PumasAI DeepPumas Bristol Myers Squibb®

Early go/no-go decisions in clinical trials

using a DeepNLME joint tumour growth dynamics and overall survival model

PumasAI Lorenzo Contento, Mohamed Tarek, Amit Roy

Bristol Myers Squibb Kiyoto A. Tanemura, Lu Chen, Chuanpu Hu, Anna G. Kondic

Underline = presenter

MOTIVATION

Build a tool to inform early go/no-go decisions

Reduce data and time without sacrificing confidence

WHY STOP EARLY

- clinical trials are expensive
- participants may fare worse than they would on the SoC

OBJECTIVE

Stop the trial **as early as possible**, to either:

- **abandon** a failing new drug
- **advance** a promising one

with confidence!

REQUIREMENTS

- **concretely evaluable**
  we can measure how well it works, so we can trust it
- **model-informed**
  leverages prior knowledge, reduces required data

OUR CASE · NSCLC

treatment arm		control arm
TX	vs	**CTX**
new drug		SoC (chemo)

Do subjects on TX survive longer than subjects on CTX?

AVAILABLE DATA

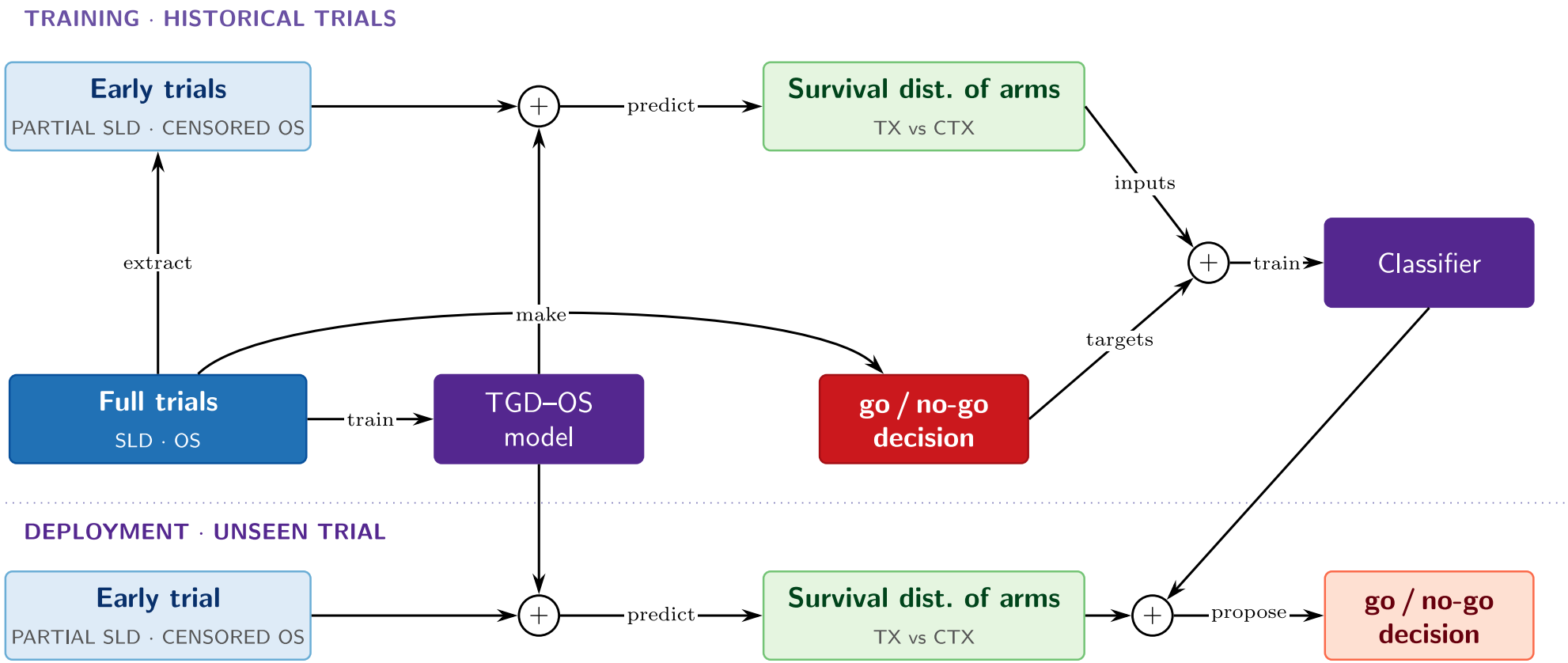
- **TGD** tumour growth dynamics
- **OS** overall survival (potentially censored)

in our case: **TGD** = **SLD** = sum of the longest diameters of all tumour lesions

TRAINING	EVALUATION
7 studies	2 studies
14 arms	TX & CTX arms
4677 subjects	decision labels available: one positive and one negative

DeepPumas

Pipeline for model-informed decisions on early-trial data



DeepPumas

Survival distribution for a subject

given observed data and covariates

INPUT · SUBJECT i

SLD	**OS**	**covariates**
as a function of time	potentially censored	baseline and/or time-dependent

OUTPUT · ITS SURVIVAL DISTRIBUTION S_i

conditional on the input data

or, even better,

OUTPUT WITH UQ · THE DISTRIBUTION OF S_i

conditional on the input data

captures the model's own uncertainty

ASSUMPTIONS

DISEASE-SPECIFIC

the **SLD** ⇔ **OS** relationship is likely to be highly disease-specific

extension: tumour type as a covariate,
to better capture patterns shared by different diseases

TREATMENT-AGNOSTIC

transfers to drugs absent from training

extension: MoA as a covariate (does not transfer to unseen MoAs)

SLD alone cannot determine OS
the model will be biased

but

our task is classification
bias can be tolerated if the classifier works

Data-driven TGD NLME model

built with DeepPumas

learning the space of all possible TGD trajectories

DYNAMICS

$$\mathrm{SLD}'(t) = \mathrm{NN}(\mathrm{SLD}(t), t; \theta, \eta)$$

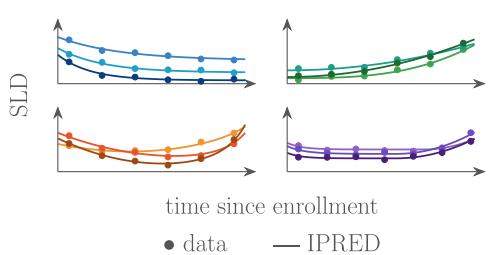
where

- NN neural network: last-layer activation keeps SLD positive
- θ population parameters
- η random effects: individualize the NN output per subject

EFFICIENT TWO-STAGE FIT

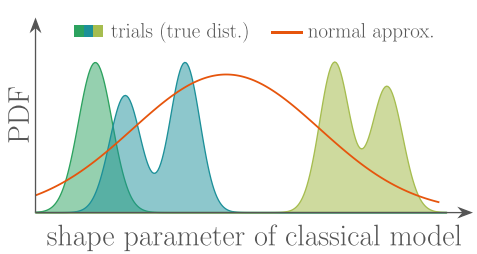
1. fit by maximizing the joint likelihood of θ and η
2. fit the distribution of η with a **normalizing flow**

A SINGLE DeepPumas MODEL HANDLES MULTIPLE STUDIES



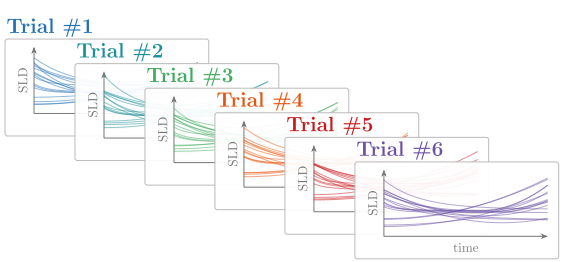
all SLD trajectories can be described

neural networks are universal approximators



no overestimation of IIV

unlike classical models with Gaussian REs, especially across studies



fit multiple studies with thousands of subjects

thanks to the efficient two-stage fit

DeepPumas

A proportional hazards OS model

built with DeepPumas

using covariates, SLD, and data-driven features

HAZARD

$$\lambda(t) = \lambda_0(t)\exp(\beta_c\, c(t) + \beta_{\text{SLD}}\, F_{\text{SLD}}(t) + \beta_{\text{NN}}\, \text{NN}_{\text{OS}}(c(t), F_{\text{SLD}}(t)))$$

where

λ	hazard function
λ_0	log-logistic baseline hazard
c	baseline + time-varying covariates
F_{SLD}	local SLD-derived features: SLD, relative change from baseline, derivative
NN_{OS}	neural network capturing residual nonlinearity

The TGD and OS submodels are trained and evaluated **sequentially**

COMPARISON WITH A CLASSICAL MIXTURE NLME TGD-OS MODEL

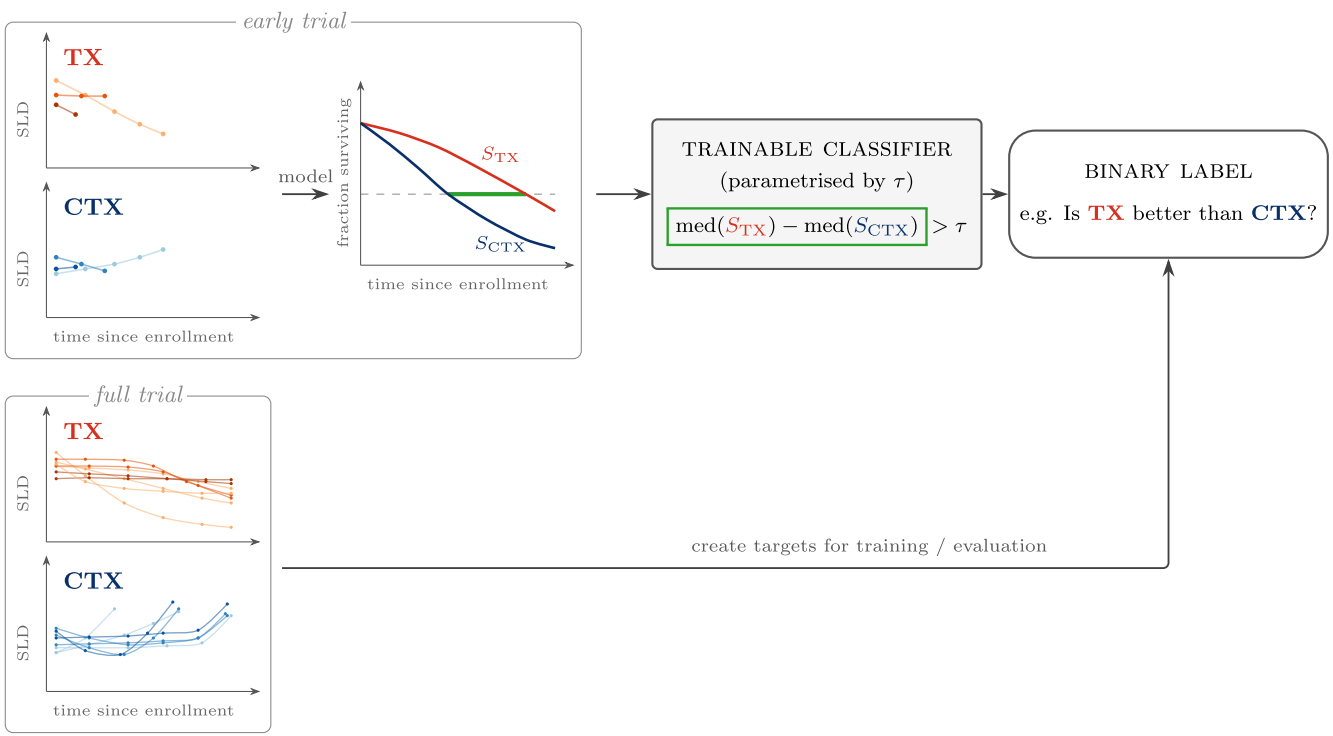
✓ **the DeepPumas model has better OS log-likelihood**
on both training and test data

~ **both models show stronger bias in chemotherapy arms**
potential ways to reduce it: more data, more expressive hazard, …

CLASSIFIER

Discriminating trial arms early

using model-predicted survival: a very simple example



CLASSIFIER

INPUT · $S_{\mathrm{TX}}, S_{\mathrm{CTX}}$

model-predicted survival distributions based on early-trial data

RULE

compare two scalar statistics of survival

τ: classifier threshold
trainable

OUTPUT · **IS TX BETTER THAN CTX?**

decision that would have been made had the full trial been available

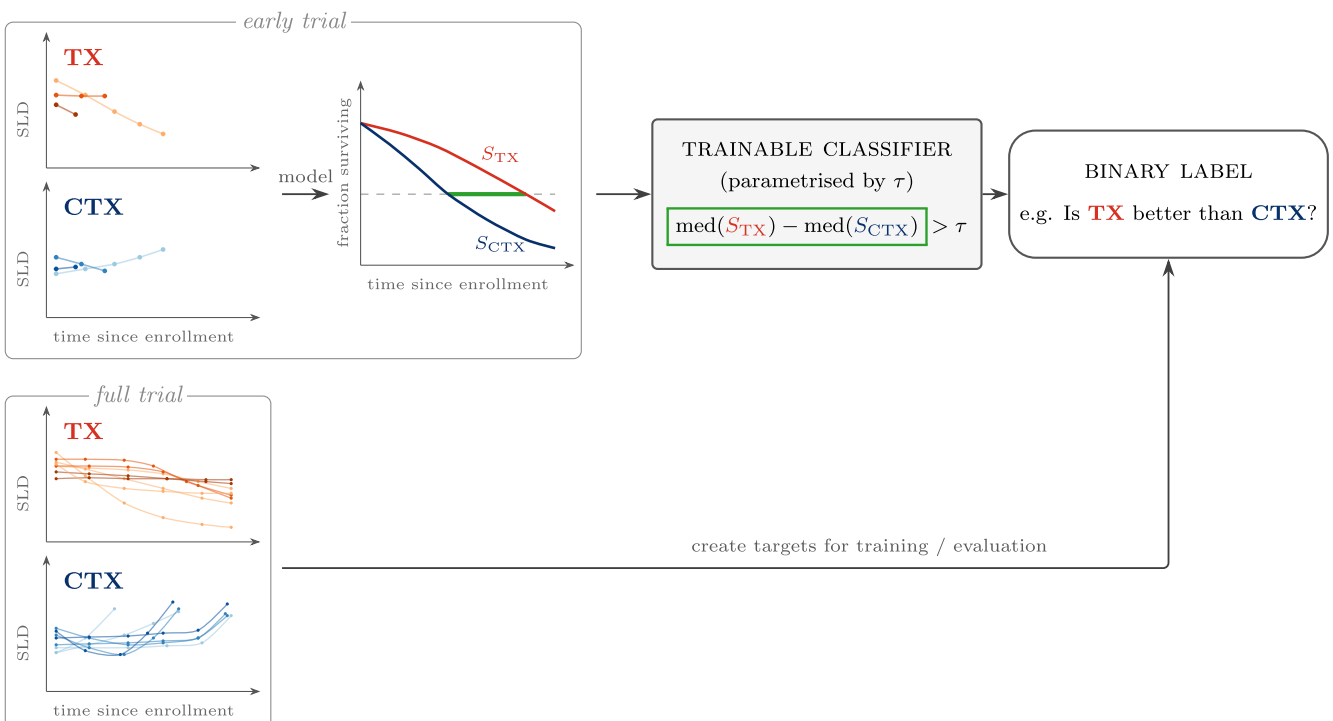
annotated or derived by a rule

DeepPumas

CLASSIFIER

Discriminating trial arms early

using model-predicted survival: a very simple example



TRAINING THE CLASSIFIER

tune τ to trade off

false negatives ↔ false positives

e.g. for go/no-go we do not want to abandon unless we are sure the new treatment will fail

→ we want low FNR

→ τ must be small (potentially negative)

BEYOND A SINGLE SCALAR METRIC

- multiple metrics combined
- full distributional comparisons

1 sample

=

1 pair of an **early trial** +

the **full trial it would have evolved into**

FNR = False Negative Rate

DeepPumas

What is the definition of an “early trial”?

It has to match the early trials we encounter at deployment

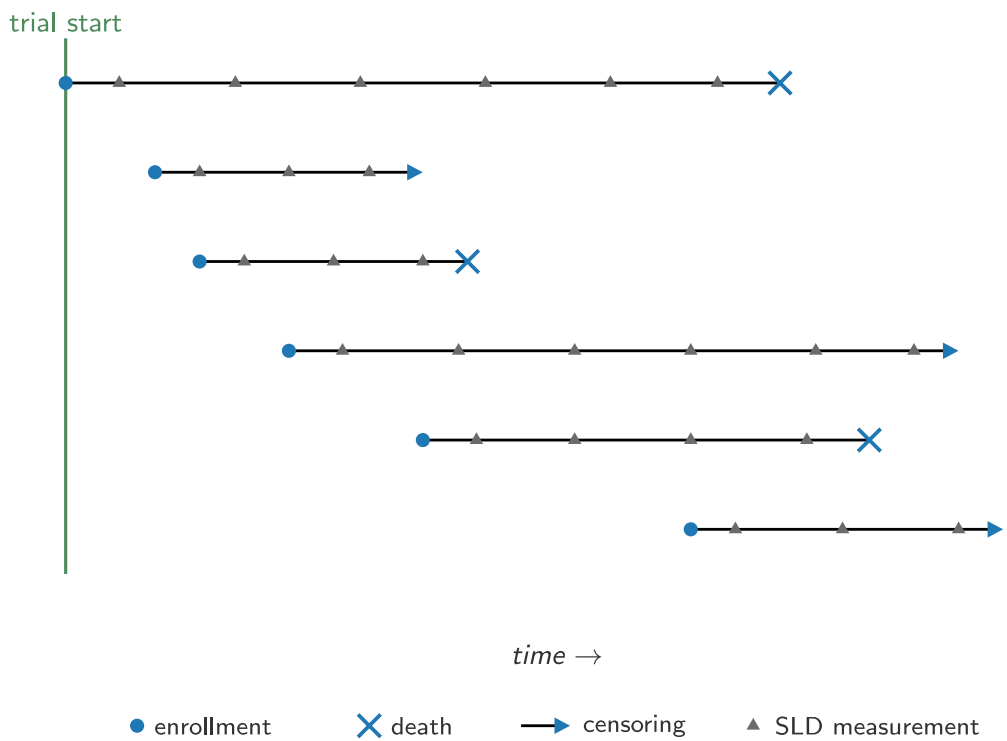
WHEN SHOULD WE STOP AND CALL THE CLASSIFIER?

We must choose a **stopping criterion**, e.g.:

- total trial duration
- number of subjects enrolled
- number of subjects with a given follow-up duration
- …

EXTRACT AN EARLY TRIAL FROM A FULL TRIAL

1. Pick a full trial



DeepPumas

What is the definition of an “early trial”?

It has to match the early trials we encounter at deployment

WHEN SHOULD WE STOP AND CALL THE CLASSIFIER?

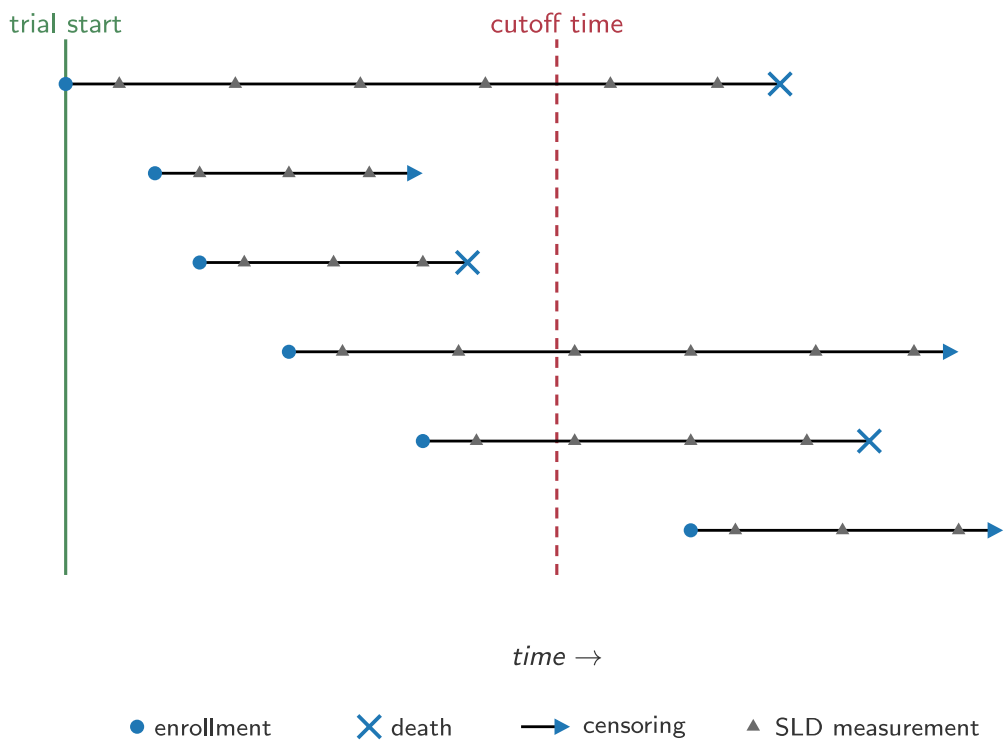
We must choose a **stopping criterion**, e.g.:

- total trial duration
- number of subjects enrolled
- number of subjects with a given follow-up duration
- …

EXTRACT AN EARLY TRIAL FROM A FULL TRIAL

1. Pick a full trial
2. Find the cutoff: earliest moment the criterion is met

Make the data match what was available at the cutoff



DeepPumas

What is the definition of an “early trial”?

It has to match the early trials we encounter at deployment

WHEN SHOULD WE STOP AND CALL THE CLASSIFIER?

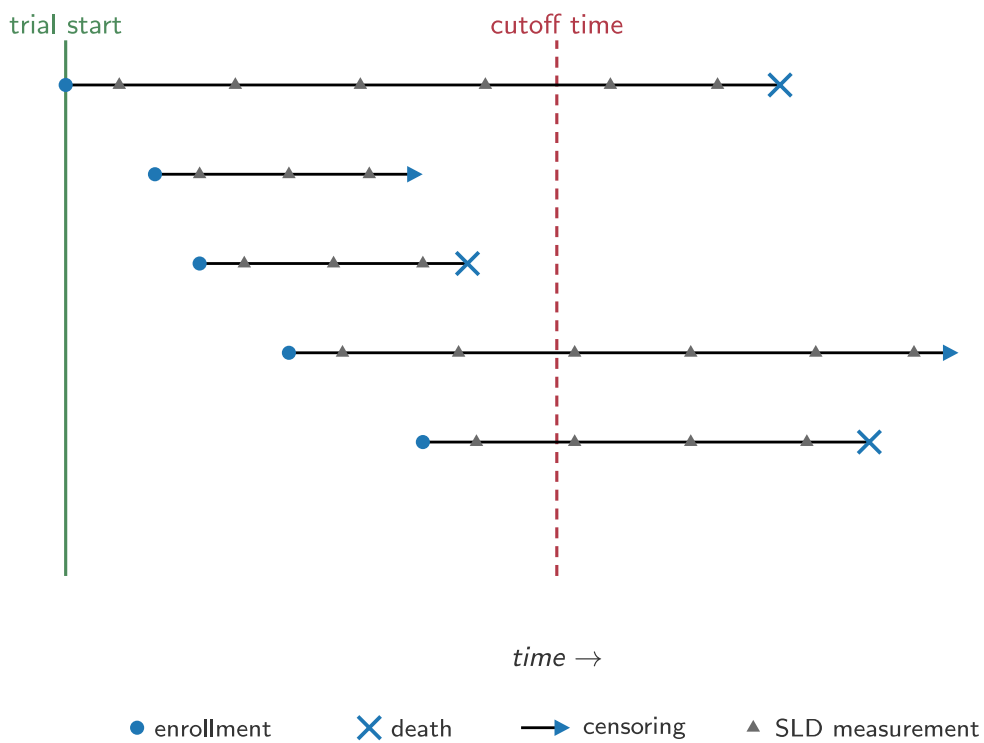
We must choose a **stopping criterion**, e.g.:

- total trial duration
- number of subjects enrolled
- number of subjects with a given follow-up duration
- …

EXTRACT AN EARLY TRIAL FROM A FULL TRIAL

1. Pick a full trial
2. Find the cutoff: earliest moment the criterion is met

   Make the data match what was available at the cutoff
3. Drop subjects enrolled after the cutoff



DeepPumas

What is the definition of an “early trial”?

It has to match the early trials we encounter at deployment

WHEN SHOULD WE STOP AND CALL THE CLASSIFIER?

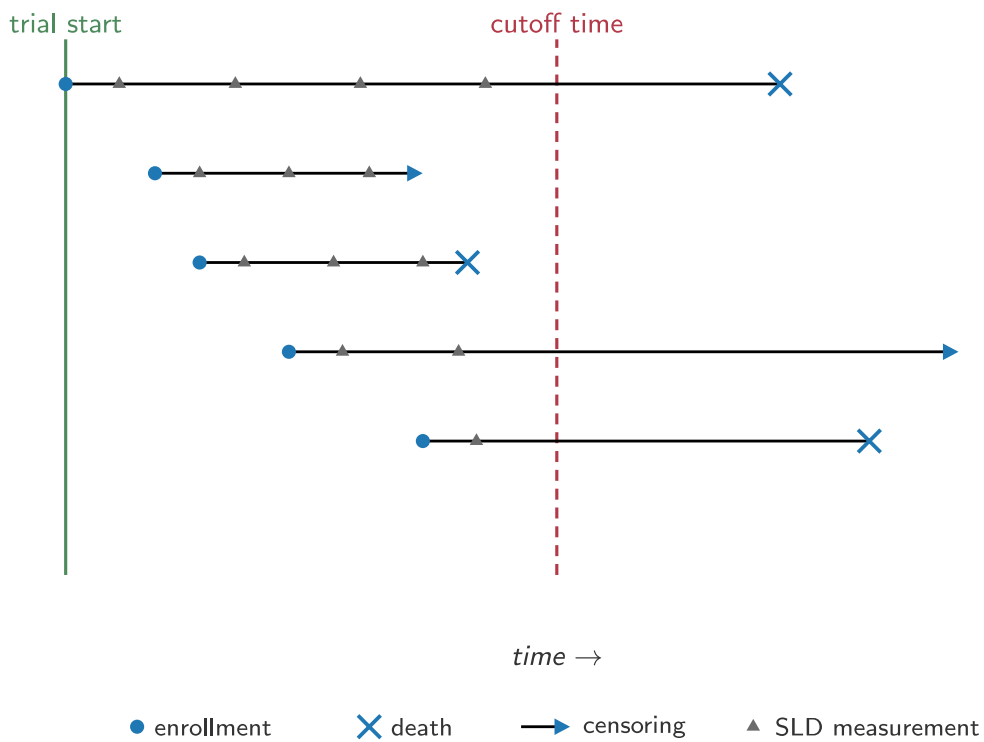
We must choose a **stopping criterion**, e.g.:

- total trial duration
- number of subjects enrolled
- number of subjects with a given follow-up duration
- …

EXTRACT AN EARLY TRIAL FROM A FULL TRIAL

1. Pick a full trial
2. Find the cutoff: earliest moment the criterion is met

   Make the data match what was available at the cutoff
3. Drop subjects enrolled after the cutoff
4. Drop SLD observations after the cutoff



DeepPumas

What is the definition of an “early trial”?

It has to match the early trials we encounter at deployment

WHEN SHOULD WE STOP AND CALL THE CLASSIFIER?

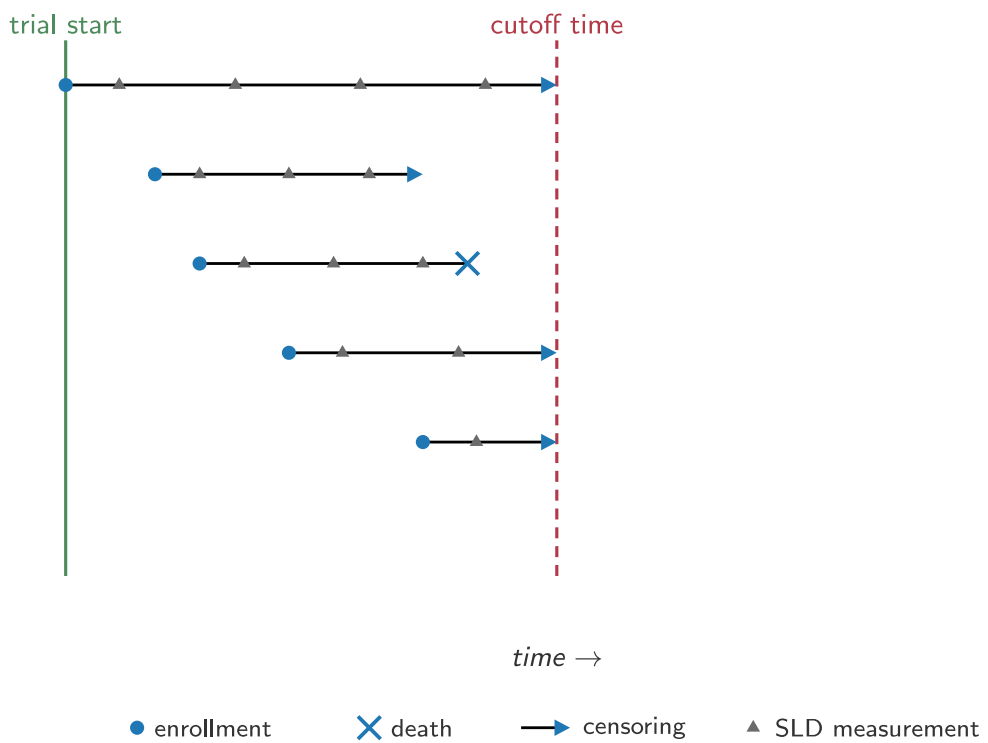
We must choose a **stopping criterion**, e.g.:

- total trial duration
- number of subjects enrolled
- number of subjects with a given follow-up duration
- …

EXTRACT AN EARLY TRIAL FROM A FULL TRIAL

1. Pick a full trial
2. Find the cutoff: earliest moment the criterion is met

   Make the data match what was available at the cutoff
3. Drop subjects enrolled after the cutoff
4. Drop SLD observations after the cutoff
5. Censor at the cutoff any subject still alive at that time



DeepPumas

Not enough data to train the classifier

TWO BOTTLENECKS IN THE TRAINING DATA FOR THE CLASSIFIER

AVAILABLE TRIALS ARE FEW

trials used to train the TGD-OS model should ideally not be reused:
risk of overestimating the performance of the classifier

1 TRIAL → 1 SAMPLE

with the stopping criterion fixed,
each full trial yields exactly one early trial

the number of training samples is very small: **not enough to tune even a scalar τ!**

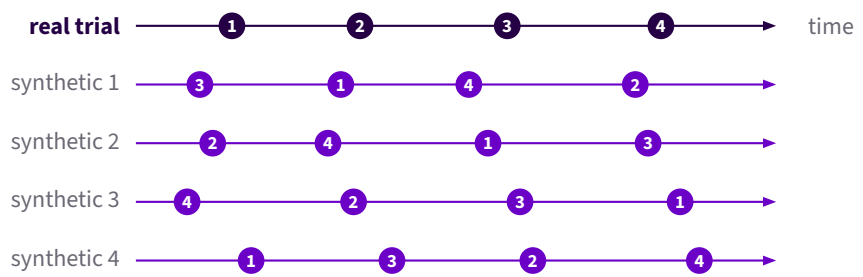
Augmenting the classifier dataset with synthetic trials

Replay subjects through alternative enrollment schedules

IDEA only one trial **actually** happened, but many **plausibly could have**

SYNTHETIC EARLY TRIAL GENERATION

1. **resample subject enrollment times**
   → *how the **full** trial may have looked*
2. **apply the same truncation procedure as before**
   → *how the **early** trial may have looked*



one full trial → **hundreds** of synthetic early trials

enough to calibrate the classifier

A SORT OF DATA AUGMENTATION, NOT A SYNTHETIC DATASET

Enrollment schedule	**Measurements**	**No synthetic subjects**
resampled	real, only truncated at a different cutoff	not generated by any model

DeepPumas

How do we resample an enrollment schedule?

Sample a new subject order, then new inter-arrival times

DECOMPOSING THE DISTRIBUTION OF ENROLLMENT SCHEDULES

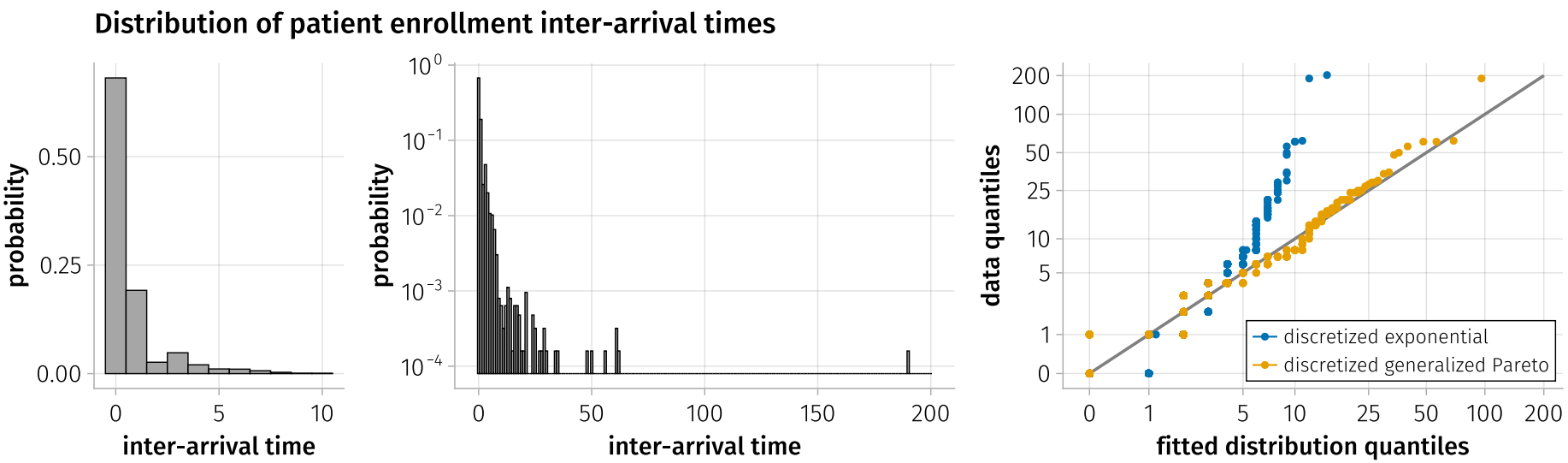
1 **SUBJECT ORDER**
uniform random permutation
*assumption: already-enrolled subjects do not influence **which** subjects enroll next*

2 **INTER-ARRIVAL TIMES (IAT)**
parametric IAT distribution
assumption: IAT independent of the current trial state

both plausible in a randomized trial — **and they only need to hold in the early phase**

Choosing a family for the IATs

Compare exponential vs generalized Pareto by fitting to the IATs from all full trials in the training set



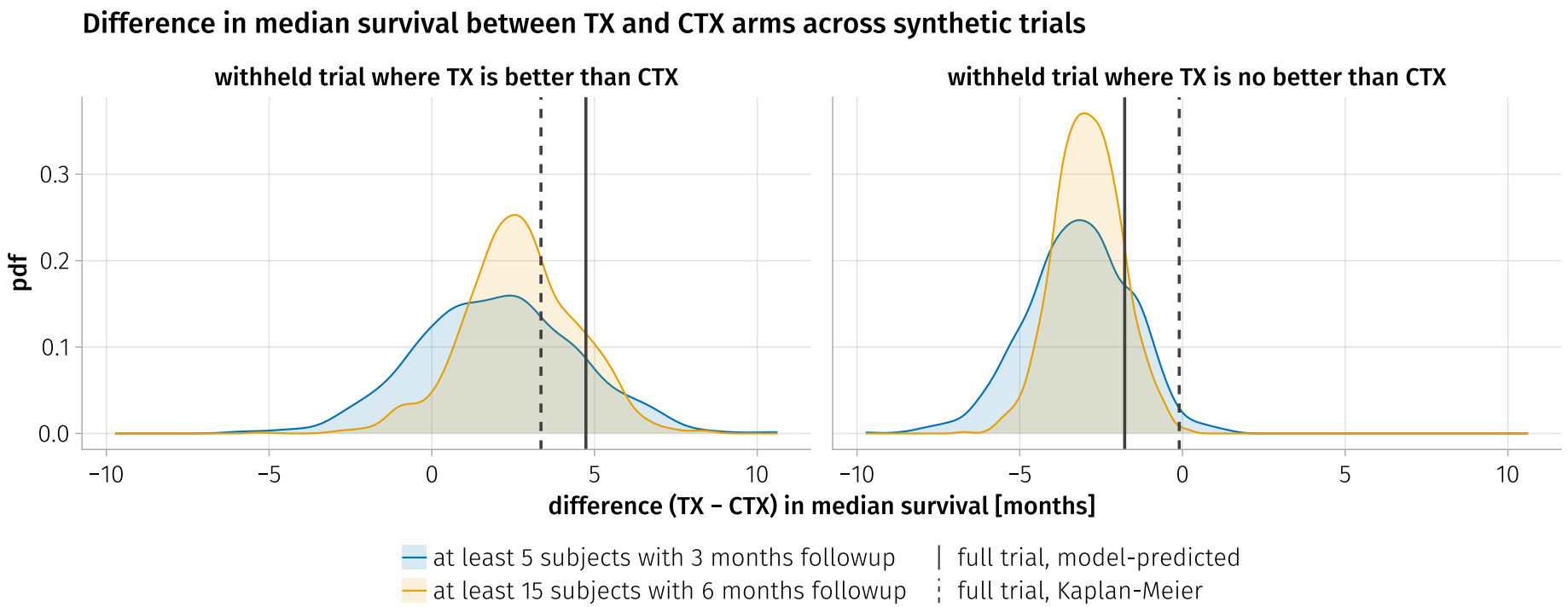
IAT = 0 has high probability: subjects tend to enroll in groups?

Generalized Pareto performs much better on the **heavy tails**

DeepPumas

The score discriminates go from no-go trials

even if survival predictions are biased



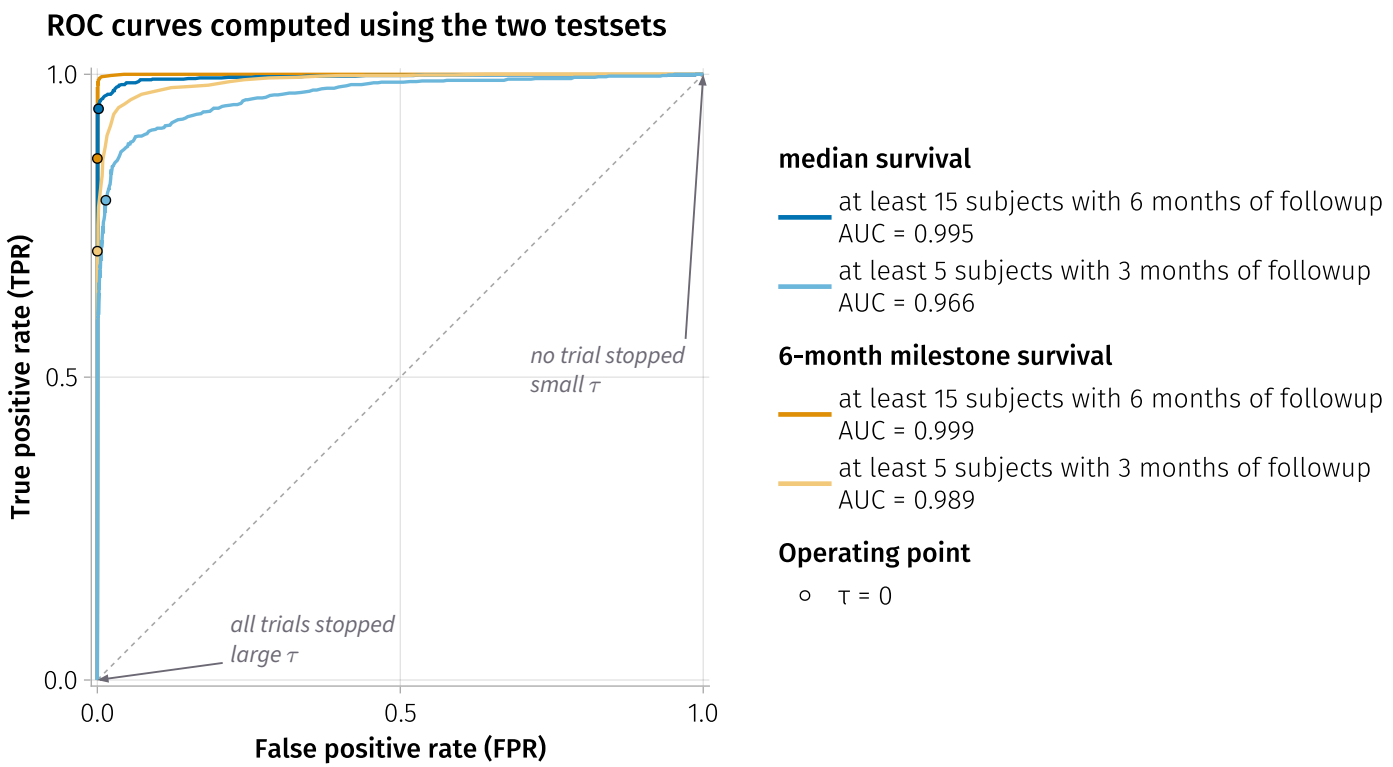
Clean separation — mass on the correct side of zero

More data, sharper signal — more confident prediction

DeepPumas

ROC curves are good across metrics and stopping criteria

big caveat: only two full trials in the classifier training set!



FNR = 1 − TPR = fraction of potentially successful trials we mistakenly stop

DeepPumas

Reliable early go/no-go decisions in NSCLC trials

Model-agnostic framework based on data augmentation by synthetic trials

TAKEAWAYS

1. **Synthetic-trial augmentation**
   one full trial → hundreds of plausible early ones
2. **Real subjects, no synthetic patients**
   enrollment schedules resampled; measurements stay real
3. **Tune τ and estimate error rates**
   from a small handful of full trials
4. **Model-agnostic pipeline**
   any joint TGD-OS model plugs in
5. **Good ROC curves** (with caveat)
   on held-out NSCLC, with the DeepPumas TGD-OS model

OUTLOOK

1. Evaluate the classifier on more test trials
2. Try more expressive OS models (e.g. UDE/NN) to reduce bias
3. Apply the framework to other tumour types
4. Propagate EBE uncertainty into the survival distribution
5. Use historical SoC data in place of a same-trial control arm
6. Package the workflow for practitioner use (e.g. web app)

THANK YOU FOR LISTENING

WEB pumas.ai
PUBLICATIONS pumas.ai/resources/publications

DeepPumas