Determination of a sample size necessary to identify a predictive signature in cancer survival

L. Villain\textsuperscript{1,2,3*}, R. Thiébaut\textsuperscript{1,2,3}, A. Bikfalvi\textsuperscript{4}, B. Hejblum\textsuperscript{1,2,3}

\textsuperscript{1}Univ. Bordeaux, Inserm, Bordeaux Population Health Research Center, UMR 1219, Bordeaux, France
\textsuperscript{2}INRIA Bordeaux Sud Ouest, SISTM team, Talence, France
\textsuperscript{3}Vaccine Research Institute, Créteil, France
\textsuperscript{4}INSERM U1029, Angiogenesis and Cancer Microenvironment Laboratory, Bordeaux, France

October, 10\textsuperscript{th}
Gliomas are the most common primary brain tumors

*GliomaPRD objective:* predict the progression of Lower Grade Gliomas (LGGs)
Gliomas are neuroepithelial tumours from the supporting glial cells of the central nervous system (CNS).

They are derived from the astrocytic or oligodendrocytic lineage.

They are divided into high-grade and low-grade gliomas.
High-grade gliomas (HGG) exhibit neoangiogenesis, necrosis and invasion at tumor borders and are mainly IDH1/2 wild-type.

Low grade gliomas (LGG) have slow infiltrative growth, absence of pronounced neoangiogenesis and are generally IDH1/2 mutated.

HGG have short survival whereas LGG survive longer, albeit some LGG patients exhibit a more rapid recurrence after surgery.
Data: RNA-seq from the tumor: N patients with different PET-SCAN outcome (cold/diffuse)

Objective of the project: predict survival of patient based on RNA-seq and imaging data
Problematic: Sample size calculation

How many subjects to identify a predictive signature of genes involved in the patients survival?
Sample size studies: general method

Pilot data

- Generation of simulated data
  - Generation of N patients and their variable of interest (eg: genes)
  - Generation of the outcome of interest, depends on the chosen amplitude of effect

Detection of genes
  - Detection of the variables having an effect
  - Calcul of the statistical power: percentage of correctly detected variables

Repeat for multiple N size and amplitude of effect
The structure of data can change the analysis and the statistical power.

Pilot data: simulation of similar data and adequate method → more accurate sample size study.
Pilot data: Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas

Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas

The Cancer Genome Atlas Research Network*
Pilot data from TCGA

- 278 adults patients having Lower Grade Gliomas of types II and III
- Gene expression of 20000 genes
Pilot data from TCGA

- 278 patients × 20000 genes
- RNA-seq counts highly correlated
- Clinical outcome not available
- [The Cancer Genome Atlas Research Network, NEJM, 2015] identified a signature of 1500 genes associated with LGG survival times
Sample Size Considerations of Prediction-Validation Methods in High-Dimensional Data for Survival Outcomes

Herbert Pang and Sin-Ho Jung*

Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, North Carolina

Received 24 September 2012; Revised 21 January 2013; accepted revised manuscript 9 February 2013.
Published online 7 March 2013 in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/gepi.21721
Survival times simulations [Pang and Jung, 2013]

- Generation of the RNA-seq count for N simulated patients on $p = 20000$ genes.

- Pilot data $\{(x_{k1}, \ldots, x_{kp}), k = 1, \ldots, n\}$ on the $n = 278$ observed patients

- We generate
  \[ z_{ij} = (x_{k'j} - \bar{x}_j)\epsilon_i / s_j \]
  with $k'$ a random number from $(1, \ldots, n)$, $\epsilon \sim \mathcal{N}(0, 1)$, $\bar{x}_j$ and $s_j$ respectively the mean and sd of the gene $j$
Pilot data is used to respect the correlation structure

- $\hat{\Sigma}_{p\times p}$ from $z_{ij}$, with $i \in [1, N]$ and $j \in [1, p]$

- Generation of $Z'_i \sim MVN(0_p, \hat{\Sigma}_{p\times p})$

⚠️ For computational reason, we split into batches of 1000 genes

→ To generate the signal, we create a special batch of 1500 genes
Survival times simulations

We generate the $T_i$ times of survival:

$$T_i = -\frac{\log(U)}{l_0 \exp(\beta^T Z_i)}$$

- $U \sim U(0, 1)$
- $l_0$ scale parameter determined from the survival of the patients in the pilot data
- $\beta^T$ the vector of effects being 0 except in the special batch
Once the times $T_i$ are generated, we can apply an univariate Cox model to each gene $j$:

$$
\lambda(t, Z_j') = \lambda_0(t) \exp(\beta_j Z_j')
$$

The gene $j$ is detected if the $\beta_j$ is statistically different from 0.

Statistical power is then calculated with the proportion of correctly detected genes.
Sample size study realised

Pilot data

Generation of simulated data

- Generation of N patients and their RNA-seq counts based on pilot data
- Generation of survival time based on the RNA-seq counts, depends on the chosen amplitude of effect (Cox formula)

Detection of genes

- Detection of the genes having an effect: univariate Cox model
- Calcul of the statistical power: percentage of correctly detected genes

Repeat for multiple N size and amplitude of effect
Results of the simulation study

Statistical power vs Sample size for different Relative risk per gene values.
1. The high correlation implies that:
   - Even a **small relative risk is detected** with a satisfying statistical power
   - With 50 patients, we can detect more than 50% of the selected genes, but even by increasing the relative risk or the sample size, we can not achieve more than 75% statistical power

2. A statistical power of 50% means detecting **at least 750 genes** which is **good enough for a predictive goal**.

3. We might improve the statistical power by **grouping the genes by pathways**.
Having a robust method for gene expression analysis

Data: 21 Breast cancer studies to predict tumor grade from gene expression
Genesets from their transcription factors: 593 sets (from MSigDB)

Identify the genesets differently expressed between tumor grade 1 and 3

Method is applied to the 21 datasets to evaluate the consistency of the selected genesets
Geneset enrichment principle

- Adapting the method to our case: score in **survival context**

Figure adapted from [Gaynor et al, 2019]
50 patients is a good sample size to detect enough gene for good prediction performance.

Increasing the sample size would not imply a huge improvement of the statistical power, thus it is not necessary.

The high correlation between the genes suggest the use of method using pathways

By formulating a score value in a survival context, we could adapt the GBJ method for our purpose

It might be difficult to formulate a score that takes into account the correlation between genes → Berk-Jones method
GBJ method: model and score

- Logit model for \( \mu_i = E(Y_i|G_i) \), with \( Y_i \) the tumor grade of the patient \( i \) and \( G_i \) the gene expression of the \( d \) genes of a given geneset:

\[
\log\left(\frac{\mu_i}{1 - \mu_i}\right) = \alpha_0 + G_i^T \beta, \quad i = 1, \ldots, n
\]

- \( H_0 : \beta = 0_{d \times 1} \)
GBJ method: model and score

- Logit model for $\mu_i = E(Y_i|G_i)$, with $Y_i$ the tumor grade of the patient $i$ and $G_i$ the gene expression of the $d$ genes of a given geneset:

  $$\log\left(\frac{\mu_i}{1-\mu_i}\right) = \alpha_0 + G_i^T \beta, \ i = 1, \ldots, n$$

- $H_0: \beta = 0_{d \times 1}$

- The score value $Z_j$ of each gene is calculated:

  $$Z = \frac{\sum_{i=1}^{n} G_{ij}(Y_i - \bar{Y})}{\sqrt{(\bar{Y}(1-\bar{Y})} \sqrt{\sum_{i=1}^{n} G_{ij}^2 - (\sum_{i=1}^{n} G_{ij})^2 / n}}$$

  With $Z \sim MVN(0, \Sigma)$. The absolute values $|Z|_{(j)}$ are ordered, with $|Z|_{(1)}$ being the smallest value.
A threshold function is defined: \( S(t) = \sum_{j=1}^{d} \mathbb{1} \{ |Z|_{(j)} \geq t \} \)

When the \( Z \) are independents, then \( S \sim \text{Bin}(d, \pi) \) with \( \pi = 2\phi(t) \)

When the \( Z \) are not independent, \( S(t) \) can be approximated by an Extended Beta-Binomial random variable [Sun and Lin, 2019]
GBJ method: statistic value

- The value of the GBJ statistic is then calculated:

\[
\text{GBJ}_d = \max_{1 \leq j \leq d/2} \log \left( \frac{\Pr \left\{ S \left( |Z|_{d-j+1} \right) = j \bigg| E(Z) = \hat{\mu}_{j,d} \cdot \mathbf{J}_d, \text{cov}(Z) = \Sigma \right\}}{\Pr \left\{ S \left( |Z|_{d-j+1} \right) = j \bigg| E(Z) = 0 \cdot \mathbf{J}_d, \text{cov}(Z) = \Sigma \right\}} \right).
\]

With \( \mathbf{J}^T_d = (1, 1, ..., 1)_{d \times 1} \) and \( \hat{\mu}_{j,d} > 0 \) solving the equation:

\[
\frac{j}{d} = 1 - \left\{ \phi(|Z|_{d-j+1} - \hat{\mu}_{j,d}) - \phi(-|Z|_{d-j+1} - \hat{\mu}_{j,d}) \right\}
\]

- It can be seen as "a maximum of a set of likelihood ratio tests performed on \( S(t) \) at all observed test statistic magnitudes greater than of equal to the median observed magnitude"