Wilstrup and Cave
BMC Medical Informatics and Decision Making (2022) 22:196
https://doi.org/10.1186/s12911-022-01943-1

BMC Medical Informatics and Decision Making

RESEARCH

Open Access

Combining symbolic regression with the Cox proportional hazards model improves prediction of heart failure deaths

Casper Wilstrup* and Chris Cave

Abstract

Background: Heart failure is a clinical syndrome characterised by a reduced ability of the heart to pump blood. Patients with heart failure have a high mortality rate, and physicians need reliable prognostic predictions to make informed decisions about the appropriate application of devices, transplantation, medications, and palliative care. In this study, we demonstrate that combining symbolic regression with the Cox proportional hazards model improves the ability to predict death due to heart failure compared to using the Cox proportional hazards model alone.

Methods: We used a newly invented symbolic regression method called the QLattice to analyse a data set of medical records for 299 Pakistani patients diagnosed with heart failure. The QLattice identified non-linear mathematical transformations of the available covariates, which we then used in a Cox model to predict survival.

Results: An exponential function of age, the inverse of ejection fraction, and the inverse of serum creatinine were identified as the best risk factors for predicting heart failure deaths. A Cox model fitted on these transformed covariates had improved predictive performance compared with a Cox model on the same covariates without mathematical transformations.

Conclusion: Symbolic regression is a way to find transformations of covariates from patients' medical records which can improve the performance of survival regression models. At the same time, these simple functions are intuitive and easy to apply in clinical settings. The direct interpretability of the simple forms may help researchers gain new insights into the actual causal pathways leading to deaths.

Keywords: Proportional hazards model, Symbolic regression, Qlattice, Machine learning, Cardiovascular heart diseases, Heart failure

Background

Heart failure (HF) is a clinical syndrome characterised by a reduction in the ability of the heart to pump or fill with blood. HF can be defined physiologically as an inadequate cardiac output to meet metabolic demands, often manifesting as increased left ventricular filling pressure [1]. Among the causes of HF are coronary heart disease, hypertension, diabetes, obesity, and smoking [2]. HF affects at least 26 million people globally and has a high mortality rate [3].

Various methods have been developed to estimate the risk of death for patients with HF. Well-known models include the ADHERE model [4] and the Seattle Heart Failure Model [5]. Although these models are accurate, they are unintuitive and rely on extensive medical records, making them hard to apply in a clinical setting.

Ahmad et al. published a study of 299 patients with HF admitted to Faisalabad Institute of Cardiology or Allied Hospital Faisalabad, Punjab, Pakistan [6]. In their study,

*Correspondence: casper.wilstrup@abzu.ai

Abzu, Orient Plads 1, 2150 Copenhagen, Denmark

Ahmad et al. used a Cox model to predict survival of patients using much fewer covariates than the Seattle Heart Failure Model. The data set was made freely available by Ahmad et al., and it has subsequently been used in additional analyses using both survival models [7] and machine learning techniques [8].

Cox proportional hazards model

The Cox model [9] is widely used to predict survival in health science. When using the method, the researcher needs to choose which risk factors to include in the model, and the final model performance is obviously dependent on the appropriate choice of risk factors. One such example of a Cox regression model to model HF is the MAGGIC model in [10]

The general form of the Cox model is [9]:

$$h(t) = h_0(t) \cdot e^{b_1 x_1 + b_2 x_2 + \ldots + b_p x_p}$$

where

- t is the survival time
- $h(t)$ is the hazard function determined by a set of p covariates $(x_1, x_2, \ldots, x_p)$
- $(b_1, b_2, \ldots, b_p)$ are the coefficients which measure the impact of the covariates
- the term $h_0(t)$ is the baseline hazard, i.e. the hazard if all the x_i are equal to zero

This can also be written as:

$$h(t) = h_0(t) \cdot e^{b_1 x_1} \cdot e^{b_2 x_2} \cdot \ldots \cdot e^{b_p x_p}$$

If we let $h(t)$ and $h'(t)$ be the hazard function for two individuals that differ only in x_1 with a difference δx, their relative hazard is:

$$\frac{h'(t)}{h(t)} = \frac{e^{b_1(x_1+\delta x)}}{e^{b_1 x_1}} = e^{b_1(x_1+\delta x - x_1)} = e^{b_1 \delta x}$$

This shows an important limitation of Cox models: A unit change in a covariate will have the same effect on the hazard regardless of the origin of the change. For example: a change in x_1 of 0.1 will have the same effect on survival whether from 1.0 to 1.1 or from 5.0 to 5.1.

Mathematical transformations

In this paper, the term *mathematical transformation*, sometimes just *transformation*, means to apply a mathematical function to a covariate: $x_T = f(x)$ for some function f. Perhaps the most widely used mathematical transformation is log-transformation: $x_T = log(x)$. But other typical transformations include x^2, $\sqrt{x}$, and e^x.

When fitting a Cox model using mathematically transformed covariates, we have:

$$\frac{h'(t)}{h(t)} = e^{b_1(f(x_1+\delta x) - f(x_1))}$$

which is clearly no longer independent of the origin of x_1 unless f is a linear function. Thus the effect of the transformation is to overcome this specific limitation in the generality of Cox models.

It is reasonable to hypothesise that well-chosen transformations of the available covariates will improve the predictive power of Cox models.

In this study, we demonstrate that this is indeed the case, and that a useful method to identify an appropriate mathematical transformation is to use symbolic regression.

Symbolic regression

Symbolic regression is a machine learning method that attempts to explain some Y in terms of some X using a mathematical expression composed of a set of basic functions. However, the search space of possible expressions grows exponentially with the length of the expression, which makes a direct search infeasible. Traditionally, genetic programming has been used to search this space selectively [11–13]. Recent approaches have been more physics-inspired [14].

The QLattice is a symbolic regressor inspired by quantum field theory [15]. The QLattice runs on a dedicated high-performance computing cluster and models the list of possible mathematical expressions—in principle: infinite—that could link Y to X as a superposition of an infinite set of spatial paths. The QLattice searches this list of all mathematical expressions, including parameters, for the expressions that best model the output from the input. The result of the search is a list of expressions sorted by how well they match observations.

The QLattice can be used to generate either classification models or regression models. In this study, we only use it to generate classification models. Mathematically, this means that the QLattice will wrap each expression in a logistic function, thereby allowing the output to be interpreted as a probability. In other words: if X is an input vector, $f(X)$ is the mathematical equation, and Y is the event we want to predict, then the QLattice will search for functions f such that the predictive power of:

$$\widehat{Y} = \frac{1}{1 + e^{-f(X)}}$$

is maximised.

In this study, we used symbolic regression to mathematically transform the available covariates from Ahmad

et al. We subsequently fitted a Cox model to these transformed values.

To our knowledge, the combination of symbolic regression and Cox models has not previously been studied in research.

Data

The data set used in this study consists of medical records of 299 HF patients admitted to the Faisalabad Institute of Cardiology and the Allied Hospital in Faisalabad (Punjab, Pakistan) between April and December 2015.

The data set contains 105 women and 194 men with left ventricular systolic dysfunction. The patients had HF in the classes III or IV of New York Heart Association (NYHA) classification scheme.

The patients were between 40 and 95 years old (mean 60.8) at the time of admission. The follow-up period was between 4 days and 285 days with a mean of 130 days.

The data set contains the following potential risk factors: age, serum sodium, serum creatinine, gender, smoking, blood pressure (BP), ejection fraction (EF), anaemia, platelets, and creatinine phosphokinase (CPK). Table 1 shows the summary statistics for the data.

Anemia in patients was assessed by their haematocrit level. Patients with haematocrit less than 36% (minimum normal level of haematocrit) were taken as anemic. Information on a patient's smoking status and high blood pressure were taken from physician's notes.

Methods

In this study we picked the top three significant features under the Mann–Whitney U test from [8] and then used the QLattice to find how these covariates can best be mathematically transformed to model risk of death. Finally, we fitted a Cox model on the transformed covariates.

The top three significant features according to the Mann–Whitney U test are: *ejection fraction*, *serum creatinine* and *age*. Interestingly typical factors contributing to HF such as gender, smoking and blood pressure are not significant in this test. See the discussion in Ahmad et al. [6] and references therein .

QLattice for finding the best transformation of covariates

For each of the selected covariates, we used the QLattice to find the mathematical transformation that best predicted death events given that covariate alone.

Cox regression has an important limitation: a unit change in a covariate will have same effect on the hazard regardless of the origin of change. Thus by exploring the possibility of transformations of covariates we can overcome this limitation and potentially improve the predictive power of Cox models.

Since this search only involved a single covariate, we limited the QLattice to work with the following unary function set: $\frac{1}{x}$, e^x, $\log(x)$, $\sqrt{x}$.

Table 1 Summary statistics for the data set

		Overall	Min	25%	50%	75%	Max
n		299					
TIME, mean (SD)		130.3 (77.6)	4	73	115	203	285
Event, n (%)	0	203 (67.9)					
	1	96 (32.1)					
Gender, n (%)	0	105 (35.1)					
	1	194 (64.9)					
Smoking, n (%)	0	203 (67.9)					
	1	96 (32.1)					
Diabetes, n (%)	0	174 (58.2)					
	1	125 (41.8)					
BP, n (%)	0	194 (64.9)					
	1	105 (35.1)					
Anaemia, n (%)	0	170 (56.9)					
	1	129 (43.1)					
Age, mean (SD)		60.8 (11.9)	40	51	60	70	95
EF, mean (SD)		38.1 (11.8)	14	30	38	45	80
Sodium, mean (SD)		136.6 (4.4)	113	134	137	140	148
Creatinine, mean (SD)		1.4 (1.0)	0.5	0.9	1.1	1.4	9.4
Pletelets, mean (SD)		263K (97K)	25K	212K	262K	303K	850K
CPK, mean (SD)		581.8 (970.3)	23	116.5	250	582	7861

Cox model

Finally, we fitted a Cox model with the output of the mathematically transformed covariates.

The resulting model was validated using concordance index (C-index). The C-index for a survival model can be thought of as the weighted average of the area under time-specific Receiver Operating Characteristic (ROC) curves. We also computed the time-specific ROC curve and area under the curve (AUC) for 285 day survival.

Comparison Cox model

For comparison, we fitted a Cox model on the same covariates but *without* the identified mathematical transformations. The two models used the same covariates and only differed in the transformations. The models were compared using C-index, AUC (285 days), log-likelihood, and Akaike Information Criterion (AIC) [16].

All calculations were done in the programming language Python. For symbolic regression, we used QLattice version 1.4.6 [15]. For Cox Modelling, we used lifelines version 0.25.7 [17].

Results

Best transformation of Ejection Fraction

The best mathematical relation between ejection fraction and the probability of death was:

$$f(E) = -2.1 + \frac{1.6}{0.066E - 0.89}$$

or, combined with the logistic wrapper:

$$P(death \mid E) = \frac{1}{1 + e^{2.1 - \frac{1.6}{0.066E - 0.89}}}$$

The probability of death was found to be closer associated with the inverse of the ejection fraction than with the ejection fraction directly.

Best transformation of serum creatinine

The best mathematical relation between serum creatinine and the probability of death was:

$$f(C) = 1.5 - \frac{0.91}{0.42C - 0.084}$$

or, combined with the logistic wrapper:

$$P(death \mid C) = \frac{1}{1 + e^{-1.5 + \frac{0.91}{0.42C - 0.084}}}$$

As with the ejection fraction, the probability of death is closer associated with the inverse of serum creatinine than to serum creatinine directly.

Best transformation of age

The best mathematical relation between age and the probability of death was:

$$f(C) = 0.02e^{0.056A} - 1.5$$

or, combined with the logistic wrapper:

$$P(death \mid A) = \frac{1}{1 + e^{-0.02e^{0.056A} + 1.5}}$$

The probability of death grows exponentially with age, with a growth constant of 0.056.

Kaplan Meier curves

In Fig. 1 we plot the Kaplan Meier curve on all the data and compare between Kaplan Meiers curves segmented

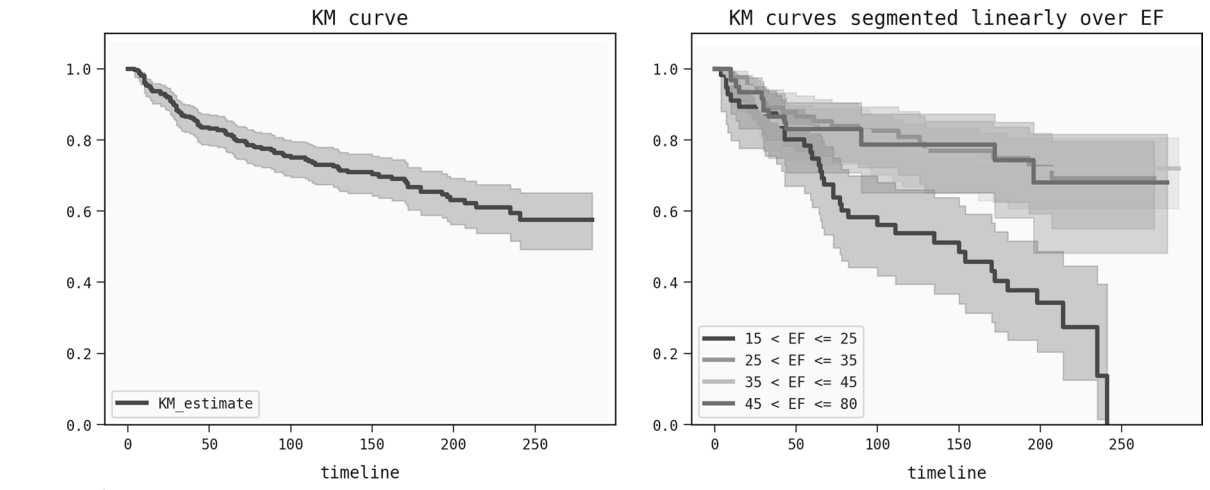

Fig. 1 Kaplan Meier curves

Table 2 Significance of mathematically transformed covariates in the Cox model

	coef	HR	HR lower 95%	HR upper 95%	z	*p*
Exp(0.056A)	0.014	1.014	1.009	1.019	5.851	< 0.0005
1/E	0.537	1.711	1.418	2.064	5.609	< 0.0005
1/C	-1.515	0.220	0.108	0.446	−4.198	< 0.0005

Table 3 Significance of covariates without transformation in the Cox model

	coef	HR	HR lower 95%	HR upper 95%	z	*p*
A	0.044	1.045	1.027	1.064	4.934	< 0.0005
E	-0.049	0.952	0.933	0.971	−4.803	< 0.0005
C	0.358	1.430	1.251	1.635	5.244	< 0.0005

linearly over the ejection fraction. The aim is to demonstrate the non-linear effect the ejection fraction has on survival. Below are the curves with 95% confidence interval bands.

Within linear increments between 25 and 80 of ejection fraction, the change on survival is limited. However within the range of $15 < EF \leq 25$ then ejection fraction has a larger impact on survival. This points towards a non-linear effect the ejection fraction has on survival.

Cox model with transformed risk factors

Having determined that $e^{0.056A}$, $1/E$, and $1/C$ are closer associated with risk of death than E, C, and A directly, we fitted a Cox model on these transformed covariates. Table 2 shows the coefficients (coef), hazard ratios (HR), and p-values for each of the risk factors. (Where A is the age, E is the ejection fraction, and C is serum creatinine.)

The log-likelihood of the model was found to be 83.8, and the C-index was 0.75. The model's discrimination ability was also tested with a ROC curve, and the AUC was found to be 0.82 at 285 days, meaning that the model could correctly predict death within 285 days for 82% of the patients.

Cox model with unmodified covariates

The comparison model was fitted on the same three covariates used above, but in untransformed form. Table 3 shows the coefficients (coef), hazard ratios (HR), and p-values for each of the risk factors for this comparison model.

The log-likelihood of the model was found to be 66.5, and the C-index was 0.72. The model's discrimination ability with AUC was found to be 0.78 at 285 days, meaning that the model could correctly predict death within 285 days for 78% of the patients.

Table 4 Comparison of performance metrics

	Transformed covariates	Untransformed covariates
C-index	0.75	0.72
AUC (285 days)	0.82	0.78
Log-likelihood	83.8	66.5
Partial AIC	941	958

Comparing the two models

Table 4 shows a comparison of all four performance metrics for the two Cox models on the three covariates (ejection fraction, serum creatinine, and age) with and without the mathematical transformations.

The mathematical transformations improved the Cox model on all metrics without adding additional covariates or any other information.

Comparison to model with all untransformed covariates

In the model used in Ahmad et al. [6], all covariates are used and have not been transformed. The AUC (250 days) of this model is 0.81 so we can see that the performance is very similar to the model with transformed covariates.

Discussion

Beyond the linear assumption in Cox models

Cox models are a powerful tool for survival regression. They are conceptually simple, easy to interpret, and can often accurately model the survival probability over time–even given censored data. These properties have made Cox models one of the most popular methods for survival modelling in health science.

As described above, a limitation of Cox models is that each risk factor is modelled to affect risk independent

of the origin. This study has shown than using mathematically transformed covariates can overcome this limitation.

For example: with the mathematical transformation $1/E$, the hazard ratio after a unit change in ejection fraction is:

$$\frac{h'(t)}{h(t)} = e^{b_1(f(E+\delta E)-f(E))} = e^{b_1(\frac{1}{E+\delta E}-\frac{1}{E})}$$

which is clearly no longer independent of the origin of E.

Interpretability

One of the advantages of symbolic regression over other machine learning methods is that the models are easier to interpret. This is true when the full model is expressed as a symbolic relation, and it is also true when a symbolic transformation is combined with another interpretable model such as the Cox model.

For example, the fact that the transformation $1/E$ is a better risk factor than E directly has at least two possible interpretations:

- One is simply that the lower the ejection fraction, the larger the additional hazard given an additional reduction, i.e. a change from 0.3 to 0.29 is more dangerous than a change from 0.6 to 0.59
- Another interpretation can be seen by observing that if a fraction E is pumped out of the heart in every heartbeat, the average number of heartbeats that a unit of blood stays in the heart before it is pumped out will be $1/E$, although certainly with large variation across the cavity of the ventricles.
  The mathematical transformation $1/E$ of ejection fraction means that the Cox model is based on a covariate which is the average number of heartbeats blood stays in the heart.

Conclusions

In this study, we demonstrated that combining symbolic regression with survival regression models improves the predictive power without sacrificing the interpretability that comes from a clear mathematical formulation of the model.

Other machine learning methods exist that may provide the same—or perhaps better—performance improvements than what can be achieved with the combination approach used here, but the price will typically be that the resulting model is black-box, or at least difficult to interpret.

In this study, we focused on predicting heart failure deaths, which is important in its own right, but the perspective of our research is that mathematical transformations based on symbolic regression may be applied to any survival model with improved predictive accuracy and new insights as a result.

Abbreviations

HF: Hearth failure; BP: Blood pressure; EF: Ejection fraction; CPK: Creatinine phosphokinase; C-index: Concordance index; AUC: Area under the curve; ROC: Receiver operating characteristic; AIC: Akaike information criterion; HR: Hazard ratio.

Acknowledgements

The authors would like to thank the following for their extensive help in reviewing the article and their many suggestions for clarifications and improvements: From Abzu: Jaan Kasak, Meera Machado, and Elyse Sims. From Aalborg University: Martin Siemienski Andersen. From Zealand University Hospital: Helene Søgaard Andersen.

Author contributions

CW designed the method and performed the analysis reported in this article. CC and CW co-authored the text of this article. All authors read and approved the final manuscript.

Funding

No external funding was received for this research project

Availability of data and materials

All data analysed in this study are included in the original published study by Ahmad et al. [6]. It can be directly accessed from https://doi.org/10.1371/journal.pone.0181001.s001 (CSV) Code to reproduce this study is publicly available under the BSD 3-Clause License at: https://github.com/wilstrup/sr-heartfailure The code needs access to a QLattice to run. Abzu grants access to QLattices for free for research purposes. If you need access, please contact the author or the company.

Declarations

Ethics approval and consent to participate

The original study containing the data used in this study was approved by the Institutional Review Board of Government College University (Faisalabad, Pakistan). It states that the Helsinki Declaration were followed [6]. The original study containing the data used in this study states that "Informed consent was taken by the patients from whom the information on required characteristics were collected/accessed." [6]

Consent for publication

Not applicable

Competing interests

CW and CC are employees of Abzu, which is the company developing the QLattice technology. CW is also co-founder of the company and the inventor of the patent-pending QLattice technology.

Received: 18 January 2021 Accepted: 20 July 2022
Published online: 25 July 2022

References

1. Tan LB, Williams SG, Tan DKH, Cohen-Solal A. So many definitions of heart failure: are they all universally valid? A critical appraisal. Expert Rev Cardiovasc Ther. 2010. https://doi.org/10.1586/erc.09.187.
2. Virani SS, Alonso A, Benjamin EJ, Bittencourt MS, Callaway CW, Carson AP, Chamberlain AM, Chang AR, Cheng S, Delling FN, Djousse L, Elkind MSV, Ferguson JF, Fornage M, Khan SS, Kissela BM, Knutson KL, Kwan TW, Lackland DT, Lewis TT, Lichtman JH, Longenecker CT, Loop MS, Lutsey PL, Martin SS, Matsushita K, Moran AE, Mussolino ME, Perak AM, Rosamond WD, Roth GA, Sampson UKA, Satou GM, Schroeder EB, Shah SH, Shay CM, Spartano NL, Stokes A, Tirschwell DL, VanWagner LB, Tsao CW, Wong SS, Heard DG. Heart disease and stroke statistics-2020 update: a report

from the american heart association. Circulation. 2020. https://doi.org/10.1161/CIR.0000000000000757.
3. Savarese G, Lund LH. Global public health burden of heart failure. Card Fail Rev. 2017;3:7–11. https://doi.org/10.15420/cfr.2016:25:2.
4. Fonarow GC, Adams KF, Abraham WT, Yancy CW, Boscardin WJ. Risk stratification for in-hospital mortality in acutely decompensated heart failure: classification and regression tree analysis. J Am Med Assoc. 2005. https://doi.org/10.1001/jama.293.5.572.
5. Levy WC, Mozaffarian D, Linker DT, Sutradhar SC, Anker SD, Cropp AB, Anand I, Maggioni A, Burton P, Sullivan MD, Pitt B, Poole-Wilson PA, Mann DL, Packer M. The seattle heart failure model: prediction of survival in heart failure. Circulation. 2006. https://doi.org/10.1161/CIRCULATIONAHA.105.584102.
6. Ahmad T, Munir A, Bhatti SH, Aftab M, Raza MA. Survival analysis of heart failure patients: a case study. PLoS One. 2017. https://doi.org/10.1371/journal.pone.0181001.
7. Zahid FM, Ramzan S, Faisal S, Hussain I. Gender based survival prediction models for heart failure patients: a case study in pakistan. PLoS One. 2019. https://doi.org/10.1371/journal.pone.0210602.
8. Chicco D, Jurman G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Med Inform Decis Mak. 2020. https://doi.org/10.1186/s12911-020-1023-5.
9. Cox DR. Regression models and life-tables. J R Stat Soc Ser B. 1972. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x.
10. Pocock SJ, Ariti CA, McMurray JJV, Maggioni A, Køber L, Squire IB, Swedberg K, Dobson J, Poppe KK, Whalley GA, Doughty RN. On behalf of the meta-analysis global group in chronic heart failure (MAGGIC): predicting survival in heart failure: a risk score based on 39 372 patients from 30 studies. Eur Heart J. 2012;34(19):1404–13. https://doi.org/10.1093/eurheartj/ehs337.
11. Schmidt M, Lipson H. Distilling free-form natural laws from experimental data. Science. 2009. https://doi.org/10.1126/science.1165893.
12. Vladislavleva EJ, Smits GF, den Hertog D. Order of nonlinearity as a complexity measure for models generated by symbolic regression via pareto genetic programming. IEEE Trans Evolut Comput. 2009. https://doi.org/10.1109/TEVC.2008.926486.
13. Dabhi VK, Vij SK. Empirical modeling using symbolic regression via postfix genetic programming. In: 2011 International Conference on Image Information Processing. 2011, p. 1–6. https://doi.org/10.1109/ICIIP.2011.6108857.
14. Udrescu SM, Tegmark M. AI Feynman: a physics-inspired method for symbolic regression. Sci Adv. 2020. https://doi.org/10.1126/sciadv.aay2631.
15. Abzu: Feyn software. 2020. https://pypi.org/project/feyn/ Accessed 2021-01-01.
16. Akaike H. A new look at the statistical model identification. IEEE Trans Autom Control. 1974. https://doi.org/10.1109/TAC.1974.1100705.
17. Davidson-Pilon C, Kalderstam J, Jacobson N, Reed S, Kuhn B, Zivich P, Williamson M, AbdealiJK Datta D, Fiore-Gartland A, Parij A, Wilson D, Gabriel Moneda L, Moncada-Torres A, Stark K, Gadgil H, Jona Singaravelan K, Besson L, Peña MS, Anton S, Klintberg A, Noorbakhsh J, Begun M, Kumar R, Hussey S, Seabold S, Golland D. CamDavidsonPilon/lifelines: v0257 Zenodo. 2020. https://doi.org/10.5281/zenodo.4313838.

Publisher's Note