ding and Yu Chen) at University of Pittsburgh. The conference had a budget surplus which will be used for the 2021 LiDS conference. It also allows us to establish our Section's student paper awards, offer future webinars, and provide other activities to benefit members. This is a luxury for a new Section and we are grateful. LiDS also sponsored/co-sponsored two invited sessions and five topic-contributed sessions at the 2019 JSM in Denver, and served as a co-sponsor for the 2019 Annual Meeting of the Workshop on Biostatistics and Bioinformatics in Atlanta.

In September of 2019, a webinar survey was launched to collect members' opinion on webinars and topics of interest. Of the 85 members responded to the survey, over 90% of the respondents were at least moderately interested (33% extremely interested, 41% very interested, 16% moderately interested, 6% slightly interested). The top five topics of interest for the webinars are: (a) Machine learning and high dimensional survival data with application in genomics studies, imaging data, etc, (b) Causal inference, mediation analysis, and precision medicine with survival data, (c) Risk assessment and dynamic prediction for survival data, (d) Joint modeling of survival and longitudinal data, and (e) Competing risks. Based on this feedback, the Executive Committee has decided to offer webinars. More information will be communicated as it develops. I want to take this opportunity to thank Shanshan Zhao for helping prepare the very informative webinar survey. Thank you, Shanshan! This past year, the Executive Committee also established a Student Paper Award associated with the annual JSM. The Student Paper Award Committee is chaired by Guoqing Diao with Zhezhen Jin, Lu Mao, Tony Sit, Yifei Sun, and Ronghui Xu as committee members. The Committee is currently very busy reviewing the papers. Thank you Guoqing, Zhezhen, Lu, Tony, Yifei, and Ronghui for all your hard work!

The Nomination Committee for the 2020 Election was led by Richard Cook (Past Chair 2019) and included Yu Shen (Past Program Chair 2019) and Ying Ding as committee members. The Nomination Committee did a wonderful job and we are very fortunate to have a very strong slate of candidates for all the open positions. I encourage everyone to participate in the election, which will be held in the spring of this year.

As the Past-Chair, I will be responsible for organizing the slate of candidates for the 2021 elections. A nominating committee of three members will identify candidates for election of Chair-Elect, Secretary, Program Chair, and Council of Sections Representative for 2022. If you have ideal candidates in mind or would like to be nominated for any of these positions, please let me know and I will bring these names forward to the nominating committee for discussion.

We have an exciting year in 2020 ahead of us and we are in the great hands of Nick Jewell as the 2020 Chair. It has been a great pleasure to work with you in the past year, Nick, and I wish you the best in 2020!

I look forward to meeting more members this year and wish all LiDS members a very happy and healthy new year in 2020!

Jianwen Cai, Chair 2019

2020 Election Candidates

This year we have a number of positions open on the Executive Committee of the Section on Lifetime Data Science including Chair Elect, Treasurer, and Program Chair-Elect. We are in the

Figure 1: Participants of the annual LiDS business meeting at JSM 2019 in Denver, Colorado.

fortunate position of having a very strong slate of candidates who have agreed to stand for election. They are as follows:

Chair-Elect
Douglas Schaubel, University of Pennsylvania
Tony (Jianguo) Sun, University of Missouri

Treasurer
Adin-Cristian Andrei, Northwestern University
Yu Cheng, University of Pittsburgh

Program Chair-Elect
Kevin He, University of Michigan
Jing Ning, University of Texas, MD Anderson Cancer Center

As the Section of Lifetime Data Science (LiDS) is now officially part of the American Statistical Association (ASA), the ASA will handle the election process and so we can expect to hear more about the candidates including their biographies and personal statements in the coming weeks. On behalf of the LiDS Nomination Committee comprised of Ying Ding and Yu Shen and myself, sincere thanks to the membership for participating in the nomination process, and to the candidates for their commitment to the section and willingness to consider these leadership positions.

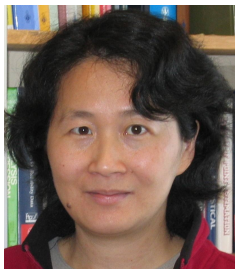
Richard Cook, Chair, 2019 LiDS Nomination Committee

New Section Officers

Chair-Elect 2020 Dr. Mimi Kim is a Professor in the Department of Epidemiology and Population Health at the Albert Einstein College of Medicine and has been Head of the Division of Biostatistics since 2003. She also directs the Biostatistics Shared Resource of the Institute of Clinical and Translational Research, and the Center for Quantitative Sciences. She is a Fellow of the American Statistical Association, on the Board of Trustees of the National Institute of Statistical Sciences, was Vice Chair of the American Statistical Association Council of Chapters Governing Board, and past President of the Korean International Statistical Society. She has participated on numerous grant review panels for the National Institutes of Health including the Epidemiology of Cancer Study Section and the Arthritis, Musculoskeletal and Skin Diseases Clinical Trials Review Committee. Her research interests include multivariate and interval-censored survival data, the design and analysis of clinical trials, and epidemiologic methods. Her publications have appeared in Biometrics, Lifetime Data Analysis, Statistics in Medicine, NEJM, and other statistics and medical journals.

Program Chair-Elect 2020 Dr. Haoda Fu is a Research Fellow and an Enterprise Lead for Machine Learning, Artificial Intelligence, and Digital Connected Care from Eli Lilly and Company. Dr. Haoda Fu is a Fellow of ASA (American Statistical Association). He is also an adjunct professor of biostatistics department, Indiana university school of medicine. Dr. Fu received his Ph.D. in statistics from University of Wisconsin - Madison in 2007 and joined Lilly after that. Since he joined Lilly, he is very active in statistics methodology research. He has more than 90 publications in the areas, such as Bayesian adaptive design, survival analysis, recurrent event modeling, personalized medicine, indirect and mixed treatment comparison, joint modeling, Bayesian decision making, and rare events analysis. In recent years, his research area focuses on machine learning and artificial intelligence. His research has been published in various top journals including JASA, JRSS, Biometrics, ACM, IEEE, JAMA, Annals of Internal Medicine etc.. He has been teaching topics of machine learning and AI in large industry conferences including teaching this topic in FDA workshop. He was board of directors for statistics organizations and program chairs, committee chairs such as ICSA, ENAR, and ASA Biopharm session.

Report from the Section Secretary



Starting as an interest group (LiDA-IG) of the ASA in 2014, we became the ASA-LiDS section in 2018 and has now 375 section members. Our official webpage is maintained by Ker-Ai Lee (ka2lee@uwaterloo.ca): its link is `https://community.amstat.org/lids/home`.

Over 40 people attended the section's 2019 annual meeting chaired by Section Chair Jianwen Cai at JSM 2019, Colorado Convention Center, on July 30, 2019. Matthew Amboy, Senior Editor of Springer, presented Springer top-cited and top-downloaded papers awards for articles from the Lifetime Data Analysis. Two section mem-

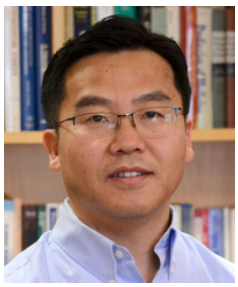
bers, A. Tsodikov and M. Yu, and two members of LiDA-IG, H. Fu and B. Li, were among the newly elected ASA fellows. Other highlights of our section's participation in JSM 2019 included the section sponsored invited session "Statistical Methods for Composite Time-to-Event Endpoints" and the two topic-contributed sessions "Recent Advances in Lifetime Data Analysis" and "Estimand Framework and its Impact on Drug Development in Oncology".

The 2019 LiDS Conference took place at University of Pittsburgh May 29-–31 2019, with its Scientific Program Committee co-chaired by Richard Cook, the section's Past Chair 2019, and Jianwen Cai. The conference had 50 parallel sessions, three short courses, and three keynote presentations. It attracted 230 attendants from 10 countries. Eight out of over 30 participants in the competition won the LiDS 2019 Student Paper Awards; three, the LiDS 2019 Student Poster Awards.

The section has selected to sponsor two invited sessions and two topic-contributed sessions at JSM 2020 from the received proposals. The preparation for the next section's conference, LiDS 2021, has started off. There are/will be a lot going on with the section. We welcome you to become a section member, to participate in the section's various activities, and to serve on the section associated committees in 2020.

Joan Hu, Secretary 2019–2021

JSM 2020 Program Update



The 2020 Joint Statistical Meetings (JSM) will be held in Philadelphia, Pennsylvania during August 1–6, 2020. For the 2020 JSM, our section has allocations of one invited session and two topic-contributed sessions. We have received seven invited session proposals: three with LiDS being the primary sponsor and four have LiDS as the second or third sponsor. Due to the excellence of the proposals, our section was able to win an open-competing invited session, so we will be the primary sponsor for two invited sessions:

+ (Organized by Dr. Li-Shan Huang) "Current and Future Challenges in Analyzing Composite Endpoints";
+ (Organized by Dr. Jialiang Li) "Survival Analysis in Causal Inference Studies".

We have also received five excellent topic-contributed session proposals as the primary sponsor. The two LiDS selected topic-contributed sessions are:

+ (Organized by Dr. Tony Sit) "Statistical Methods for New Challenges in Lifetime/Complex Data";
+ (Organized by Dr. Limin Peng) "Advanced Survival Analysis Tools for Statistical Learning from Complex Scientific Studies".

One proposal, became a topic-contributed session with WNAR being its primary sponsor:

+ "Novel methods for high-dimensional and large-scale survival data" organized by Dr. Ludovic Trinquart.

Another proposal became a topic-contributed session with Section on Statistics in Marketing being its primary sponsor:

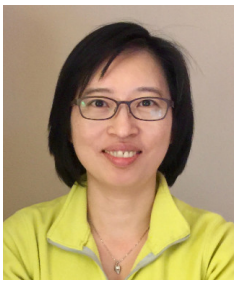
+ "Causal inference when the outcome is truncated by death" organized by Dr. Jessica G. Young.

I would like to thank all the organizers for their support and participation.

For the 2020 JSM, the abstract submission was open on December 3, 2019 and will close on February 4, 2020, and a draft manuscript must be received by May 15, 2020. It is imperative that all abstracts be submitted by February 4, 2020.

Zhezhen Jin, Program Chair 2020

Treasurer's Report



The beginning balance of the LiDS Section account on July 1, 2019 was $43,802.88. The total income during the period between July 1 and December 31 was $863.61, which included membership dues of $204. The expenses incurred ($2543.95) during the same period were mainly for the business and general meetings held at JSM. Then ending balance on December 31, 2019 was $42,122.44.

Beginning Balance	**6/30/2019**	**$43,802.88**
Income:		
Membership (6/2/2019 to 12/31/2019)		$204.00
Interest (6/1/2019 to 10/31/2019)		$659.51
Total Income		**$863.51**
Expense:		
Photocopying (newsletter distributed at JSM)		$93.88
2019 LiDS conference adjustment		
bank fee for wire transfer		$31.94
travel reimbursement		($16.05)
Food service at JSM		$2,434.18
Total Expense		**$2543.95**
Net Total Income		**($1680.44)**
Ending Balance	**12/31/2019**	**$42,122.44**

Chiung-Yu Huang, Treasurer 2016–2020

2021 Conference on Lifetime Data Science in Raleigh, North Carolina

The third biennial LiDS conference on Lifetime Data Science will be held from May 19-21, 2021 in Raleigh, North Carolina. Please mark your calendars! The location will likely be at the Marriott Raleigh City Center although this remains to be confirmed. The Program Committee will be co-chaired by Nick Jewell and Mimi Kim. The local organizing committee will be co-chaired by Shanshan Zhao of NIH/NIEHS and Wenbin Lu of North Carolina State University. We are currently beginning to select organizers for the scientific sessions and to choose potential short courses–please email Nick or Mimi any suggestions.

Nicholas P. Jewell and *Mimi Y. Kim*
Co-chairs of Program Committee
2021 LiDS Conference

News from Lifetime Data Analysis

Lifetime Data Analysis is the only journal dedicated to statistical methods and applications for lifetime data. The journal advances and promotes statistical science in various applied fields that deal with lifetime data, including actuarial science, economics, engineering, environmental sciences, management, medicine, operations research, public health, and social and behavioral sciences. The journal can be accessed at `https://link.springer.com/journal/10985`.

Every year, Springer summarizes a publisher's report for the articles published in *Lifetime Data Analysis*. From the report of 2019, the top-cited articles published 2016–2017 during the year of 2018 were:

1. "Longevity and concentration in survival times: the log-scale-location family of failure time models" by Chiara Gigliarano, Ugofilippo Basellini, and Marco Bonetti. Lifetime Data Analysis 23.2 (April 2017)
2. "Analysis of two-phase sampling data with semiparametric additive hazards models" by Yanqing Sun, Xiyuan Qian, Qiong Shou, and Peter B. Gilbert. Lifetime Data Analysis 23.3 (July 2017)
3. "Nonparametric and semiparametric regression estimation for length-biased survival data" by Yu Shen, Jing Ning, and Jing Qin. Lifetime Data Analysis 23.1 (January 2017)

The top-downloaded articles published in 2018 were:

1. "The wild bootstrap for multivariate Nelson–Aalen estimators" by Tobias Bluhmki, Dennis Dobler, Jan Beyersmann, and Markus Pauly. Lifetime Data Analysis 25.1 (January 2019)
2. "The effect of omitted covariates in marginal and partially conditional recurrent event analyses" by Yujie Zhong and Richard J. Cook. Lifetime Data Analysis 25.1 (January 2019)
3. "Survival models and health sequences" by Walter Dempsey and Peter McCullagh. Lifetime Data Analysis 24.4 (October 2018)

The January 2020 issue (Volume 26, number 1) of *Lifetime Data Analysis* has been published:

- Estimation for an accelerated failure time model with intermediate states as auxiliary information *by* Ritesh Ramchandani, Dianne M. Finkelstein, David A. Schoenfeld. Pages 1-20
- Nested exposure case-control sampling: a sampling scheme to analyze rare time-dependent exposures *by* Jan Feifel, Madlen Gebauer, Martin Schumacher, Jan Beyersmann. Pages 21-44
- Confidence intervals for the cumulative incidence function via constrained NPMLE *by* Paul Blanche. Pages 45-64
- Robust estimation for panel count data with informative observation times and censoring times *by* Hangjin Jiang, Wen Su, Xingqiu Zhao. Pages 65-84
- Semiparametric inference for a two-stage outcome-dependent sampling design with interval-censored failure time data *by* Qingning Zhou, Jianwen Cai, Haibo Zhou, Pages 85-108
- Frailty modelling approaches for semi-competing risks data *by* Il Do Ha, Liming Xiang, Mengjiao Peng, Jong-Hyeon Jeong, Youngjo Lee. Pages 109-133
- Multiplicative rates model for recurrent events in case-cohort studies *by* Poulami Maitra, Leila D. A. F. Amorim, Jianwen Cai. Pages 134-157
- An extended proportional hazards model for interval-censored data subject to instantaneous failures *by* Prabhashi W. Withana Gamage, Monica Chaudari, Christopher S. McMahanE-mail, Edwin H. KimMichael R. Kosorok. Pages 158-182
- Function-based hypothesis testing in censored two-sample location-scale models *by* Sundarraman Subramanian. Pages 183-213
- Correction to: Improved precision in the analysis of randomized trials with survival outcomes, without assuming proportional hazards *by* Iván Díaz, Elizabeth Colantuoni, Daniel F. Hanley, Michael Rosenblum. Pages 214-220

The October 2019 Issue (Volume 25, number 4) was a special issue dedicated to Odd O. Aalen, edited by Ørnulf Borgan and Håkon K. Gjessing:

- Special issue dedicated to Odd O. Aalen *by* Ørnulf Borgan, Håkon K. Gjessing. Pages 587-592
- Defining causal mediation with a longitudinal mediator and a survival outcome *by* Vanessa Didelez. Pages 593-610
- The additive hazard estimator is consistent for continuous-time marginal structural models *by* Pål C. Ryalen, Mats J. Stensrud, Kjetil Røysland. Pages 611-638
- A causal proportional hazards estimator under homogeneous or heterogeneous selection in an IV setting *by* Ditte Nørbo Sørensen, Torben Martinussen, Eric Tchetgen Tchetgen. Pages 639-659
- Landmark estimation of transition probabilities in non-Markov multi-state models with covariates *by* Rune Hoff, Hein Putter, Ingrid Sivesind Mehlum, Jon Michael Gran. Pages 660-680
- Modeling marginal features in studies of recurrent events in the presence of a terminal event *by* Per Kragh Andersen, Jules Angst, Henrik Ravn. Pages 681-695
- Partially hidden multi-state modelling of a prolonged disease state defined by a composite outcome *by* Vernon T. Farewell, Li Su, Christopher Jackson. Pages 696-711
- Prevalent cohort studies and unobserved heterogeneity *by* Niels Keiding, Katrine Lykke Albertsen, Helene Charlotte Rytgaard, Anne Lyngholm Sørensen. Pages 712-738
- A dual frailty model for lifetime analysis in maritime transportation *by* Robin Henderson, Ralitsa Mihaylova, Paul Oman. Pages 739-756
- Extending Bayesian back-calculation to estimate age and time specific HIV incidence *by* Francesco Brizzi, Paul J. Birrell, Martyn T. Plummer, Peter Kirwan, Alison E. Brown, Valerie C. Delpech, O. Noel Gill, Daniela De Angelis. Pages 757-780

Mei-Ling Ting Lee, Editor-in-Chief, Lifetime Data Analysis

Software Review

Small Sample Inferences with Kaplan–Meier Estimates with R Package bpcp

The **bpcp** R package is designed for making inferences on the survival function for right censored data with small samples or heavy censoring. The package addresses inferences such as exact central pointwise confidence intervals for the survival function at a fixed time point t, and exact differences (or ratios) between the two survival functions at t in the two-sample case. Despite being designed for small samples, samples with thousands of observations may be analyzed in less than a second.

Fay, Brittain, and Proschan (2013) addressed the first problem, creating a $100(1-\alpha)\%$ exact central confidence interval for $S(t)$ for a fixed t, meaning an interval (L, U) such that $\Pr[L \leq S(t)] \geq 1-\alpha/2$ and $\Pr[U \geq S(t)] \geq 1-\alpha/2$. They developed the beta product confidence procedure (BPCP) and showed that it gives an exact central pointwise confidence interval for $S(t)$ when there is no censoring (a case where the BPCP interval reduces to the usual exact central interval for a binomial observation), or when the censoring is progressive type II censoring. When the censoring is independent, then they provided only an *ad hoc* justification with simulations to argue for its exact centrality. The development repeatedly uses the probability integral transformation, with adjustments for grouped or tied data. Fay and Brittain (2016) gave additional adjustments to ensure that the Kaplan–Meier estimator was inside the BPCP interval and the intervals were monotonic in time. They also developed a mid-p version of the BPCP. Confidence interals for discrete data are often conservative for most parameter values to ensure that they cover the true parameter for all parameter values. In contrast, a mid-p version allows for undercoverage for some parameter values and overcoverage for others, so that on average over the parameter space the coverage is closer to the nominal level.

Fay, Proschan, and Brittain (2015) addressed the second problem: exact central confidence intervals for $S_2(t) - S_1(t)$ or $S_2(t)/S_1(t)$, where $S_1(t)$ and $S_2(t)$ are the two survival functions from the two-sample problem. They developed the melded method, that takes an exact central confidence interval procedure on a parameter from each of two groups (e.g., $S_1(t)$ and $S_2(t)$), and melds them together to create confidence intervals on functions of the two parameters. Fay, Proschan, and Brittain (2015) conjectured that those melded confidence intervals are exact and central, and supported the conjecture with some theoretical and simulation arguments. When there is no censoring, and the melding is applied to two BPCPs for $S_1(t)$ and $S_2(t)$, then we get confidence intervals for $S_2(t) - S_1(t)$ or $S_2(t)/S_1(t)$ that are compatible with the central Fisher's exact test. Here compatibility means that the central Fisher's exact test rejects at level α if and only if the $100(1-\alpha)\%$ confidence interval for $S_2(t) - S_1(t)$ excludes 0 (or the $100(1-\alpha)\%$ confidence interval for $S_2(t)/S_1(t)$ excludes 1).

The exact centrality of the BPCP for $S(t)$ and the melded confidence intervals for $S_2(t) - S_1(t)$ and $S_2(t)/S_1(t)$ have not been rigorously proven, although both have partial theoretical justification. These are open problems where more rigorous proofs of these conjectures or counterexamples are needed. Cui and Hannig (2019) derived the BPCP using generalized fiducial methods, but their results are asymptotic and still leave the small sample problems open.

For applications, however, these methods are better than any current method in terms of guaranteeing coverage regardless of sample size. The **bpcp** uses methods that do not rely on asymptotics, while other current methods do (e.g., using counting processes [see e.g., Andersen, *et al* 1993], empirical likelihoods [see e.g., Owen, 2001], or generalized fiducial methods [see Cui and Hannig, 2019]) or rely on other adjustments for modifying the lower interval due to censoring (see **survival** R package, version 3.1-7).

From version 1.4 of **bpcp**, we use a function structure similar to the **survival** package. For an example, we consider the acute leukemia study of Freireich *et al* (1963), and as in Gehan (1965), we treat it as a two sample trial for illustration. There are $n = 21$ in the placebo group and $n = 21$ in the treated (6-MP) group. The endpoint is time until relapse in weeks, with censoring. Here is the code to calculate two Kaplan–Meier curves, each with associated 95% pointwise confidence intervals on S and print out the first few intervals of the placebo group.

```
library(bpcp)
data(leuk2)
leuk2fit <- bpcpfit(Surv(time, status) ~ treatment,
                    data=leuk2)
head(summary(leuk2fit[["placebo"]]))
```

```
##   time interval survival lower 95% CL upper 95% CL
## 1         [0,1)   1.0000       0.8389       1.0000
## 2         [1,2)   0.9048       0.6962       0.9883
## 3         [2,3)   0.8095       0.5809       0.9455
## 4         [3,4)   0.7619       0.5283       0.9178
## 5         [4,5)   0.6667       0.4303       0.8541
## 6         [5,8)   0.5714       0.3402       0.7818
```

The print method for output from the `bpcpfit()` function is modeled after `survfit()` in the **survival** package and gives the number of events, median time to event and associated 95% confidence intervals, which are calculated using BPCP methods (see Fay, Brittain, and Proschan, 2013).

```
leuk2fit
```

```
##           n events median 0.95LCL 0.95UCL
## 6-MP     21      9     23      11     Inf
## placebo 21     21      8       4      12
```

Other quantiles within each of the two treatment groups can be obtained by the `quantile()` method:

```
quantile(leuk2fit, probs=c(.25, .5, .75))
```

```
## $`6-MP`
##      S(q)  q lower upper
## [1,] 0.25 NA    22   Inf
## [2,] 0.50 23    11   Inf
## [3,] 0.75 13     6    23
##
## $placebo
##      S(q)  q lower upper
## [1,] 0.25 12     8    22
## [2,] 0.50  8     4    12
## [3,] 0.75  4     1     8
```

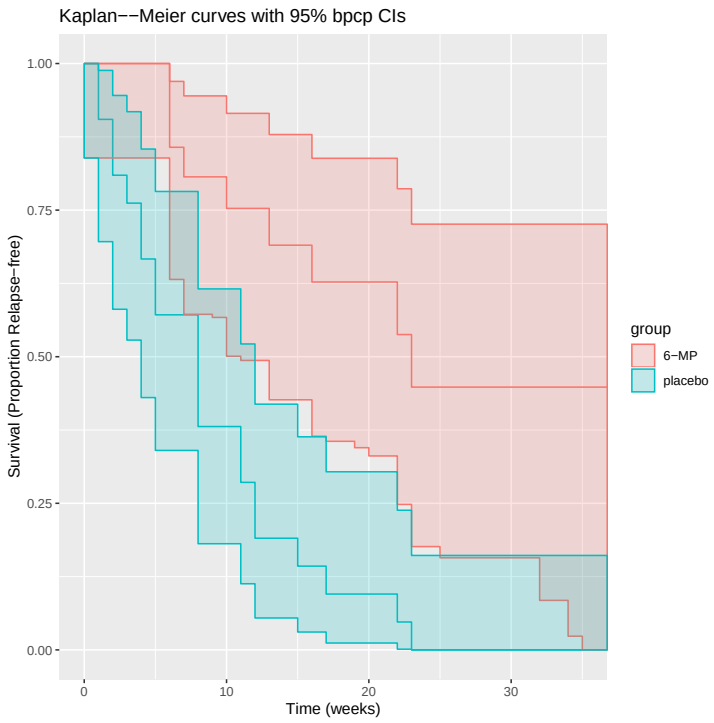
In this case, for the placebo group, because the Kaplan–Meier estimator never goes below 0.25, the 25% percentile estimator is set to missing.

The survival curves can be plotted with the `plot()` method function:

```
plot(leuk2fit)
```

The vanilla plot can be customized using `bpcp::tidykmciLR()` and the **ggplot2** package,

```
## first create a tidy object to customize ggplot
leuk2tidy <- tidykmciLR(leuk2fit)
ggplot(leuk2tidy,
       aes(x = time, y = surv, ymin = lower,
           ymax = upper, col = group)) +
    geom_line(show.legend=FALSE) +
    geom_ribbon(alpha = .2, aes(fill = group)) +
    xlab("Time (weeks)") +
    ylab("Survival (Proportion Relapse-free)") +
ggtitle("Kaplan--Meier curves with 95% bpcp CIs")
```



To calculate a confidence interval for the difference between the survival functions at 20 weeks, use:

```
## relevel treatment to make placebo the reference
leuk2$treatment <- relevel(leuk2$treatment,
                           ref = "placebo")
with(leuk2,
     bpcp2samp(time,status,treatment,
               testtime=20, parmtype="difference"))

##
##  Two-Sample Melded BPCP Test
##
## data:  S(20;group=6-MP)-S(20;group=placebo)
## S(20;group=placebo) = 0.095, S(20;group=6-MP)
## = 0.63, p-value = 0.003
## alternative hypothesis: true difference is not equal to 0
## 95 percent confidence interval:
##  0.1487 0.7751
## sample estimates:
## difference
##     0.5322
```

Calling the same function except with `parmtype="ratio"` gives the same p-value, but an estimate of the ratio $S_2(20)/S_1(20)$ equal to 6.6, with 95% confidence interval: (1.6, 55.7).

References

Andersen, P., Borgan, O., Gill, R., and Keiding, N. (1993), Statistical Models Based on Counting Processes, New York: Springer.

Cui, Y. and Hannig, J. (2019),"Nonparametric generalized fiducial inference for survival functions under censoring," Biometrika, 106, 501–518.

Fay, M. P., Brittain, E. H., and Proschan, M. A. (2013), "Pointwise Confidence Intervals for a Survival Distribution for Right Censored Data with Small Samples or Heavy Censoring," Biostatistics, 14, 723–736.

Fay, M. P. and Brittain, E. H. (2016), "Finite sample pointwise confidence intervals for a survival distribution with right-censored data," Statistics in Medicine, 35, 2726–2740

Fay, M. P., Proschan, M. A., and Brittain, E. (2015), "Combining one-sample confidence procedures for inference in the two-sample case," Biometrics, 146–156.

Freireich, E. J., Gehan, E., Frei, E., Schroeder, L. R., Wolman, I. J., Anbari, R., Burgert, E. O., Mills, S. D., Pinkel, D., Selawry, O. S., et al. (1963), "The effect of 6-mercaptopurine on the duration of steroid-induced remissions in acute leukemia: A model for evaluation of other potentially useful therapy," Blood, 21, 699–716.

Gehan, E. A. (1965), "A generalized Wilcoxon test for comparing arbitrarily singly-censored samples," Biometrika, 52, 203–224.

Owen, A. B. (2001), Empirical likelihood, Chapman and Hall / CRC.

Michael P. Fay
Mathematical Statistician
Biostatistics Research Branch
National Institute of Allergy and Infectious Diseases
Email: `mfay@niaid.nih.gov`

Regularized Cox Cure Rate Model with R package intsurv

Regularized Cox models with a cure rate are an important tool for analyzing survival data with heaving censoring and a large number of covaraites. The R Package **intsurv** (Wang et al., 2019) provides a collection of methods for integrative survival analyses with data from multiple sources. Function `cox_cure_net.fit()` in the package is an efficient implementation for regularized Cox cure rate model with elastic-net penalty (Zou and Hastie, 2005).

The cure rate models first proposed by Berkson and Gage (1952) are commonly adopted statistical methods for survival data with a cured fraction. Consider a random sample of n subjects with right-censoring data and a cured fraction. Let $T_j = \min(V_j, C_j)$ and $\Delta_j = I(V_j > C_j)$, where V_j and C_j represents the random variable of the event time and the censoring time of subject j, respectively, $I(\cdot)$ is indicator function, $j \in \{1, \ldots, n\}$. Define $Z_j = 1$ if subject j is susceptible, and $Z_j = 0$ otherwise, with probability $p_j = \Pr(Z_j = 1)$. Notice that Z_j is observed to be 1 if $\Delta_j = 1$ and is missing otherwise. Proposed by Farewell (1982), a logistic model $p_j = 1/[1+\exp(-\gamma_0 - \boldsymbol{x}_j^\top \boldsymbol{\gamma})]$ is widely used, where $\boldsymbol{x}_j$ represents the covariate vector of subject j (excluding intercept), γ_0 is unknown coefficient of intercept and $\boldsymbol{\gamma}$ is a vector of unknown covariate coefficients. Given that $Z_j = 1$, Kuk and Chen (1992) proposed modeling the conditional survival times through a Cox proportional hazard model with the hazard function

$$h_j(t \mid Z_j = 1) = h_0(t \mid Z_j = 1) \exp(\boldsymbol{x}_j^\top \boldsymbol{\beta}),$$

where $h_0(t \mid Z_j = 1)$ is an unspecified baseline function for events, and $\boldsymbol{\beta}$ is a vector of unknown coefficients of the covariate vector $\boldsymbol{x}_j$. The conditional survival function of the event time of subject j is

$$S_j(t \mid Z_j = 1) = \exp\{-H_0(t \mid Z_j = 1) \exp(\boldsymbol{x}_j^\top \boldsymbol{\beta})\},$$

where $H_0(t \mid Z_j = 1) = \int_0^t h_0(s \mid Z_j = 1) \mathrm{d}s$. Given that subject j is cured ($Z_j = 0$), the conditional survival function satisfies $S_j(t \mid Z_j = 0) = 1$, for $t < +\infty$. The observed data likelihood function can be written as

$$L(\boldsymbol{\theta}) = \prod_{j=1}^{n} \{p_j h_j(t_j \mid Z_j = 1) S_j(t_j \mid Z_j = 1)\}^{\delta_j} \{(1 - p_j) + p_j S_j(t_j \mid Z_j = 1)\}^{1-\delta_j}, \quad (1)$$

where $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \boldsymbol{\gamma}, \gamma_0, h_0(\cdot)\}$.

An estimation procedure based on the well-known EM algorithm was proposed by Sy and Taylor (2000). Recently, a few works have been proposed to perform variable selection for cure models. For example, Scolas et al. (2016) proposed variable selection with adaptive lasso penalty (Zou, 2006) for interval-censored data in a parametric cure model, where conditional survival times follow the extended generalized gamma distribution. Masud et al. (2018) proposed variable selection methods for mixture cure model and promotion cure model through regularization by the adaptive lasso penalty. Fan et al. (2017) and Shi et al. (2019) promoted structural similarity and sign consistency of $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\beta}}$, respectively, with minimax concave penalty (Zhang, 2010) for variable selection. Here, we concentrate on the following regularized estimator with elastic-net penalty,

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} -\frac{1}{n}\ell(\boldsymbol{\theta}) + P_1(\boldsymbol{\beta}; \alpha_1, \lambda_1) + P_2(\boldsymbol{\gamma}; \alpha_2, \lambda_2), \quad (2)$$

where $\ell(\boldsymbol{\theta})$ is the log-likelihood function under the observed data from (1) and

$$P_1(\boldsymbol{\beta}; \alpha_1, \lambda_1) = \lambda_1 \left(\alpha_1 \sum_{k=1}^{p} \omega_k |\beta_k| + \frac{1 - \alpha_1}{2} \sum_{k=1}^{p} \beta_k^2 \right),$$
$$P_2(\boldsymbol{\gamma}; \alpha_2, \lambda_2) = \lambda_2 \left(\alpha_2 \sum_{k=1}^{p} \nu_k |\gamma_k| + \frac{1 - \alpha_2}{2} \sum_{k=1}^{p} \gamma_k^2 \right),$$

where ω_k and ν_k represent non-negative weights (Zou, 2006), $0 \le \alpha_1 \le 1$, $0 \le \alpha_2 \le 1$, $\lambda_1 \ge 0$, and $\lambda_2 \ge 0$ are tuning parameters. The coordinate descent algorithm (Friedman et al., 2007) or local quadratic approximations (Fan and Li, 2001) may be utilized in the M-steps of the EM algorithm to obtain the regularized estimator. Under the hood, `cox_cure_net.fit()` utilizes the coordinate-majorization-descent (CMD) algorithm proposed by Yang and Zou (2013) in the M-steps due to its descent property.

To demonstrate the usage of `cox_cure_net.fit()`, we may simulate a dataset of sample size 200 as follows. 100 covariates are simulated from multivariate normal distribution with means zero and variances one. The correlation between x_k and x_l, $k \neq l$, was set to be $\rho^{|k-l|}$, where $\rho = 0.5$. For each model part, only five covariates actually have non-zero coefficients. The true non-zero coefficients are simulated from Unif$(0.6, 1)$ independently. For susceptible subjects, the event times were generated from Weibull-Cox model with baseline hazard function $h_0(t; \boldsymbol{x}) = 0.2t \exp(\boldsymbol{x}^\top \boldsymbol{\beta})$. For cured subjects, the event times were set to be infinity. The censoring times were generated independently with the event times from exponential distribution with rate 0.01 and truncated at 10. The generation of event times and censoring times takes advantage of function `intsurv::simData4cure()`.

```
library(intsurv)
set.seed(123)
p <- 100; n <- 200; rho <- 0.5
beta0 <- gamma0 <- rep(0, p)
beta0[c(1, 2, 4, 6, 8)] <- runif(5, 0.6, 1)
gamma0[c(1, 3, 5, 7, 9)] <- runif(5, 0.6, 1)
ij_mat <- expand.grid(i = seq_len(p), j = seq_len(p))
Sigma <- matrix(mapply(function(i, j) {
    rho^abs(i - j)
}, ij_mat$i, ij_mat$j), nrow = p)
x_mat <- MASS::mvrnorm(n, mu = rep(0, p), Sigma)
colnames(x_mat) <- paste0("x", seq_len(p))
dat <- simData4cure(
    n, survMat = x_mat, survCoef = beta0,
    cureCoef = gamma0, b0 = 1, lambda_censor = 0.01,
    max_censor = 10, p1 = 1, p2 = 1, p3 = 1
)
```

Similar to function `glmnet::glmnet()` for regularized generalized linear models, `cox_cure_net.fit()` fits the regularized Cox cure rate model over a specified grid of tuning parameter λ_1 and λ_2 with fixed α_1 and α_2. Instead, the desired length of each λ sequence can be specified and an equally-spaced (in logarithm scale) sequence will be generated from the smallest

"large enough" $\lambda_{\max}$ that results in all zero coefficient estimates to a specified "small enough" $\lambda_{\min}$. By default, $\lambda_{\min} = 0.1\lambda_{\max}$ is set for both model parts in `cox_cure_net.fit()`. Here we set $\alpha_1 = \alpha_2 = 0.5$ and specify a 10 by 10 grid for λ_1 and λ_2.

```
system.time({
    fit1 <- cox_cure_net.fit(
        surv_x = x_mat, cure_x = x_mat,
        time = dat$obs_time, event = dat$obs_event,
        surv_nlambda = 10, cure_nlambda = 10,
        surv_alpha = 0.5, cure_alpha = 0.5
    )
})
```

```
##    user  system elapsed
##   5.437   0.006   5.455
```

The tuning parameters may be selected based on BIC and a `coef()` method for `cox_cure_net` objects can be used to return the coefficient estimates from the selected model. We may quickly check the true positive rate and false positive rate in terms of variable selection as follows:

```
eval_vs <- function(x, beta0, gamma0) {
    foo <- function(b, b0) {
        c("% True Positive" = mean(b[b0 != 0] != 0),
          "% False Positive" = mean(b[b0 == 0] != 0))
    }
    rbind(beta = foo(coef(fit1)$surv, beta0),
          gamma = foo(coef(fit1)$cure, gamma0))
}
eval_vs(fit1, beta0, gamma0)
```

```
##       % True Positive % False Positive
## beta        1.0000000       0.09473684
## gamma       0.8333333       0.07368421
```

To reduce computational burden, the generalized EM algorithm may be used by setting one-step CMD update as follows. In this example, we are able to substantially decrease the computation time and obtain the same variable selection results.

```
system.time({
    fit2 <- cox_cure_net.fit(
        surv_x = x_mat, cure_x = x_mat,
        time = dat$obs_time, event = dat$obs_event,
        surv_nlambda = 10, cure_nlambda = 10,
        surv_alpha = 0.5, cure_alpha = 0.5,
        surv_max_iter = 1, cure_max_iter = 1
    )
})
```

```
##    user  system elapsed
##   2.319   0.002   2.322
```

```
eval_vs(fit2, beta0, gamma0)
```

```
##       % True Positive % False Positive
## beta        1.0000000       0.09473684
## gamma       0.8333333       0.07368421
```

After variable selection, a regular Cox cure rate model may be fitted by `intsurv::cox_cure()`. See `https://wenjie-stat.me/intsurv/` for the full package documents.

References

Berkson, J. and Gage, R. P. (1952), "Survival Curve for Cancer Patients Following Treatment," *Journal of the American Statistical Association*, 47, 501–515.

Fan, J. and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American statistical Association*, 96, 1348–1360.

Fan, X., Liu, M., Fang, K., Huang, Y., and Ma, S. (2017), "Promoting Structural Effects of Covariates in the Cure Rate Model with Penalization," *Statistical Methods in Medical Research*, 26, 2078–2092.

Farewell, V. T. (1982), "The Use of Mixture Models for the Analysis of Survival Data with Long-Term Survivors," *Biometrics*, 1041–1046.

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007), "Pathwise Coordinate Optimization," *The Annals of Applied Statistics*, 1, 302–332.

Kuk, A. Y. C. and Chen, C.-H. (1992), "A Mixture Model Combining Logistic Regression with Proportional Hazards Regression," *Biometrika*, 79, 531–541.

Masud, A., Tu, W., and Yu, Z. (2018), "Variable Selection for Mixture and Promotion Time Cure Rate Models," *Statistical methods in medical research*, 27, 2185–2199.

Scolas, S., El Ghouch, A., Legrand, C., and Oulhaj, A. (2016), "Variable Selection in A Flexible Parametric Mixture Cure Model with Interval-Censored Data," *Statistics in Medicine*, 35, 1210–1225.

Shi, X., Ma, S., and Huang, Y. (2019), "Promoting Sign Consistency in the Cure Model Estimation and Selection," *Statistical methods in medical research*, 1–14.

Sy, J. P. and Taylor, J. M. G. (2000), "Estimation in a Cox Proportional Hazards Cure Model," *Biometrics*, 56, 227–236.

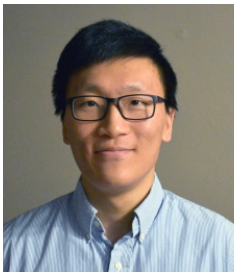
Wang, W., Chen, K., and Yan, J. (2019), *intsurv: Integrative Survival Models*, R package version 0.2.1.

Yang, Y. and Zou, H. (2013), "A Cocktail Algorithm for Solving the Elastic Net Penalized Cox's Regression in High Dimensions," *Statistics and its Interface*, 6, 167–173.

Zhang, C.-H. (2010), "Nearly Unbiased Variable Selection Under Minimax Concave Penalty," *The Annals of statistics*, 38, 894–942.

Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties," *Journal of the American statistical association*, 101, 1418–1429.

Zou, H. and Hastie, T. (2005), "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67, 301–320.



Wenjie Wang
Research Scientist
Machine Learning, Artificial Intelligence, and Connected Care
Advanced Analytics and Data Sciences
Eli Lilly and Company
Email: `wang_wenjie@lilly.com`