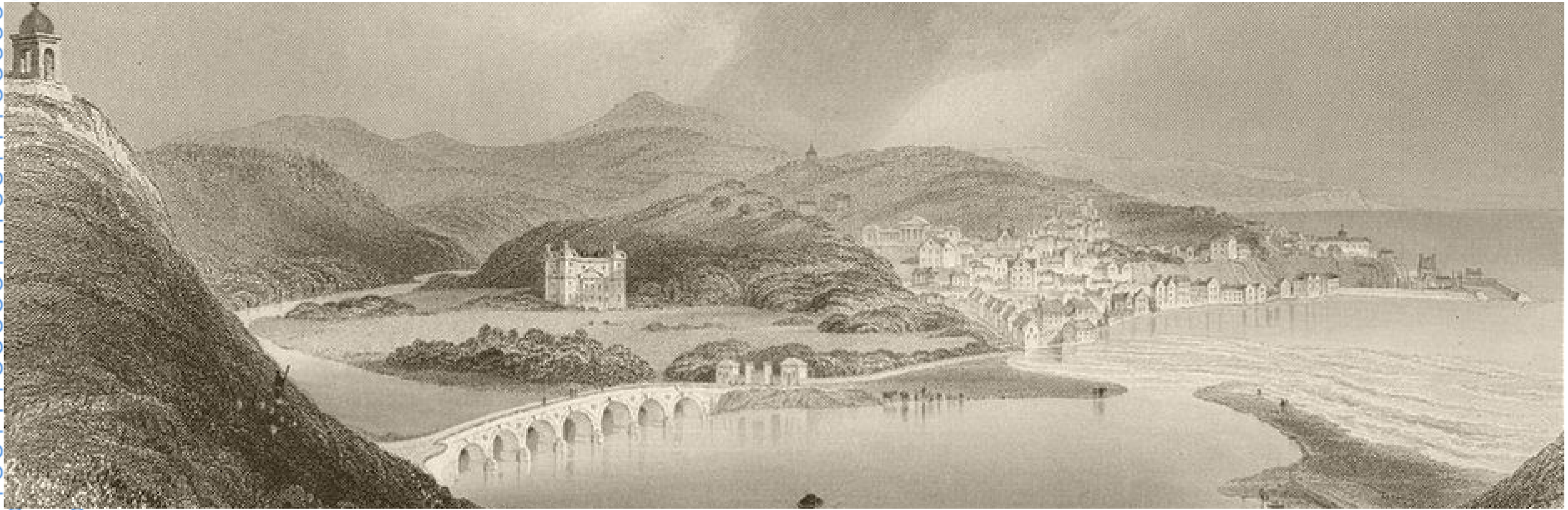


How do we make better graphs? Effective visual communication for the quantitative scientist

Mark Baillie

June 6th, 2019

npj | Digital Medicine

www.nature.com/npjdigitalmed

ARTICLE OPEN

Scalable and accurate deep learning with electronic health records

Alvin Rajkomar[1,2], Eyal Oren[1], Kai Chen[1], Andrew M. Dai[1], Nissan Hajaj[1], Michaela Hardt[1], Peter J. Liu[1], Xiaobing Liu[1], Jake Marcus[1], Mimi Sun[1], Patrik Sundberg[1], Hector Yee[1], Kun Zhang[1], Yi Zhang[1], Gerardo Flores[1], Gavin E. Duggan[1], Jamie Irvine[1], Quoc Le[1], Kurt Litsch[1], Alexander Mossin[1], Justin Tansuwan[1], De Wang[1], James Wexler[1], Jimbo Wilson[1], Dana Ludwig[2], Samuel L. Volchenboum[3], Katherine Chou[1], Michael Pearson[1], Srinivasan Madabushi[1], Nigam H. Shah[4], Atul J. Butte[2], Michael D. Howell[1], Claire Cui[1], Greg S. Corrado[1] and Jeffrey Dean[1]

npj | Digital Medicine www.nature.com/npjdigitalmed

ARTICLE OPEN

Scalable and accurate deep learning with electronic health records

Alvin Rajkomar[1,2], Eyal Oren[1], Kai Chen[1], Andrew M. Dai[1], Nissan Hajaj[1], Michaela Hardt[1], Peter J. Liu[1], Xiaobing Liu[1], Jake Marcus[1], Mimi Sun[1], Patrik Sundberg[1], Hector Yee[1], Kun Zhang[1], Yi Zhang[1], Gerardo Flores[1], Gavin E. Duggan[1], Jamie Irvine[1], Quoc Le[1], Kurt Litsch[1], Alexander Mossin[1], Justin Tansuwan[1], De Wang[1], James Wexler[1], Jimbo Wilson[1], Dana Ludwig[2], Samuel L. Volchenboum[3], Katherine Chou[1], Michael Pearson[1], Srinivasan Madabushi[1], Nigam H. Shah[4], Atul J. Butte[2], Michael D. Howell[1], Claire Cui[1], Greg S. Corrado[1] and Jeffrey Dean[1]

Scalable and accurate deep learning with electronic health records

Alvin Rajkomar[1,2], Eyal Oren[1], Kai Chen[1], Andrew M. Dai[1], Nissan Hajaj[1], Michaela Hardt[1], Peter J. Liu[1], Xiaobing Liu[1], Jake Marcus[1], Mimi Sun[1], Patrik Sundberg[1], Hector Yee[1], Kun Zhang[1], Yi Zhang[1], Gerardo Flores[1], Gavin E. Duggan[1], Jamie Irvine[1], Quoc Le[1], Kurt Litsch[1], Alexander Mossin[1], Justin Tansuwan[1], De Wang[1], James Wexler[1], Jimbo Wilson[1], Dana Ludwig[2], Samuel L. Volchenboum[3], Katherine Chou[1], Michael Pearson[1], Srinivasan Madabushi[1], Nigam H. Shah[4], Atul J. Butte[2], Michael D. Howell[1], Claire Cui[1], Greg S. Corrado[1] and Jeffrey Dean[1]

Predictive modeling with electronic health record (EHR) data is anticipated to drive personalized medicine and improve healthcare quality. Constructing predictive statistical models typically requires extraction of curated predictor variables from normalized EHR data, a labor-intensive process that discards the vast majority of information in each patient's record. We propose a representation of patients' entire raw EHR records based on the Fast Healthcare Interoperability Resources (FHIR) format. We demonstrate that deep learning methods using this representation are capable of accurately predicting multiple medical events from multiple centers without site-specific data harmonization. We validated our approach using de-identified EHR data from two US academic medical centers with 216,221 adult patients hospitalized for at least 24 h. In the sequential format we propose, this volume of EHR data unrolled into a total of 46,864,534,945 data points, including clinical notes. Deep learning models achieved high accuracy for tasks such as predicting: in-hospital mortality (area under the receiver operator curve [AUROC] across sites 0.93–0.94), 30-day unplanned readmission (AUROC 0.75–0.76), prolonged length of stay (AUROC 0.85–0.86), and all of a patient's final discharge diagnoses (frequency-weighted AUROC 0.90). These models outperformed traditional, clinically-used predictive models in all cases. We believe that this approach can be used to create accurate and scalable predictions for a variety of clinical scenarios. In a case study of a particular prediction, we demonstrate that neural networks can be used to identify relevant information from the patient's chart.

npj Digital Medicine (2018)1:18 ; doi:10.1038/s41746-018-0029-1

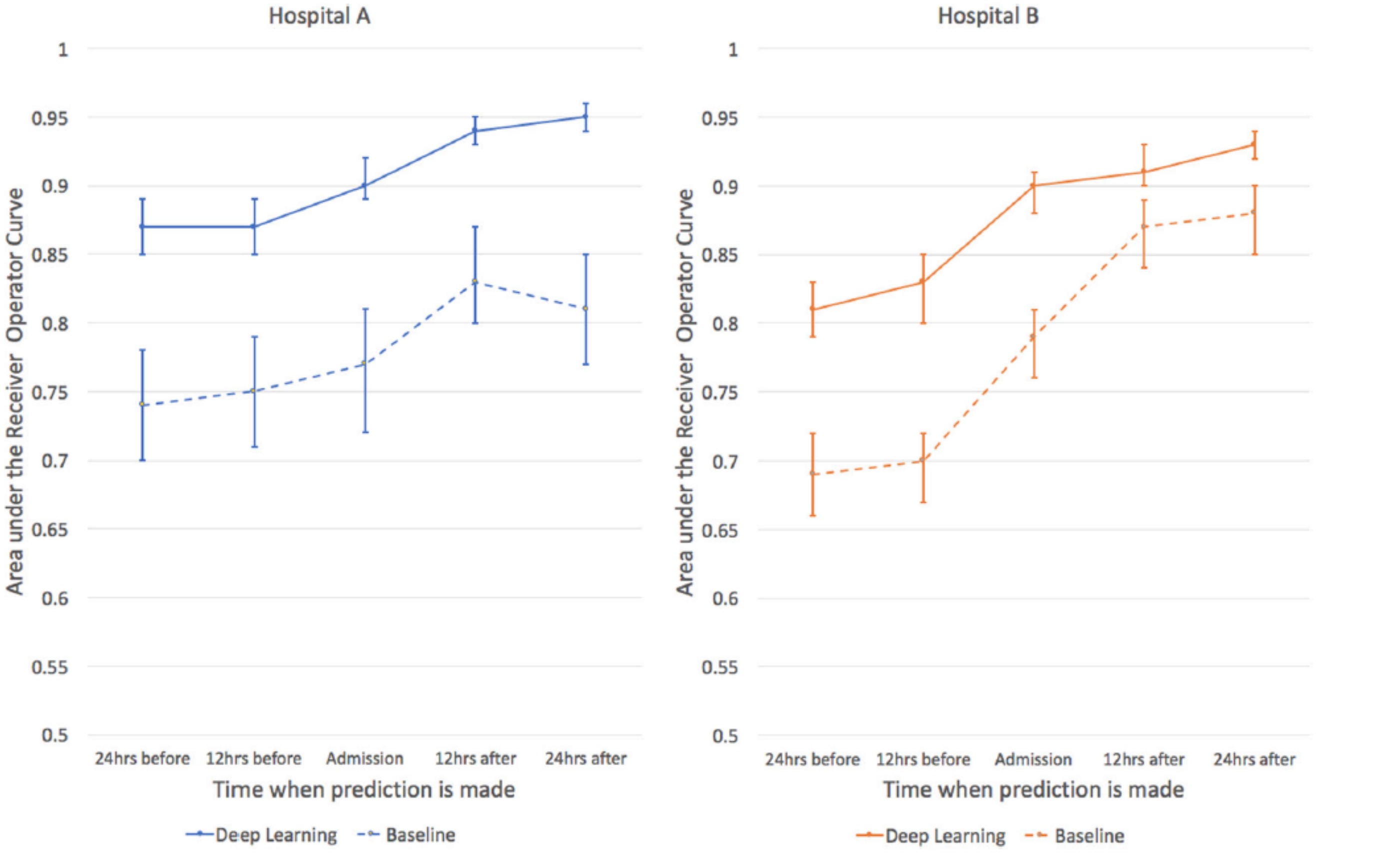


Fig. 2 The area under the receiver operating characteristic curves are shown for predictions of inpatient mortality made by deep learning and baseline models at 12 h increments before and after hospital admission. For inpatient mortality, the deep learning model achieves higher discrimination at every prediction time compared to the baseline for both the University of California, San Francisco (UCSF) and University of Chicago Medicine (UCM) cohorts. Both models improve in the first 24 h, but the deep learning model achieves a similar level of accuracy approximately 24 h earlier for UCM and even 48 h earlier for UCSF. The error bars represent the bootstrapped 95% confidence interval

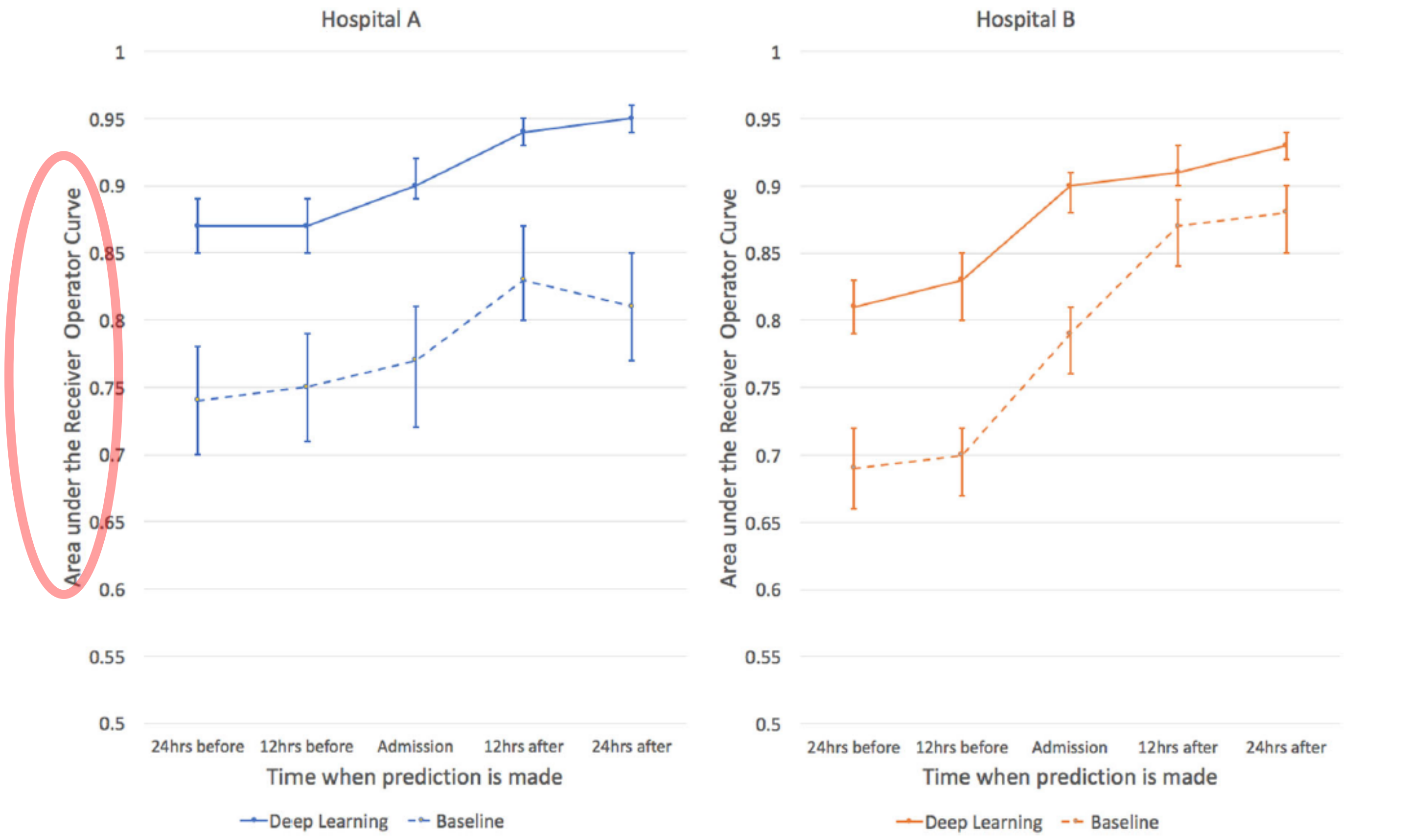


Fig. 2 The area under the receiver operating characteristic curves are shown for predictions of inpatient mortality made by deep learning and baseline models at 12 h increments before and after hospital admission. For inpatient mortality, the deep learning model achieves higher discrimination at every prediction time compared to the baseline for both the University of California, San Francisco (UCSF) and University of Chicago Medicine (UCM) cohorts. Both models improve in the first 24 h, but the deep learning model achieves a similar level of accuracy approximately 24 h earlier for UCM and even 48 h earlier for UCSF. The error bars represent the bootstrapped 95% confidence interval

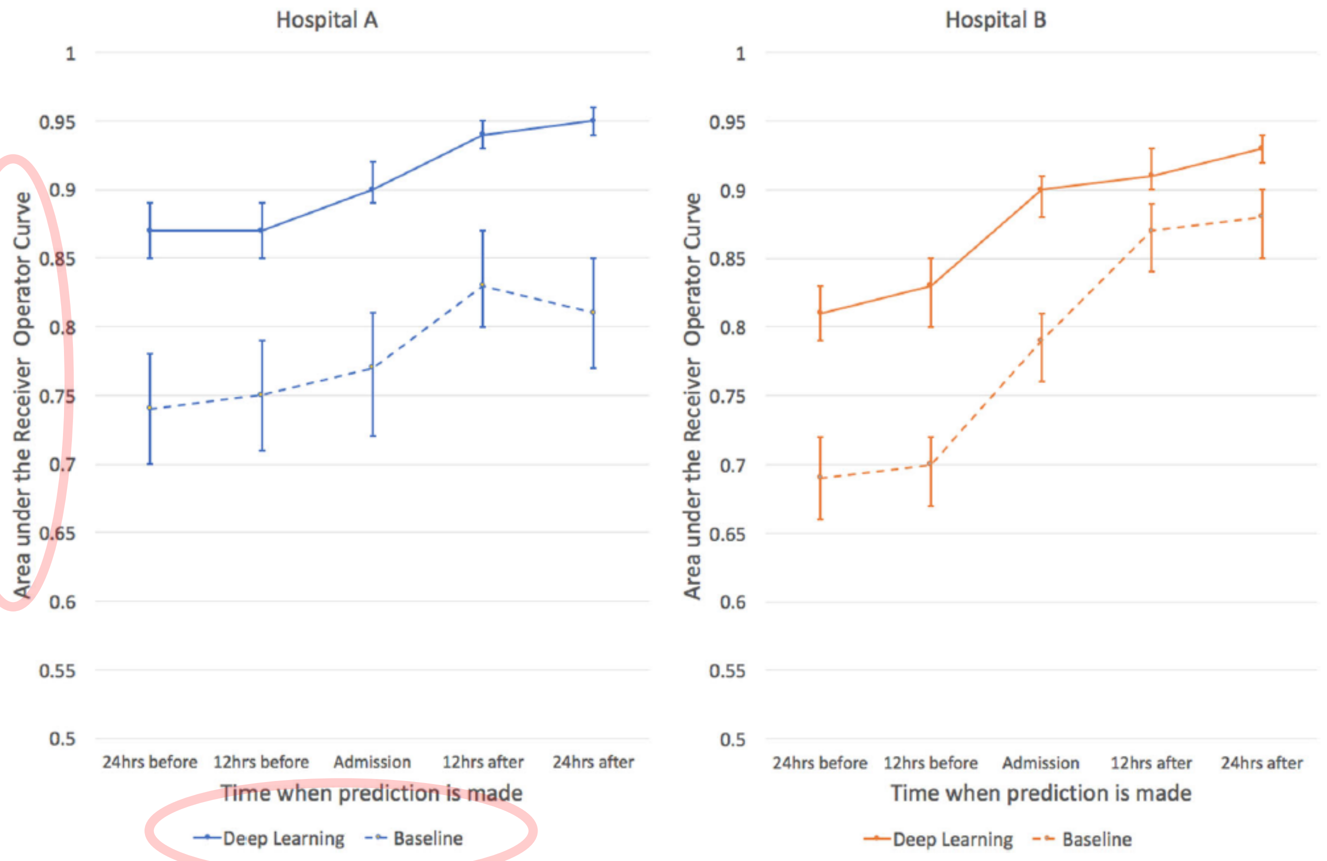


Fig. 2 The area under the receiver operating characteristic curves are shown for predictions of inpatient mortality made by deep learning and baseline models at 12 h increments before and after hospital admission. For inpatient mortality, the deep learning model achieves higher discrimination at every prediction time compared to the baseline for both the University of California, San Francisco (UCSF) and University of Chicago Medicine (UCM) cohorts. Both models improve in the first 24 h, but the deep learning model achieves a similar level of accuracy approximately 24 h earlier for UCM and even 48 h earlier for UCSF. The error bars represent the bootstrapped 95% confidence interval

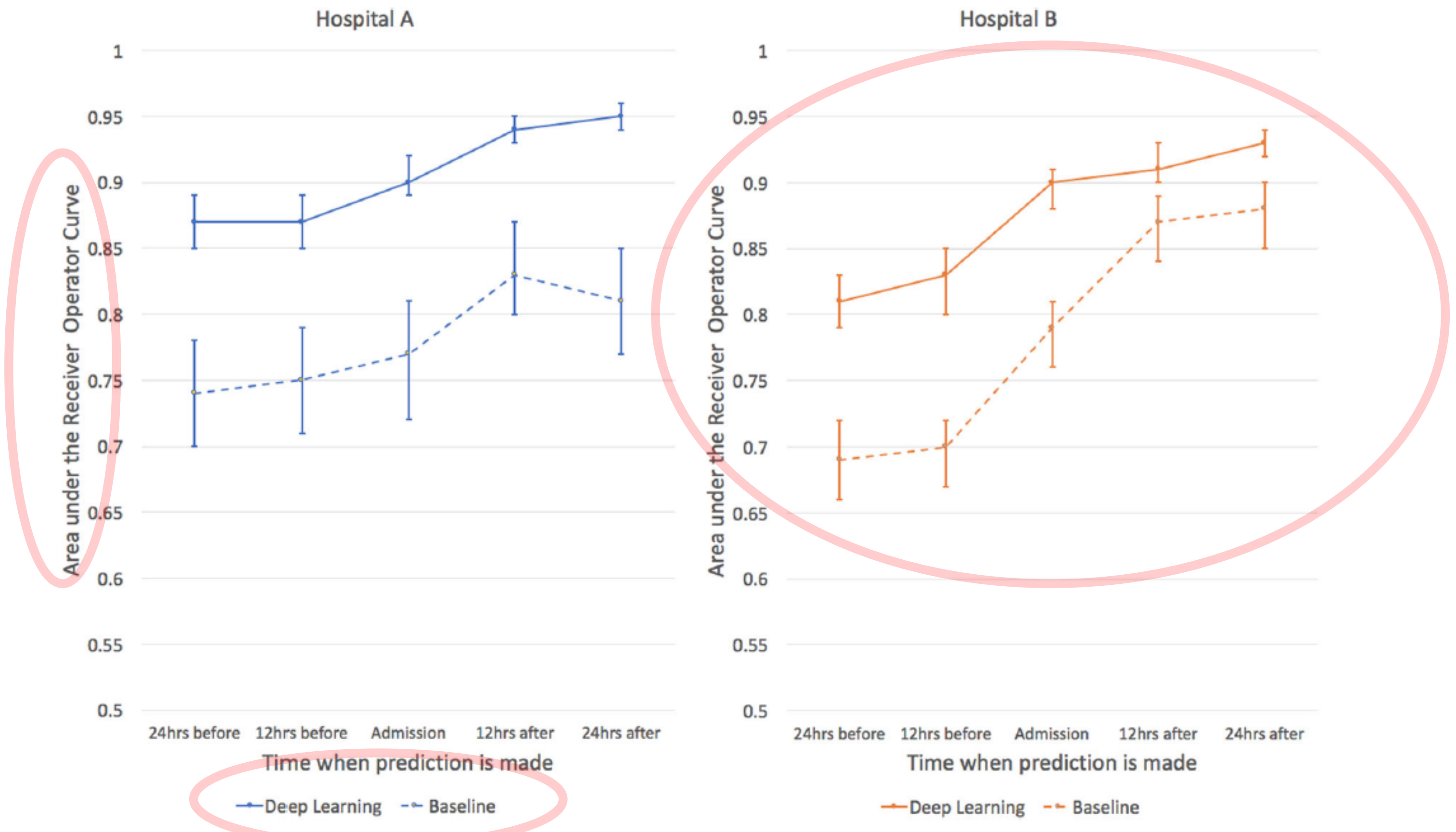


Fig. 2 The area under the receiver operating characteristic curves are shown for predictions of inpatient mortality made by deep learning and baseline models at 12 h increments before and after hospital admission. For inpatient mortality, the deep learning model achieves higher discrimination at every prediction time compared to the baseline for both the University of California, San Francisco (UCSF) and University of Chicago Medicine (UCM) cohorts. Both models improve in the first 24 h, but the deep learning model achieves a similar level of accuracy approximately 24 h earlier for UCM and even 48 h earlier for UCSF. The error bars represent the bootstrapped 95% confidence interval

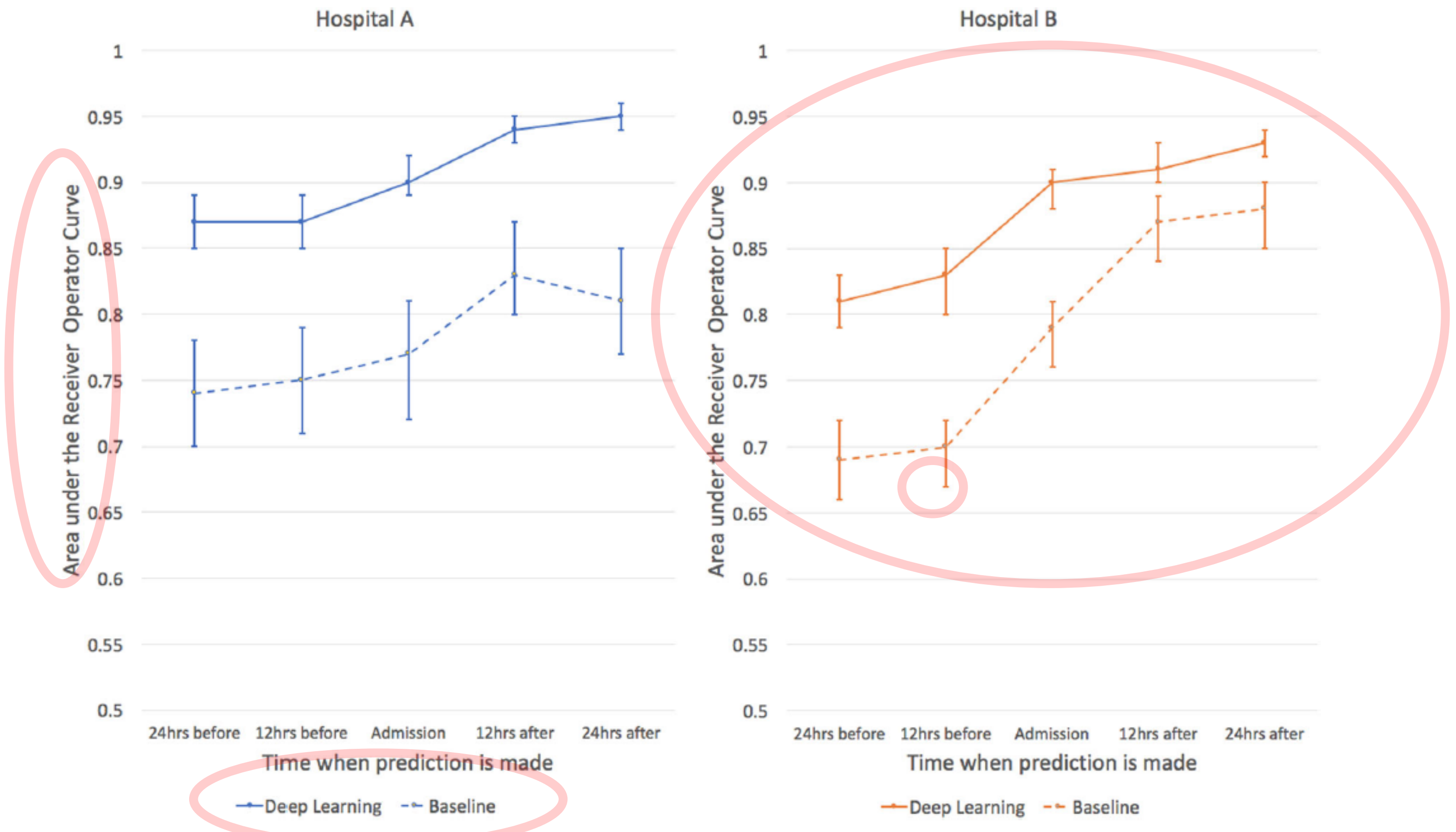


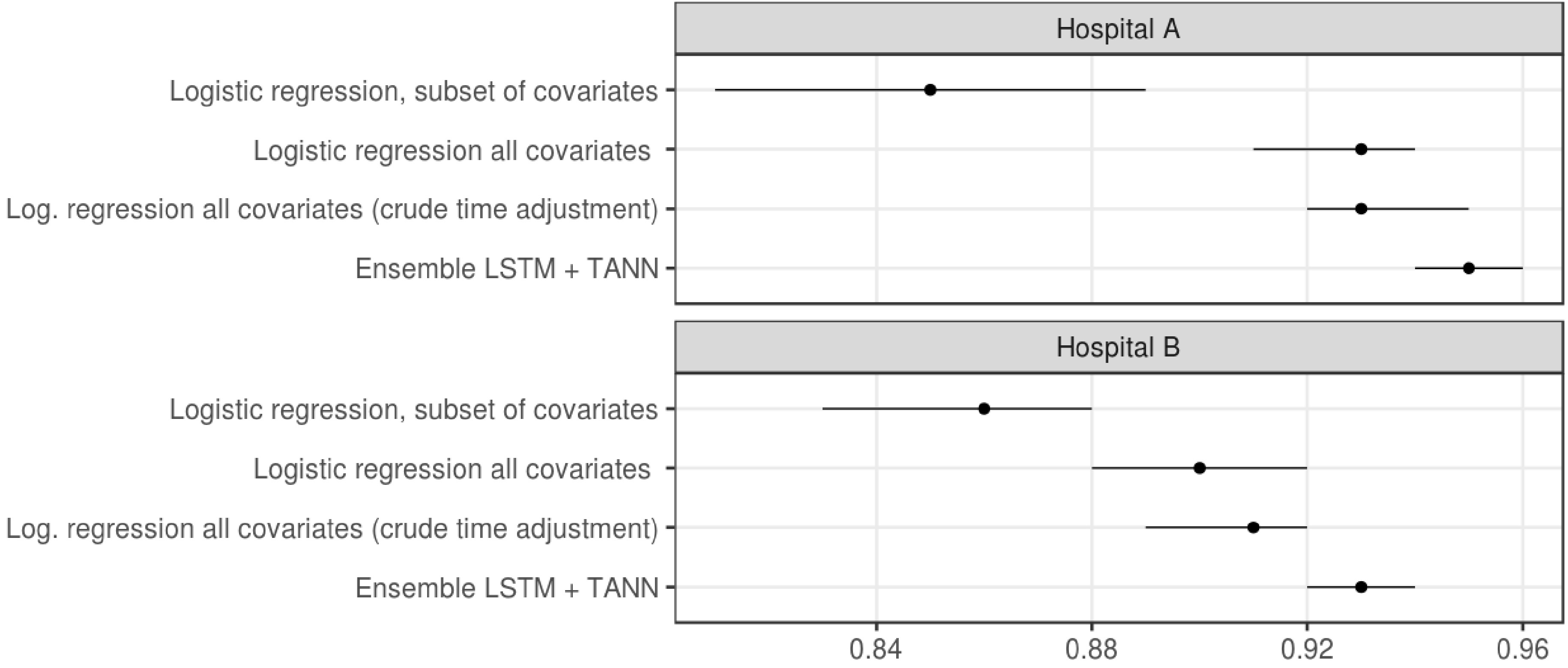
Fig. 2 The area under the receiver operating characteristic curves are shown for predictions of inpatient mortality made by deep learning and baseline models at 12 h increments before and after hospital admission. For inpatient mortality, the deep learning model achieves higher discrimination at every prediction time compared to the baseline for both the University of California, San Francisco (UCSF) and University of Chicago Medicine (UCM) cohorts. Both models improve in the first 24 h, but the deep learning model achieves a similar level of accuracy approximately 24 h earlier for UCM and even 48 h earlier for UCSF. The error bars represent the bootstrapped 95% confidence interval

Supplemental Table 1: Prediction accuracy of each task of deep learning model compared to baselines

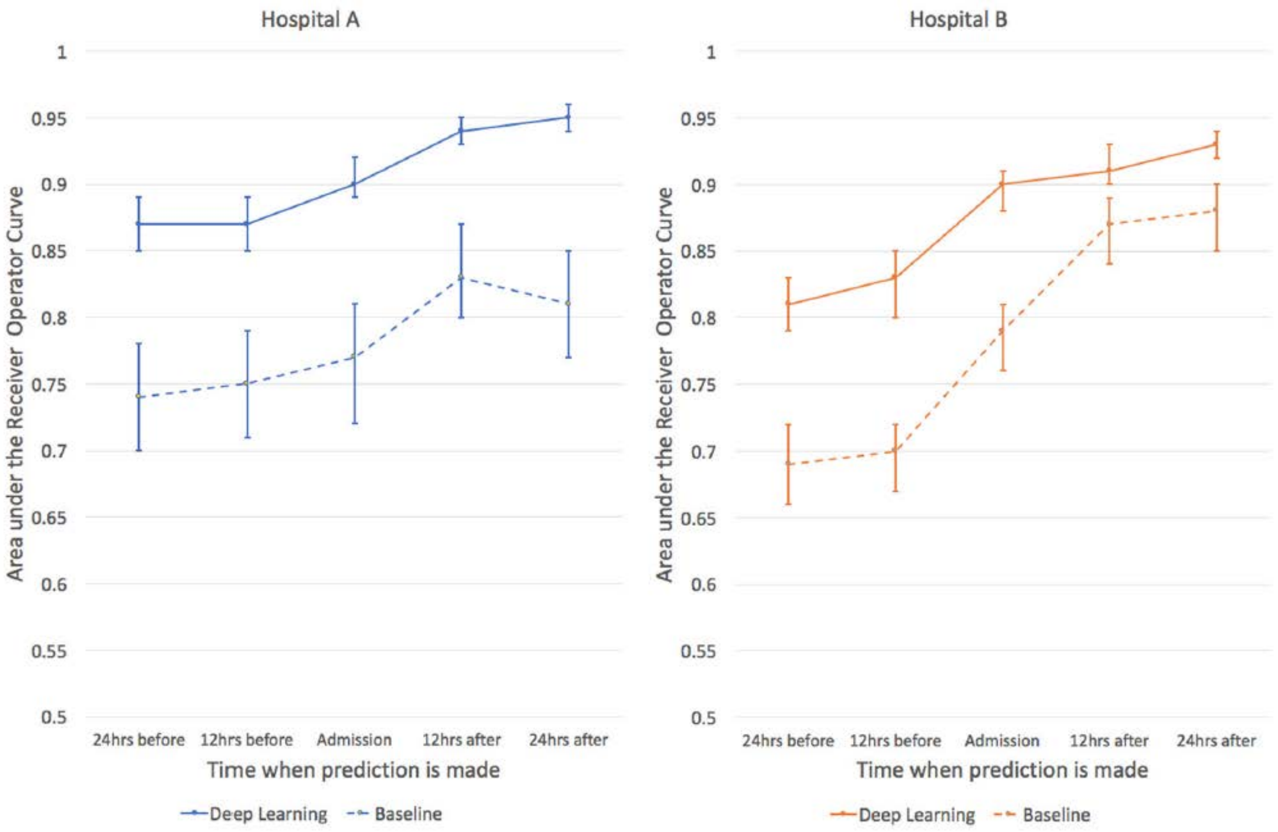
	Hospital A	Hospital B
Inpatient Mortality, AUROC[1](95% CI)		
Deep learning 24 hours after admission	**0.95**(0.94-0.96)	**0.93**(0.92-0.94)
Full feature enhanced baseline at 24 hours after admission	0.93 (0.92-0.95)	0.91 (0.89-0.92)
Full feature simple baseline at 24 hours after admission	0.93 (0.91-0.94)	0.90 (0.88-0.92)
Baseline (aEWS[2]) at 24 hours after admission	0.85 (0.81-0.89)	0.86 (0.83-0.88)
30-day Readmission, AUROC (95% CI)		
Deep learning at discharge	**0.77**(0.75-0.78)	**0.76**(0.75-0.77)
Full feature enhanced baseline at discharge	0.75 (0.73-0.76)	0.75 (0.74-0.76)
Full feature simple baseline at discharge	0.74 (0.73-0.76)	0.73 (0.72-0.74)
Baseline (mHOSPITAL[3]) at discharge	0.70 (0.68-0.72)	0.68 (0.67-0.69)
Length of Stay at least 7 days AUROC (95% CI)		
Deep learning 24 hours after admission	**0.86**(0.86-0.87)	**0.85**(0.85-0.86)
Full feature enhanced baseline at 24 hours after admission	0.85 (0.84-0.85)	0.83 (0.83-0.84)
Full feature simple baseline at 24 hours after admission	0.83 (0.82-0.84)	0.81 (0.80-0.82)
Baseline (mLiu[4]) at 24 hours after admission	0.76 (0.75-0.77)	0.74 (0.73-0.75)

[1] Area under the receiver operator curve
[2] Augmented early warning score
[3] Modified HOSPITAL score
[4] Modified Liu score

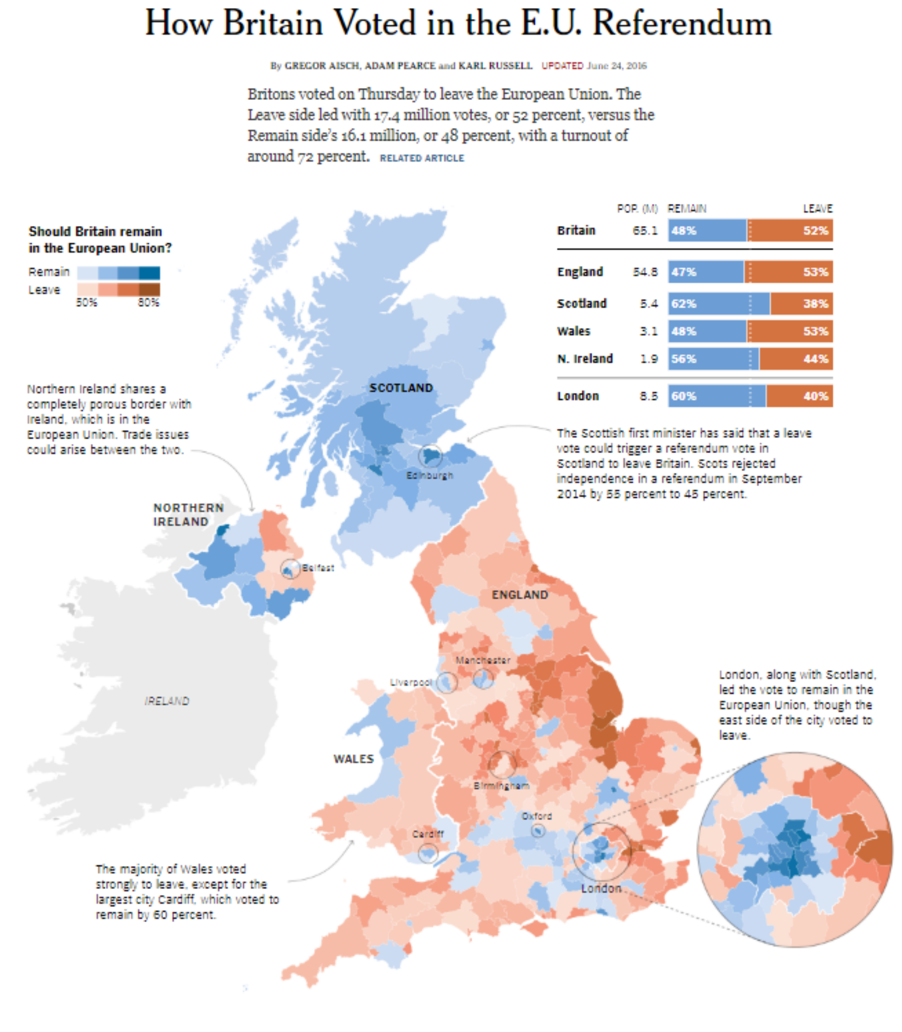
Optimised deep learning ensembles are not the driver of “accuracy”



We presented our “best” model and compared it to an inappropriate baseline



Effective visualisation is important

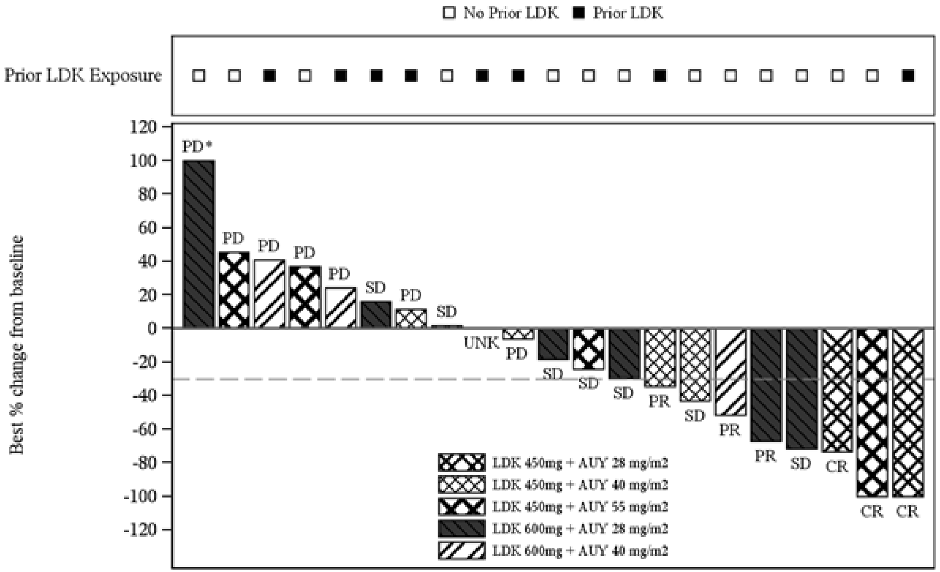


Sources: Preliminary results data from the BBC, British Office of National Statistics

https://www.nytimes.com/interactive/2016/06/24/world/europe/how-britain-voted-brexit-referendum.html

We are not always good at it

Figure 11-1 (Page 1 of 1)
Best percentage change from baseline in sum of longest diameters and best overall response as per investigator by prior LDK378 treatment
(Full analysis set)



- * Denotes the percentage change from baseline greater than 100.
Source: Table 11-4, Listing 14.2-1.2 and Listing 16.2.4-1.5

NOVARTIS

Visualisation panel

“Visualization and the use of graphics can help at every stage of an analysis, from the planning and design of an experiment, the very first data explorations, through to the communication of conclusions and recommendations. Visualization is more than "plotting data"; it can lead to a deeper understanding and inform next steps.

The role of the STRATOS visualization panel is to promote the use of good graphical principles for effective visual communication, providing guidance and recommendations covering all aspects from the design, implementation and review of statistical graphics.”

Effective visualisation is important throughout the workflow

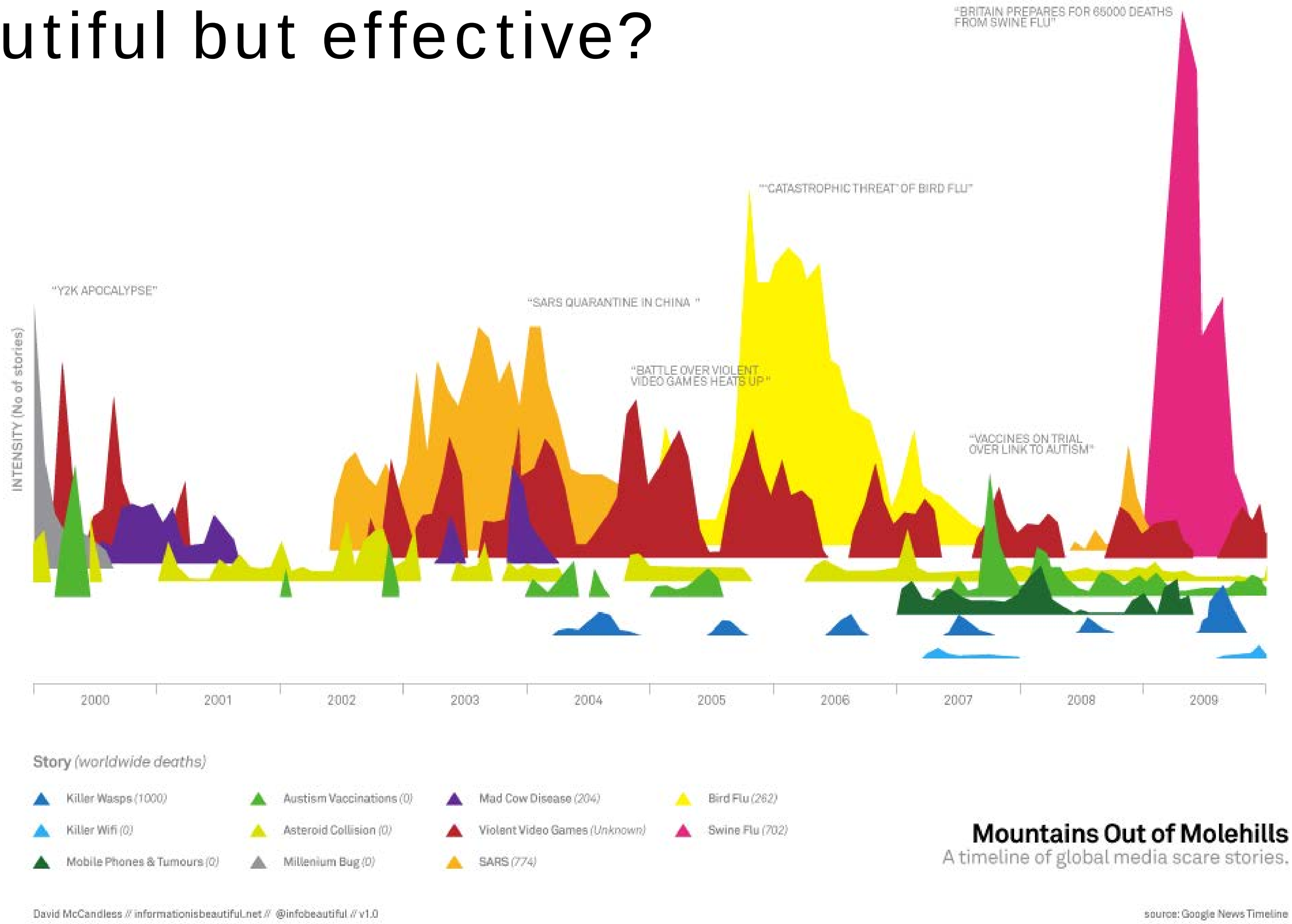
Topic groups

1	Missing data
2	Selection of variables and functional forms in multivariable analysis
3	Initial data analysis
4	Measurement error and misclassification
5	Study design
6	Evaluating diagnostic tests and prediction models
7	Causal inference
8	Survival analysis
9	High-dimensional data

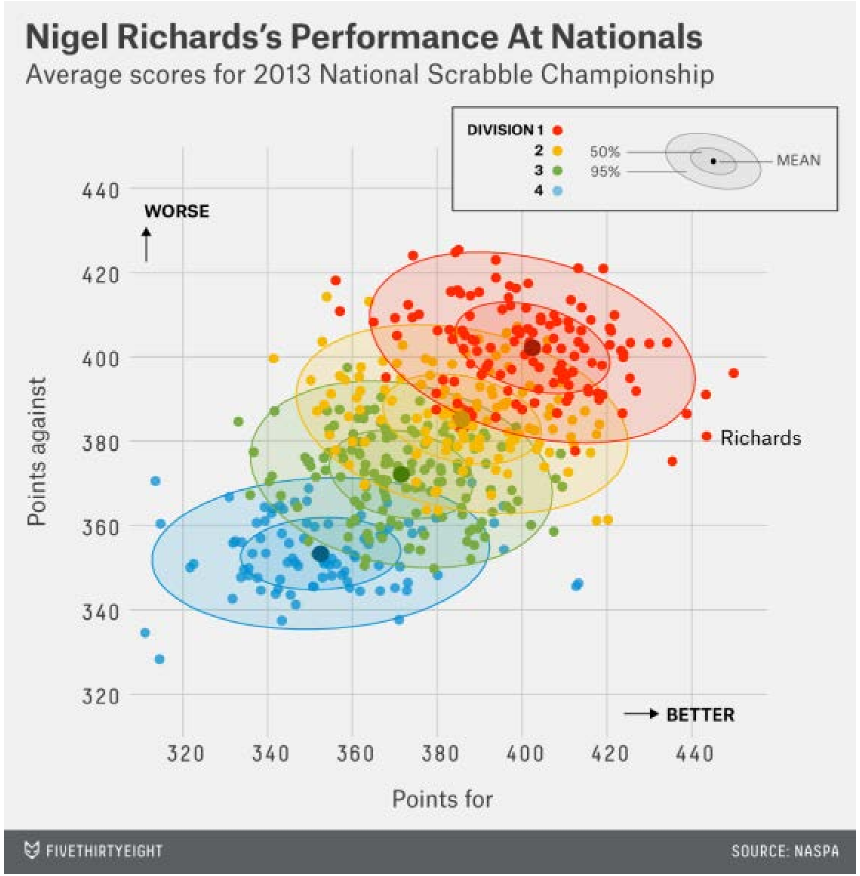
Elements of the initiative

- Graphical principles and thinking
  1. Graphics Principles Cheat Sheet
  2. Newsletter
- Easing the implementation
  3. Graph Gallery
  4. Analysis Results Datasets
  5. Standardization of most common/important graphs
- Graphics tomorrow ... or today?
  6. Question-based visualizations and interactive graphics
- ...plus overarching stakeholder management and communication

Beautiful but effective?



Beautiful and effective?



Effective data visualisation is effective visual communication

- Effective graphs...
  - are visually appealing, intuitive, legible
  - use the correct graph type and axis scales
  - use proximity & alignment to facilitate comparison
  - use labels and annotations to add clarity to the message
- Most importantly, effective use of visualisations
  - Enables clear and impactful communication
  - Elevates influence with stakeholders
  - Facilitates informed decision making

Principles for effective visual communication

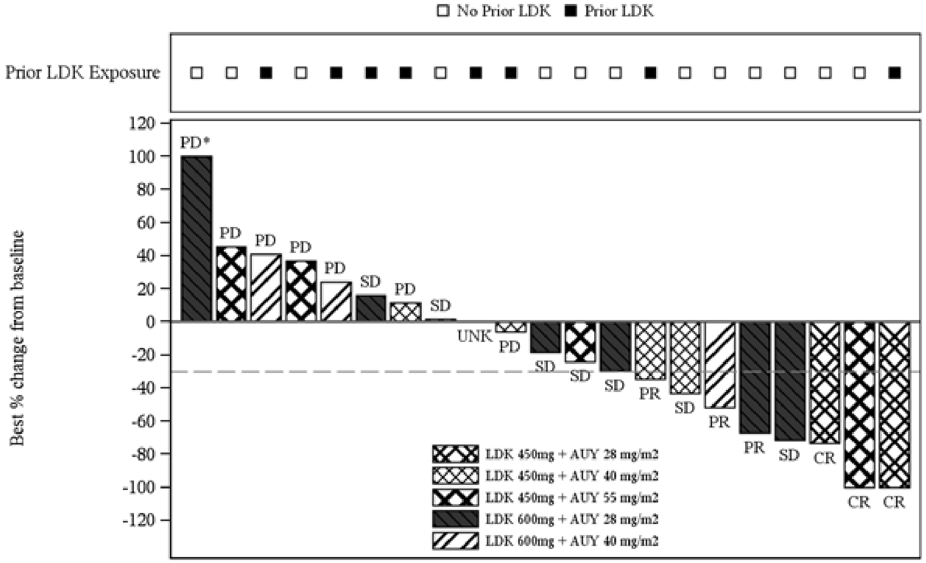
Use the cheat sheet for design and planning

- **Why** — Clearly identify the purpose of the graph, e.g. to deliver a message or for exploration?
- **What** — Identify the quantitative evidence to support the purpose
- **Who** — Identify the intended audience (specialists, non-specialists, both) and focus the design to support their needs
- **Where** — Adapt the design to space or formatting constraints (e.g. clinical report, slide deck or publication)

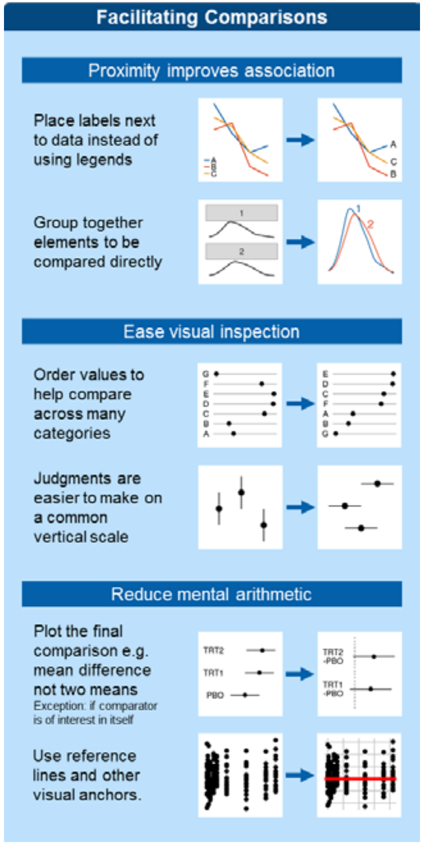
https://graphicsprinciples.github.io/

Use the cheat sheet for critical review

Figure 11-1 (Page 1 of 1)
Best percentage change from baseline in sum of longest diameters and best overall response as per investigator by prior LDK378 treatment
(Full analysis set)

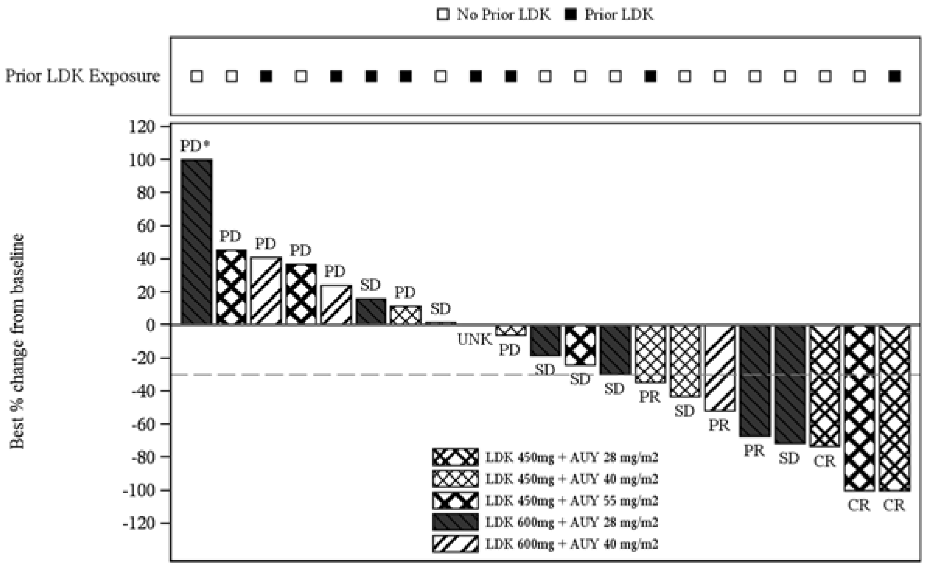


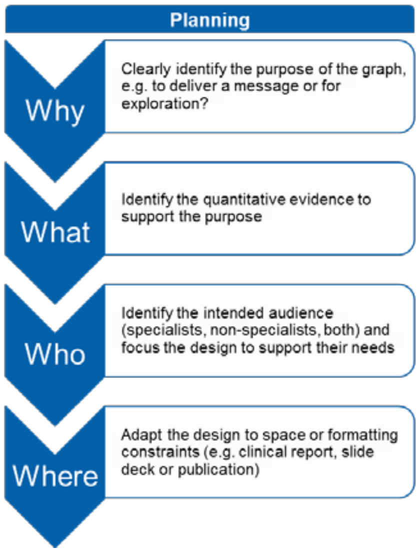
- * Denotes the percentage change from baseline greater than 100.
Source: Table 11-4, Listing 14.2-1.2 and Listing 16.2.4-1.5



Use the cheat sheet for critical review

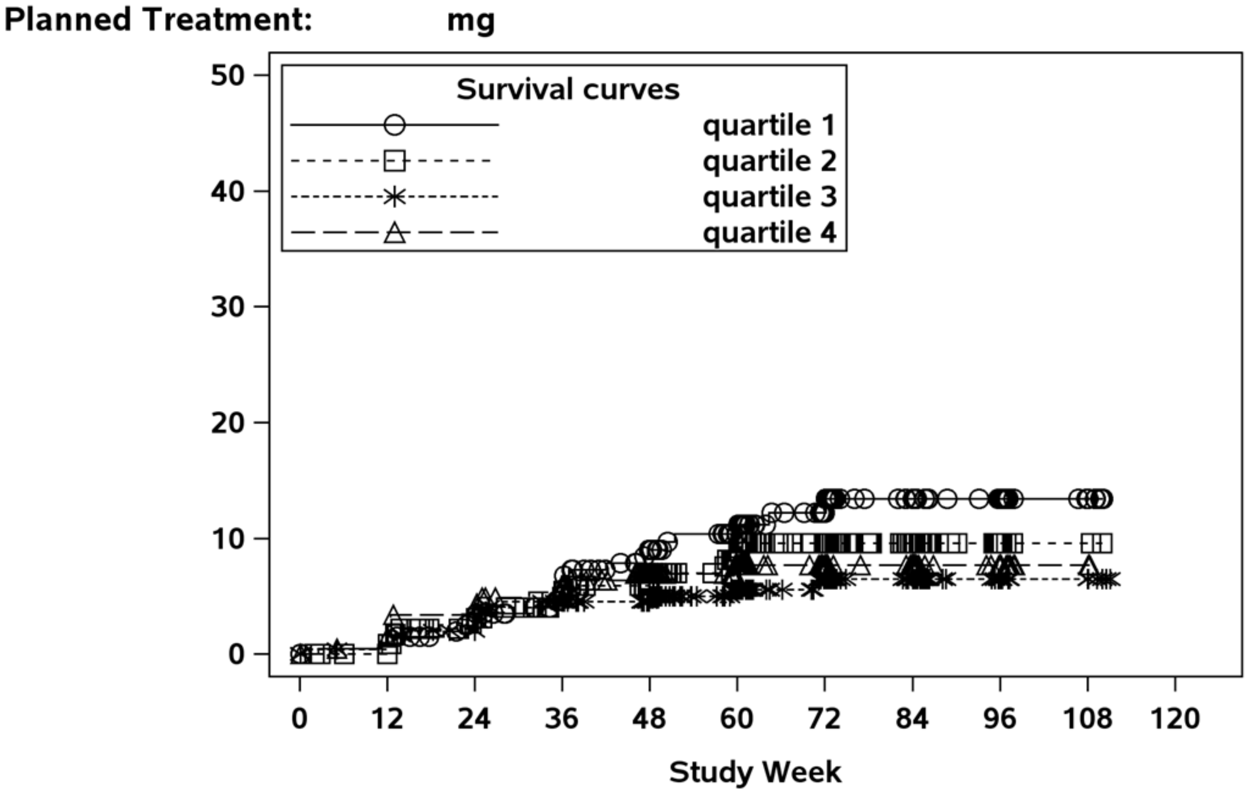
Figure 11-1 (Page 1 of 1)
Best percentage change from baseline in sum of longest diameters and best overall response as per investigator by prior LDK378 treatment
(Full analysis set)



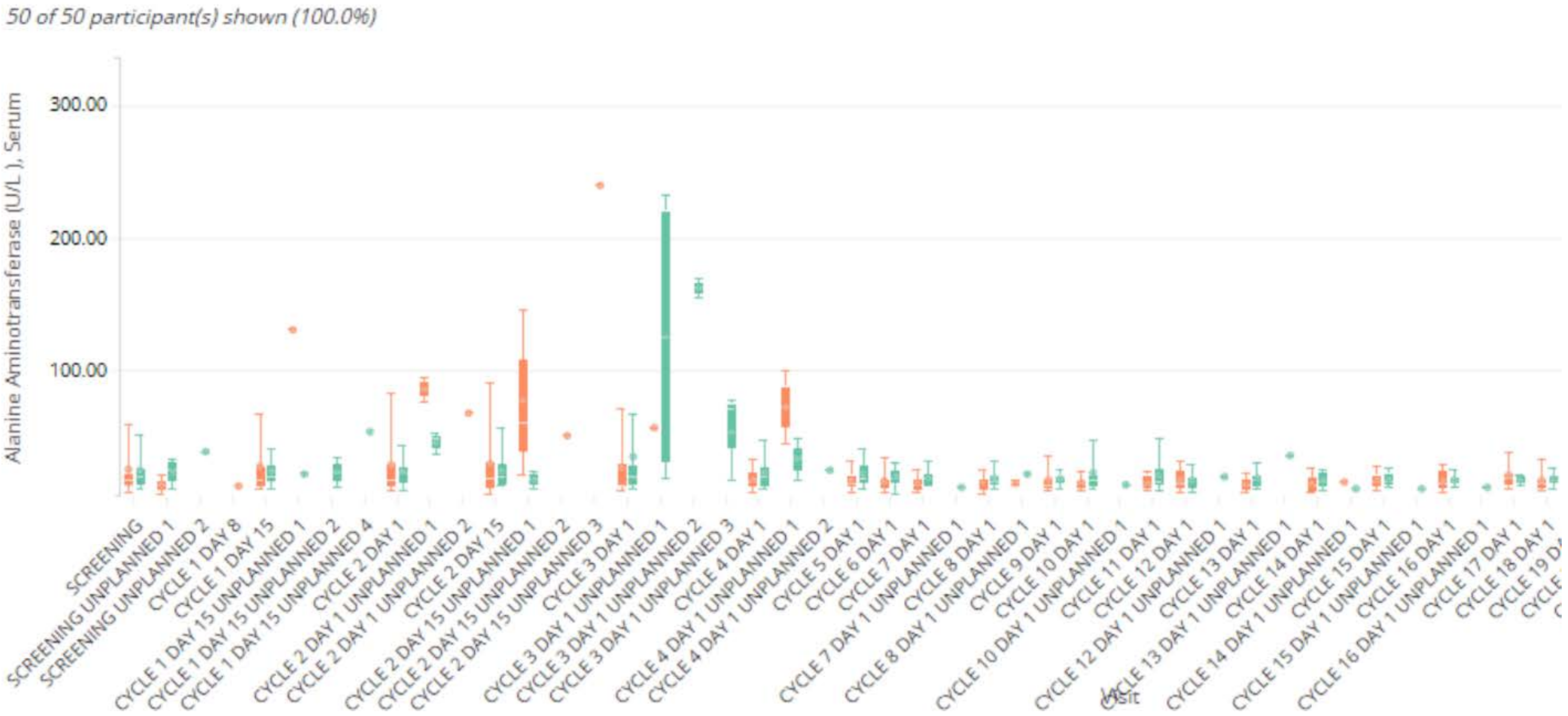


- * Denotes the percentage change from baseline greater than 100.
Source: Table 11-4, Listing 14.2-1.2 and Listing 16.2.4-1.5

This is a continual process



This is a continual process



Three laws for improving visual communication

Have a clear purpose

- Know the purpose of creating the graph
- Identify the quantitative evidence to support the purpose
- Identify the audience and focus the design to support their needs

Show the data clearly

- Choose the appropriate graph type to display your data
- Avoid misrepresentation (use appropriate scales)
- Maximize data to ink ratio (reduce distraction, less is more)

Make the message obvious

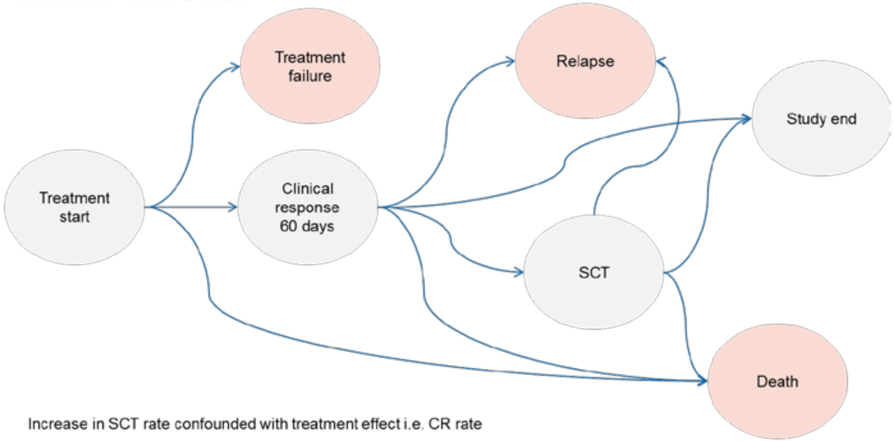
- Use proximity and alignment to aid in comparisons
- Minimize mental arithmetic (e.g. plot the difference)
- Use colors and annotations to highlight important details

https://arxiv.org/abs/1903.09512

Effective visualization is important for planning

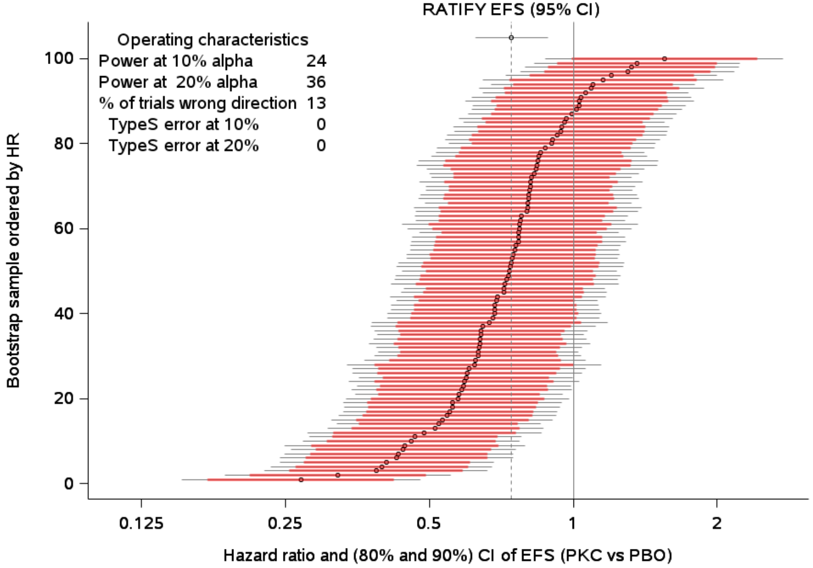
Event free survival endpoint

An EFS event for the key secondary endpoint is defined as a failure to achieve a CR within 60 days of study treatment, relapse from CR, or death due to any cause, whichever occurs first.



Increase in SCT rate confounded with treatment effect i.e. CR rate

60 sampled RATIFY patients (1:1 ratio) with poor FLT3 imbalance



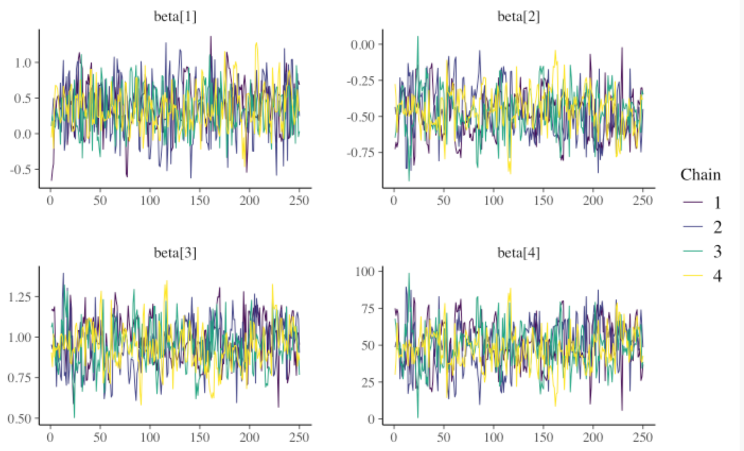
Effective visualization is important during exploratory analysis

Original Article | Free Access |

Visualization in Bayesian workflow

Jonah Gabry, Daniel Simpson, Aki Vehtari, Michael Betancourt, Andrew Gelman

First published: 15 January 2019 | https://doi.org/10.1111/rssa.12378



https://mc-stan.org/bayesplot/reference/MCMC-traces.html

Effective visualization is important during exploratory analysis

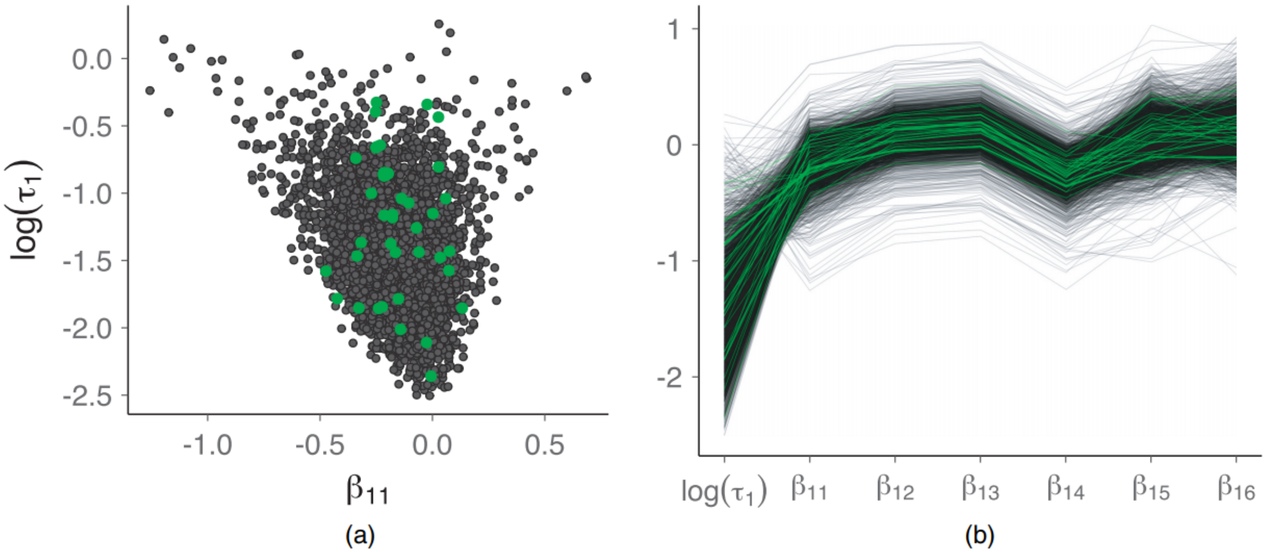
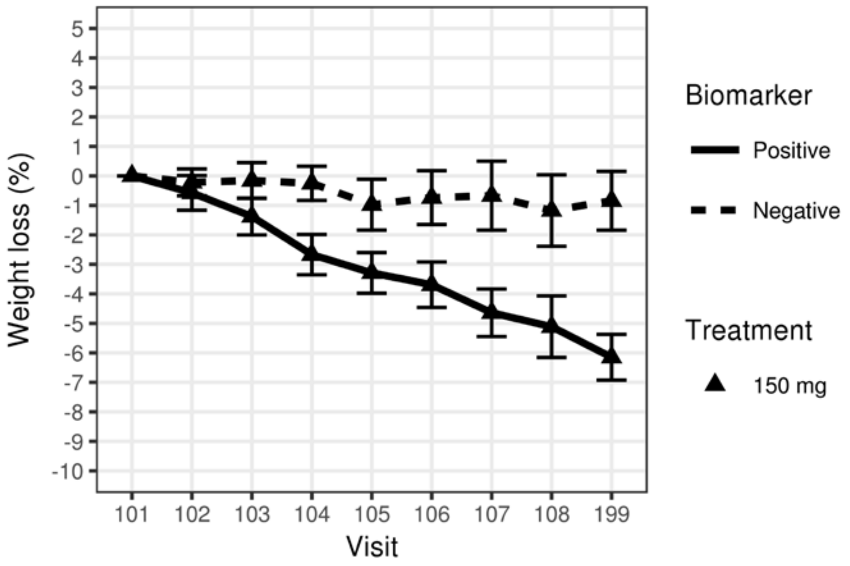


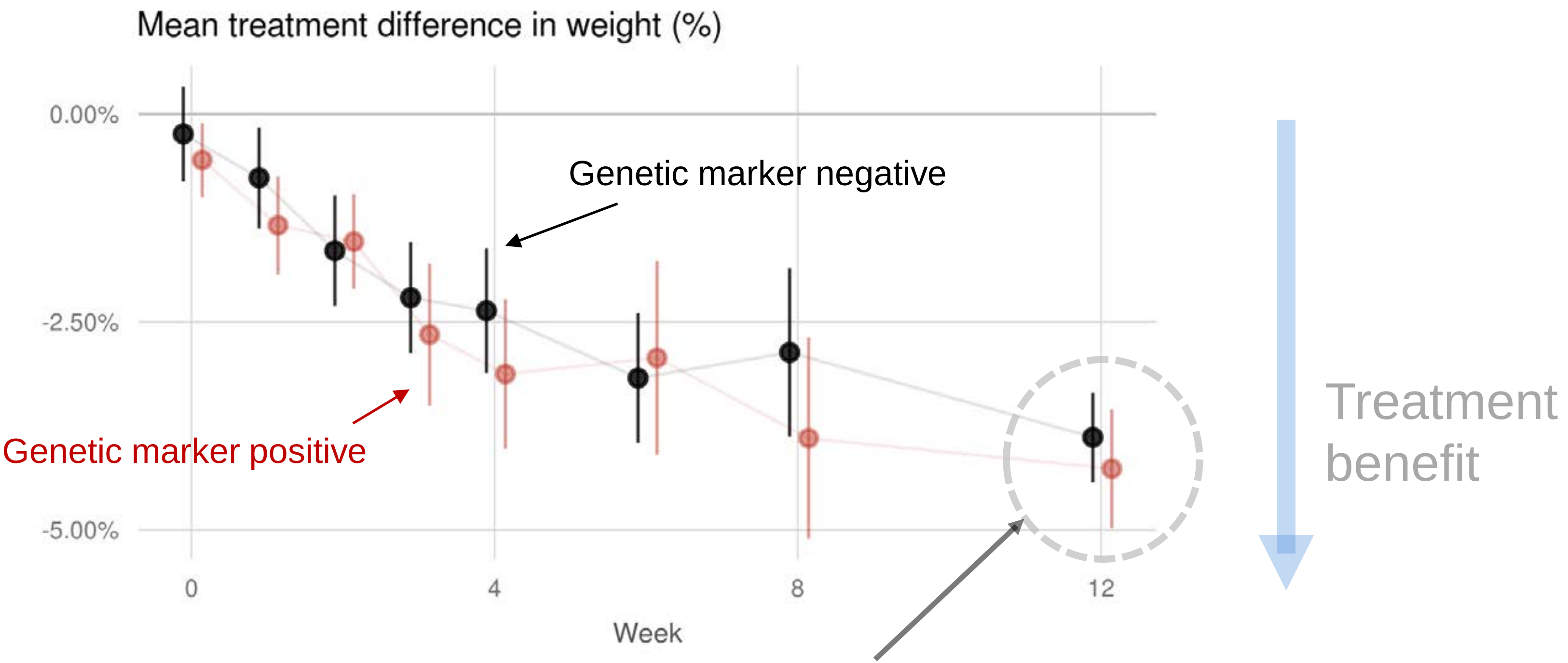
Fig. 5. Diagnostic plots for HMC sampling (models were fitted by using the RStan interface to Stan 2.17 (Stan Development Team, 2017a)): (a) for model 3, a bivariate plot of the log-standard-deviation of the cluster level slopes (y-axis) against the slope for the first cluster (x-axis) (the green dots indicate starting points of divergent transitions; this plot can be made by using mcmc_scatter in bayesplot); (b) for model 3, a parallel co-ordinates plot showing the cluster level slope parameters and their log-standard-deviation $\log(\tau_1)$ (the green lines indicate starting points of divergent transitions; this plot can be made by using mcmc_parcoord in bayesplot)

Effective visualization important for reporting

%improvement in baseline weight through week 12 by subgroup



Genetic marker positive is **not** predictive of treatment response



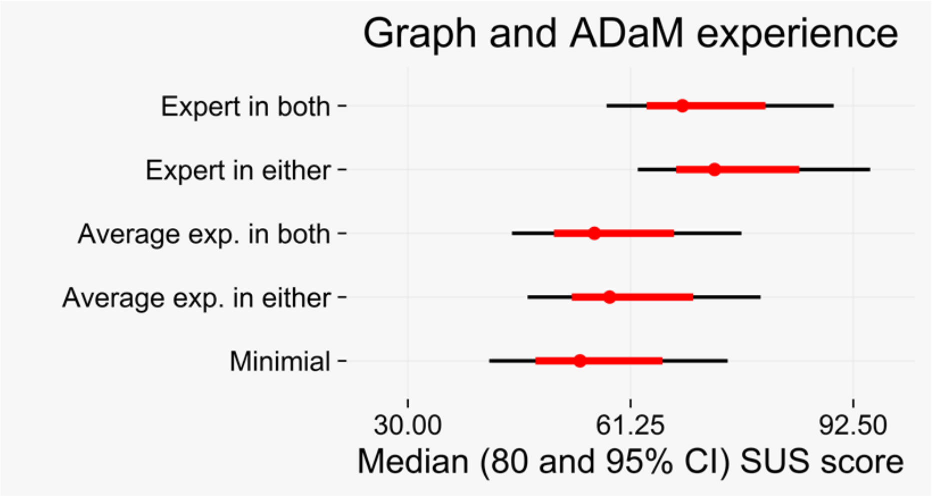
The average treatment effect is **similar** in **both** the genetic marker positive and negative subgroups and does **not** warrant further investigation

How can the VP help across skill levels?

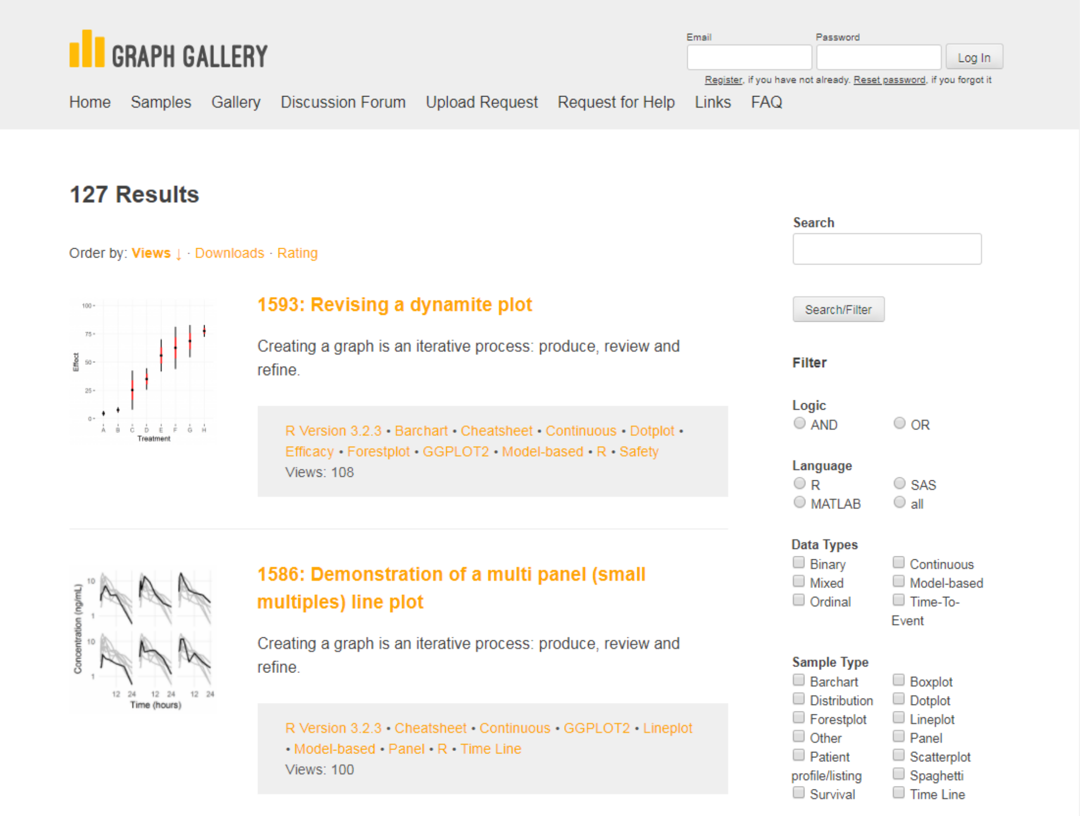
A survey was sent to associates working with clinical data

The purpose to:

- evaluate ADaM as a data standard for graph production
- Identify key issues associates currently experience
- Explore issues related to role and experience level
- 85 respondents



Provide access to examples with code and data



How can the VP help with implementation across skill levels?

(Negative) qualitative comments

- ***Complex graphs such as Forest Plots will need a fair amount of data manipulation to get all needed for input to the graph***
- ***some variables*** *needed for graphs are not in ADaM datasets*
- *Trying to figure out what the different parameters mean and* ***extracting the information relevant to my task.***

ORIGINAL ARTICLE

Aliskiren, Enalapril, or Aliskiren and Enalapril in Heart Failure

John J.V. McMurray, M.D., Henry Krum, M.B., B.S., Ph.D., William T. Abraham, M.D., Kenneth Dickstein, M.D., Ph.D., Lars V. Køber, M.D., D.M.Sc., Akshay S. Desai, M.D., M.P.H., Scott D. Solomon, M.D., Nicola Greenlaw, M.Sc., M. Atif Ali, B.A., Yanntong Chiang, Ph.D., Qing Shao, Ph.D., Georgia Tarnesby, M.B., B.Chir., et al., for the ATMOSPHERE Committees Investigators†

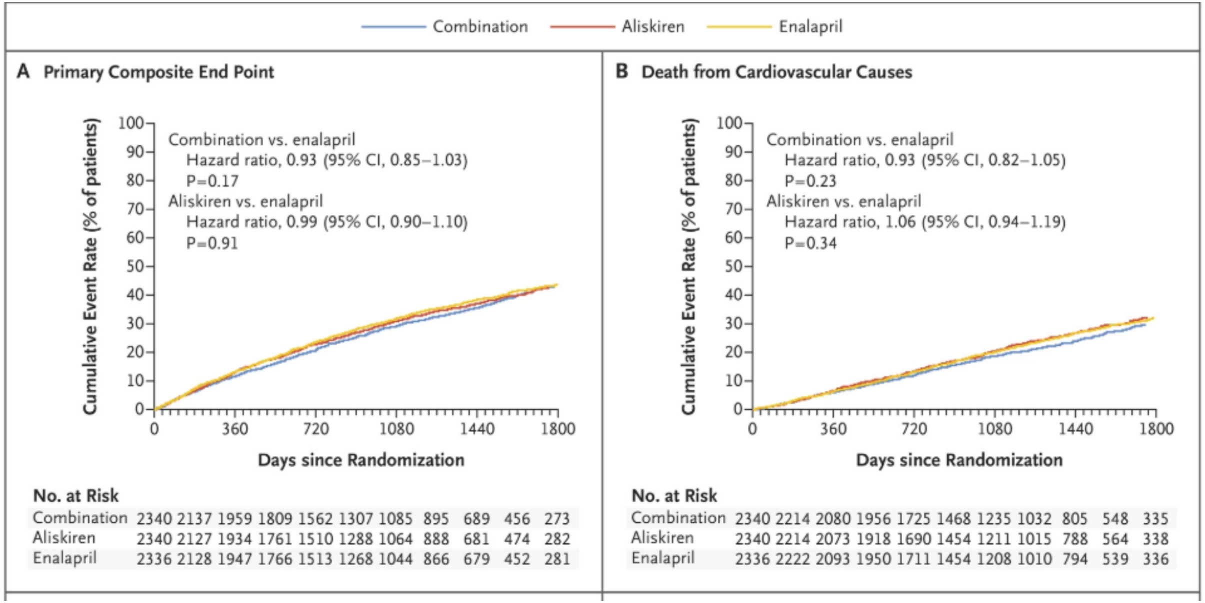
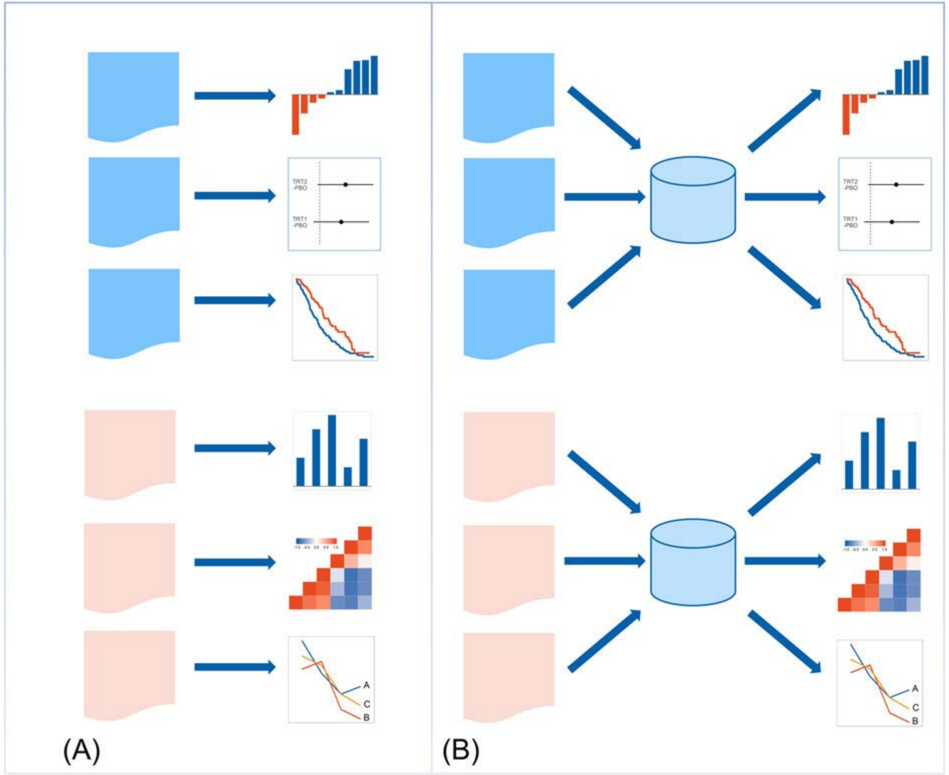
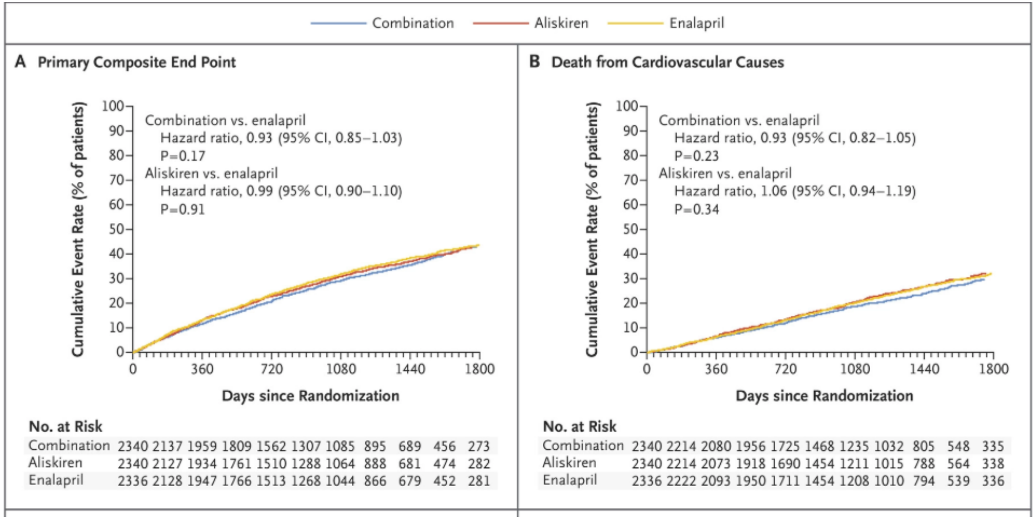


Table 2. Protocol-Specified Primary and Secondary Outcomes.*

Outcome	Combination Therapy (N=2340)	Aliskiren (N=2340)	Enalapril (N=2336)	Combination Therapy vs. Enalapril		Aliskiren vs. Enalapril	
				Hazard Ratio or Difference (95% CI)	P Value	Hazard Ratio or Difference (95% CI)	P Value
Primary composite outcome: death from cardiovascular causes or first hospitalization for worsening heart failure — no. (%)	770 (32.9)	791 (33.8)	808 (34.6)	0.93 (0.85 to 1.03)	0.17	0.99 (0.90 to 1.10)	0.91
Death from cardiovascular causes	512 (21.9)	562 (24.0)	547 (23.4)	0.93 (0.82 to 1.05)	0.23	1.06 (0.94 to 1.19)	0.34
First hospitalization for worsening heart failure	430 (18.4)	442 (18.9)	452 (19.3)	0.93 (0.82 to 1.06)	0.29	0.99 (0.87 to 1.13)	0.91
Secondary outcome: change in KCCQ clinical summary score at 12 mo†	−5.04±0.56	−6.03±0.57	−5.01±0.55	−0.03 (−1.56 to 1.50)	0.97	−1.02 (−2.56 to 0.52)	0.20
Other prespecified exploratory outcomes — no. (%)‡							
Death from cardiovascular causes, hospitalization for heart failure, nonfatal myocardial infarction, nonfatal stroke, or resuscitated cardiac arrest	841 (35.9)	874 (37.4)	877 (37.5)	0.94 (0.86 to 1.04)	0.23	1.01 (0.92 to 1.11)	0.80
Fatal or nonfatal myocardial infarction	88 (3.8)	84 (3.6)	100 (4.3)	0.87 (0.66 to 1.16)	0.36	0.85 (0.64 to 1.14)	0.28
Fatal or nonfatal stroke	87 (3.7)	103 (4.4)	93 (4.0)	0.93 (0.70 to 1.25)	0.65	1.12 (0.85 to 1.49)	0.42
First resuscitated cardiac arrest	31 (1.3)	35 (1.5)	32 (1.4)	0.96 (0.58 to 1.57)	0.86	1.10 (0.68 to 1.78)	0.69
Death from any cause	595 (25.4)	654 (27.9)	646 (27.7)	0.91 (0.82 to 1.02)	0.12	1.04 (0.93 to 1.16)	0.46
Composite renal outcome — no. (%)§	39 (1.7)	26 (1.1)	18 (0.8)	2.17 (1.24 to 3.79)	0.007	1.50 (0.82 to 2.74)	0.18

Analysis results data sets





UANALID	TRTVAR	TRTVAL	AVISITN	PARAMCD	ROWCAT1	ANLTYP1	ANLTYP2	STAT	STATVAL	ANLMETH
<STUDYID>_<RA>_XXX1	TRT01P	Combination		PCE	Combination vs. Enalapril	RESPONSE	Experimental	SMALLN	770	Lifetest KM
<STUDYID>_<RA>_XXX1	TRT01P	Combination		PCE	Combination vs. Enalapril	RESPONSE	Experimental	BIGN	2340	Lifetest KM
<STUDYID>_<RA>_XXX1	TRT01P	Combination		PCE	Combination vs. Enalapril	RESPONSE	Experimental	PERCENT	32.9	Lifetest KM
<STUDYID>_<RA>_XXX1	TRT01P	Enalapril		PCE	Combination vs. Enalapril	RESPONSE	Enalapril	SMALLN	808	Lifetest KM
<STUDYID>_<RA>_XXX1	TRT01P	Enalapril		PCE	Combination vs. Enalapril	RESPONSE	Enalapril	BIGN	2336	Lifetest KM
<STUDYID>_<RA>_XXX1	TRT01P	Enalapril		PCE	Combination vs. Enalapril	RESPONSE	Enalapril	PERCENT	34.6	Lifetest KM
<STUDYID>_<RA>_XXX1	TRT01P			PCE	Combination vs. Enalapril	COMPARISON		Hazard	0.93	Lifetest KM
<STUDYID>_<RA>_XXX1	TRT01P			PCE	Combination vs. Enalapril	COMPARISON		95CILOW	0.846	Lifetest KM
<STUDYID>_<RA>_XXX1	TRT01P			PCE	Combination vs. Enalapril	COMPARISON		95CIHIGH	1.03	Lifetest KM
<STUDYID>_<RA>_XXX1	TRT01P			PCE	Combination vs. Enalapril	COMPARISON		1sidedp	0.0862	Lifetest KM
<STUDYID>_<RA>_XXX1	TRT01P			PCE	Combination vs. Enalapril	COMPARISON		2sidedp	0.1724	Lifetest KM
<STUDYID>_<RA>_XXX1	TRT01P			PCE	Combination vs. Enalapril	COMPARISON		adj1sidep	0.3448	Lifetest KM
<STUDYID>_<RA>_XXX1	TRT01P	Combination		PCE	Non Diabetes: Combination vs Enalapril(1)	RESPONSE	Experimental	SMALLN		Lifetest KM
<STUDYID>_<RA>_XXX1	TRT01P	Combination		PCE	Non Diabetes: Combination vs Enalapril(1)	RESPONSE	Experimental	BIGN		Lifetest KM
<STUDYID>_<RA>_XXX1	TRT01P	Combination		PCE	Non Diabetes: Combination vs Enalapril(1)	RESPONSE	Experimental	PERCENT		Lifetest KM

Open challenges: communicating uncertainty

Communicating uncertainty about facts, numbers and science

Anne Marthe van der Bles, Sander van der Linden, Alexandra L. J. Freeman, James Mitchell, Ana B. Galvao, Lisa Zaval and David J. Spiegelhalter

Published: 08 May 2019 | https://doi.org/10.1098/rsos.181870

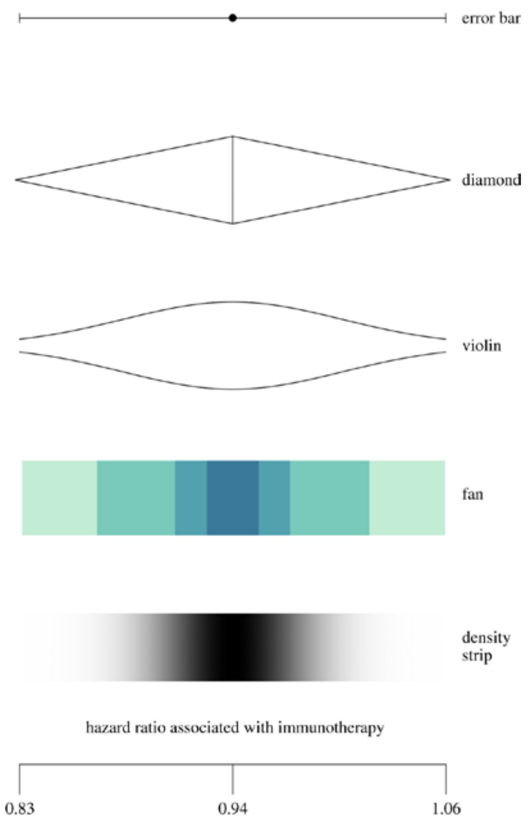
Editorial

Moving to a World Beyond "$p < 0.05$"

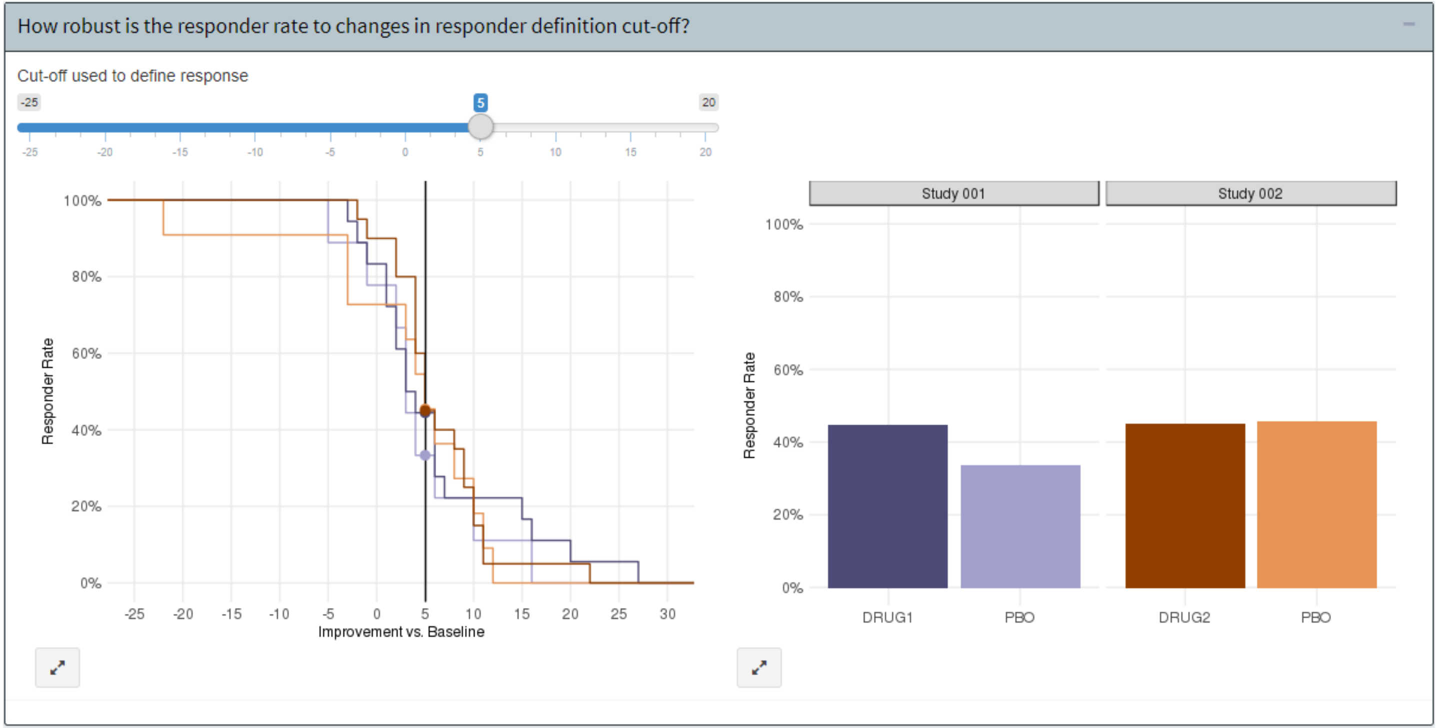
Ronald L. Wasserstein, Allen L. Schirm & Nicole A. Lazar

Pages 1-19 | Published online: 20 Mar 2019

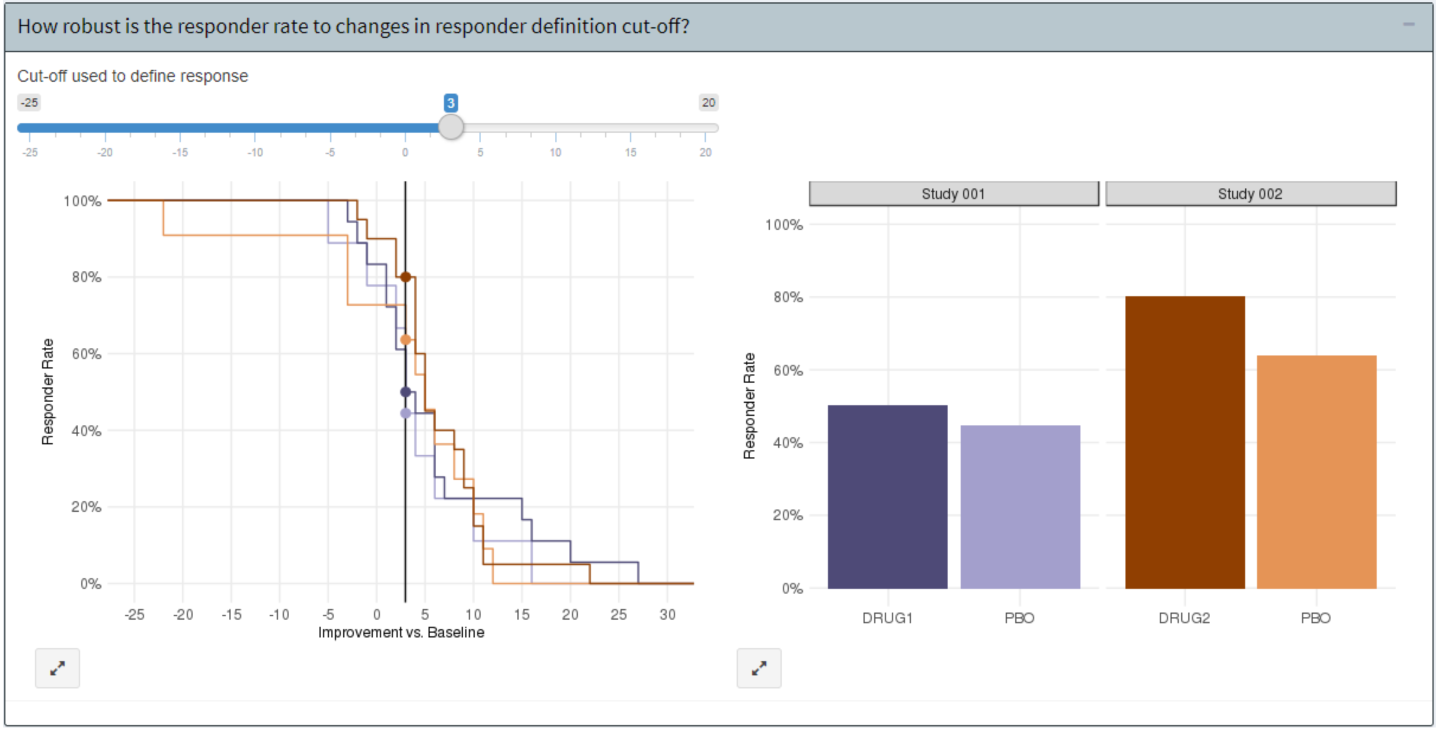
Download citation https://doi.org/10.1080/00031305.2019.1583913 Check for updates



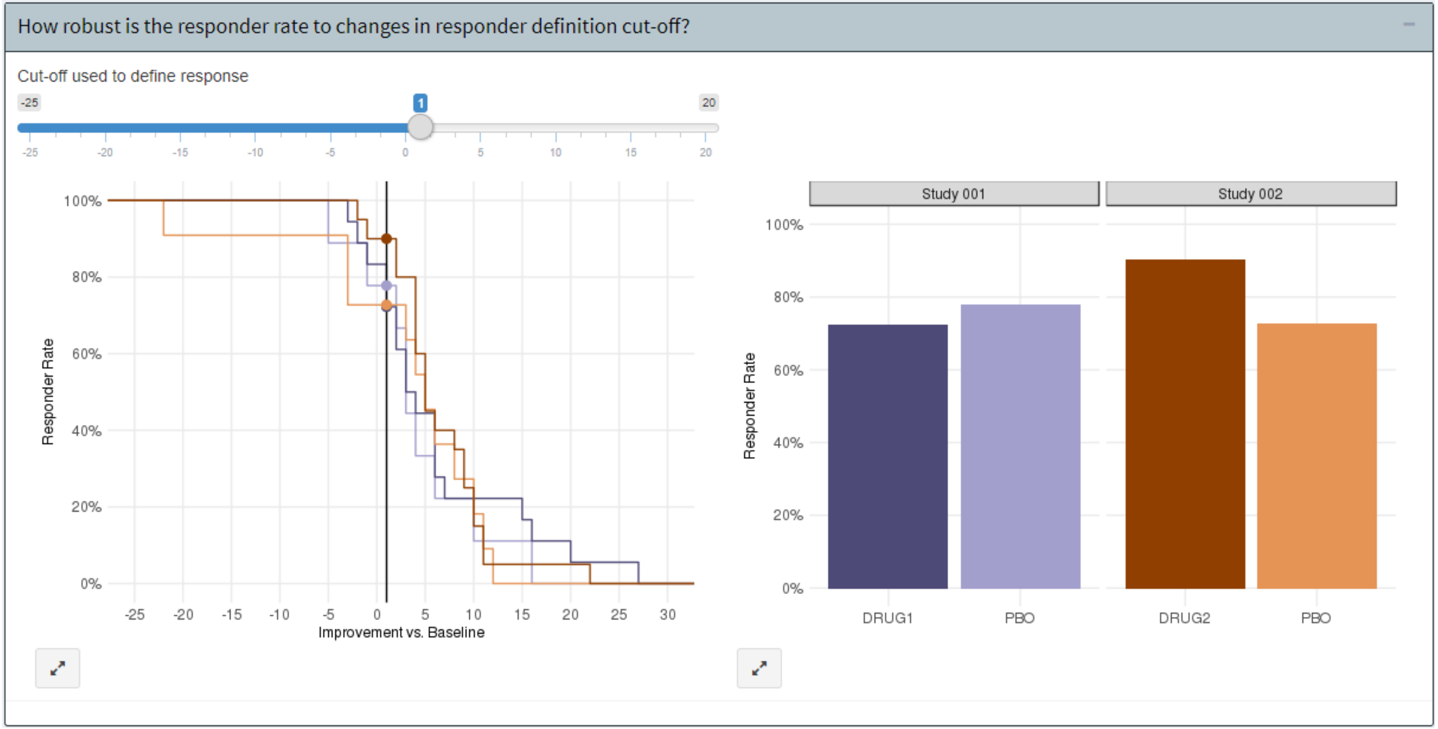
Analyse questions not data



Analyse questions not data



Analyse questions not data



Elements of a STRATOS VP initiative

Topic groups

1	Missing data
2	Selection of variables and functional forms in multivariable analysis
3	Initial data analysis
4	Measurement error and misclassification
5	Study design
6	Evaluating diagnostic tests and prediction models
7	Causal inference
8	Survival analysis
9	High-dimensional data

Effective data visualisation is effective visual communication

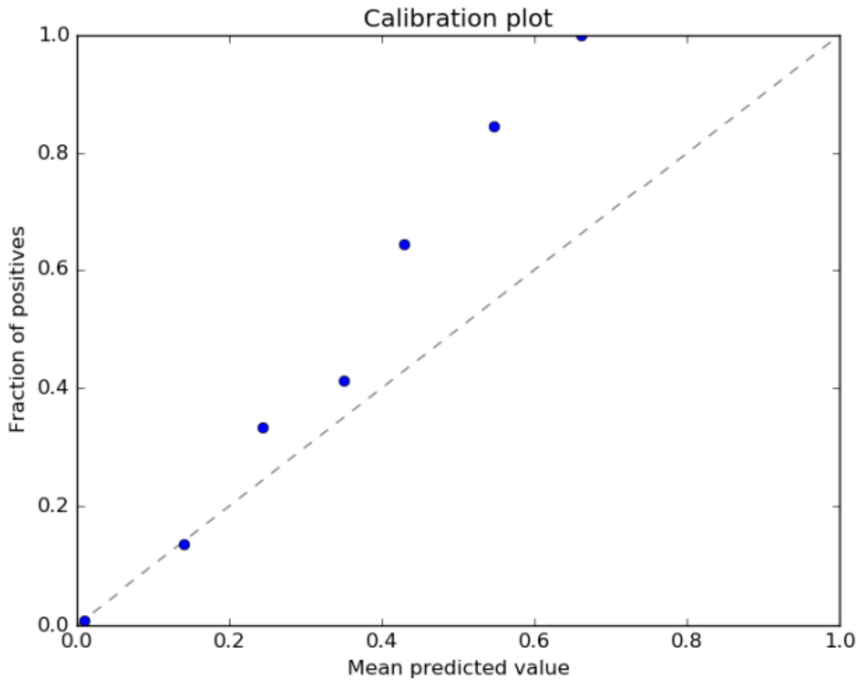
- Effective graphs...
  - are visually appealing, intuitive, legible
  - use the correct graph type and axis scales
  - use proximity & alignment to facilitate comparison
  - use labels and annotations to add clarity to the message
- Most importantly, effective use of visualisations
  - Enables clear and impactful communication
  - Elevates influence with stakeholders
  - Facilitates informed decision making

Thank you

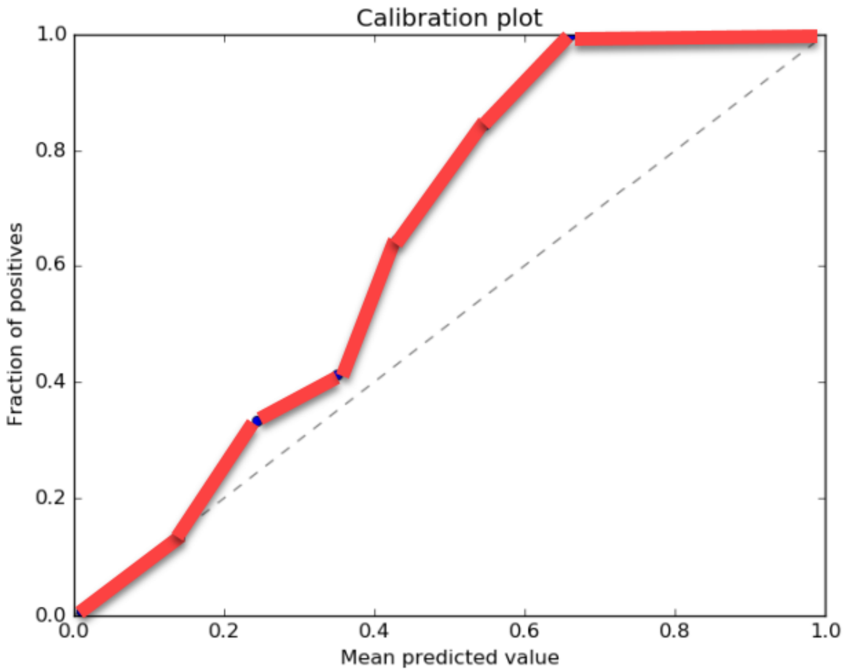
Acknowledgements

- Alison Margolskee
- Marc Vandemeulebroecke
- David Carr
- Andrew Wright
- Jürgen Löffler
- Linda Kanitra
- Baldur Magnusson
- Matthew Brierley
- David Ohlssen
- Shaun Butcher
- Julie Jones
- Walter Hufford
- Ruquan You
- Ivan-Toma Vranesic
- Ian Rees
- Nicolas Guerro
- Keo Chanthavinout
- Frank Bretz

NOVARTIS



(a) Calibration curve for inpatient mortality predicted at 24 hours into hospitalization for hospital A



(a) Calibration curve for inpatient mortality predicted at 24 hours into hospitalization for hospital A