“The Model Statistician”, Insights into model selection techniques in Clinical Trials

Peter Williams

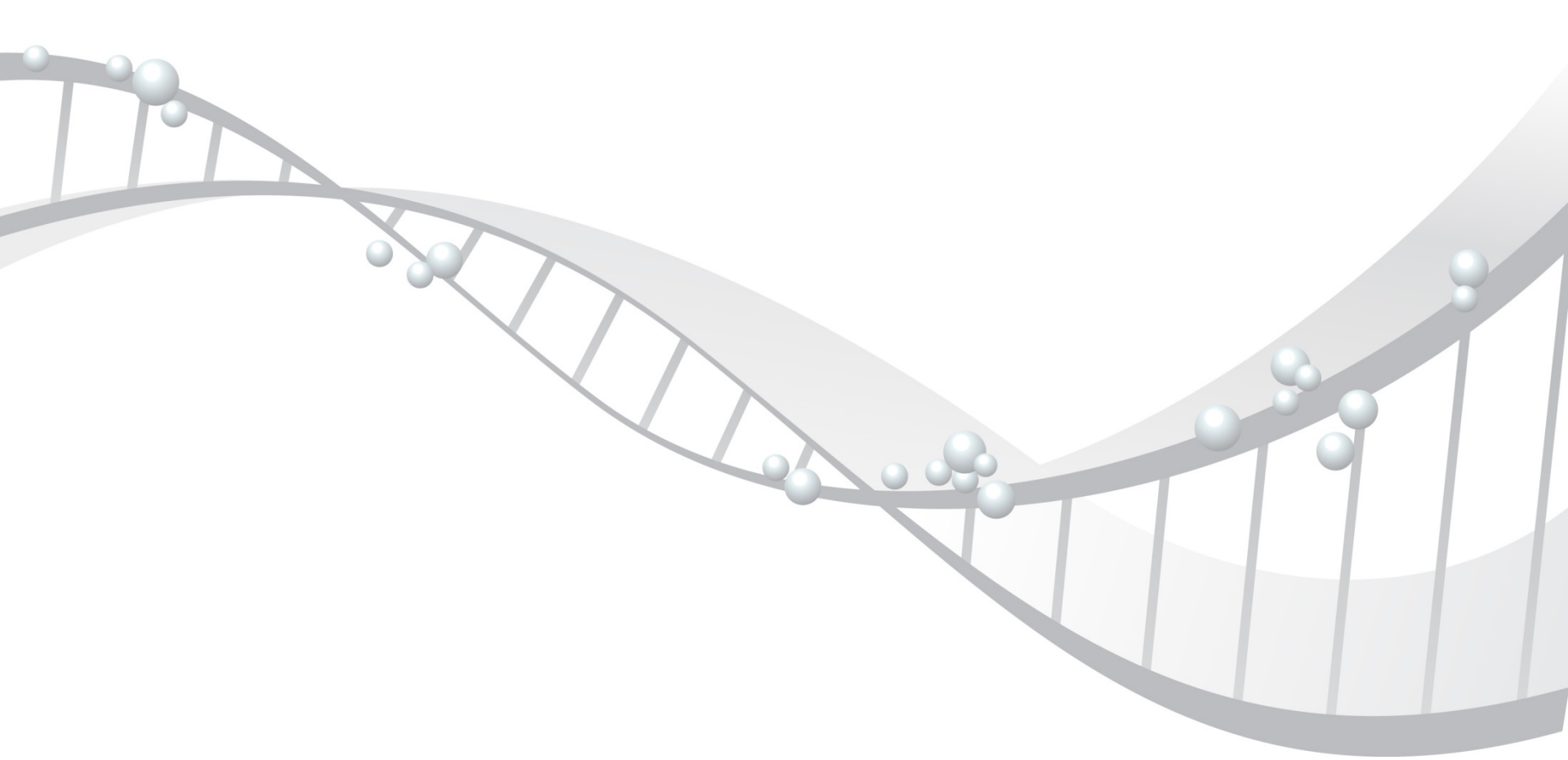
Veramed Limited

10th October 2017

veramed

Agenda

- A brief introduction to:
  - Statistical models in clinical trails
  - Model selection
- A simple motivating example
- Model selection techniques available in SAS
- When is model selection useful in the Clinical Trials setting?
- Conclusions/Discussion



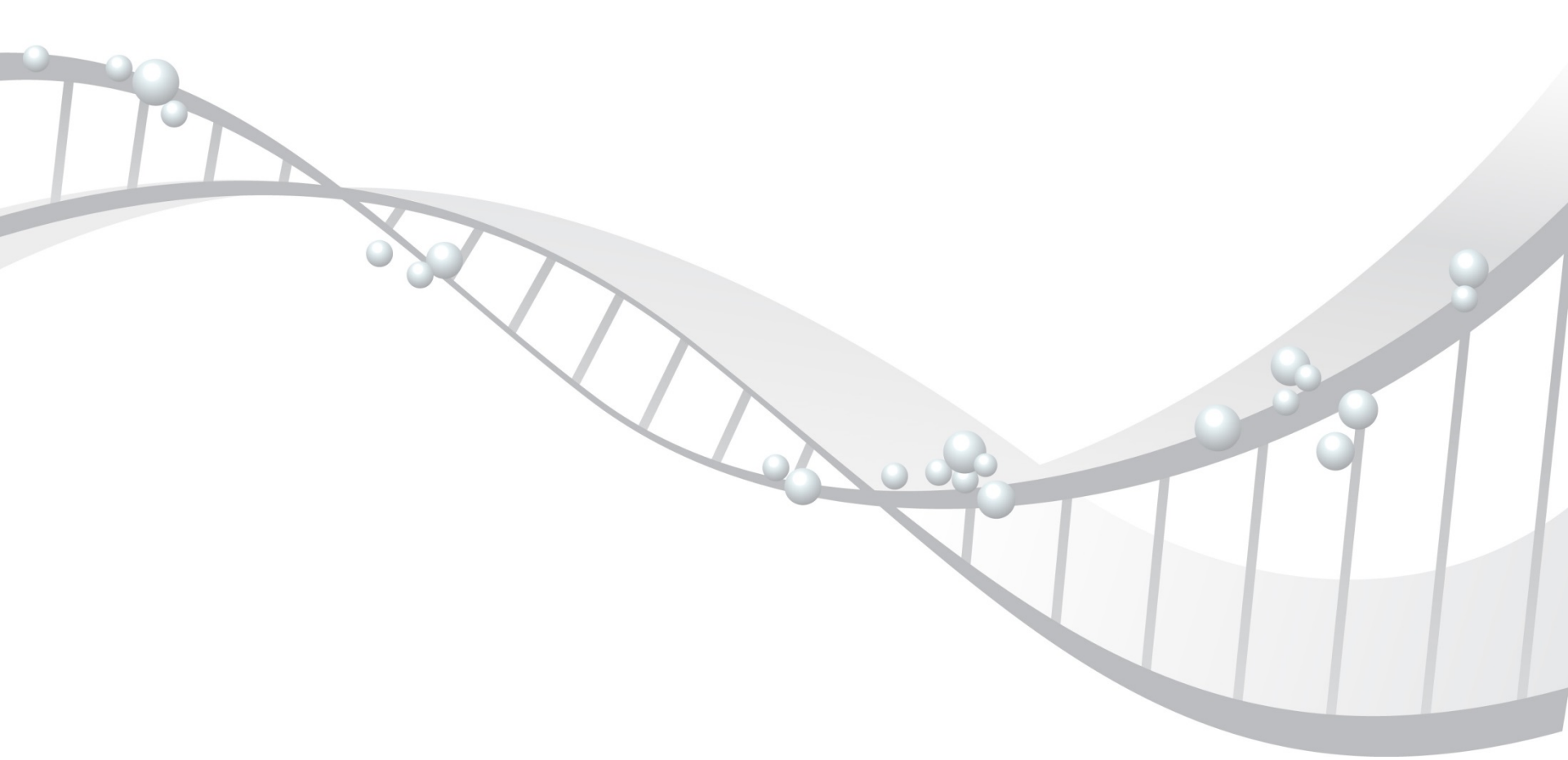
INTRODUCTION TO STATISTICAL MODELS AND MODEL SELECTION

Statistical Models

- Statistical models use data taken from a sample of individuals to estimate a dependent variable based on a selection of independent explanatory variables.
- Providing the sample is representative the model results can be used to make inferences about the population of interest.
- They underpin most efficacy and safety analysis in clinical trials.
- Type of statistical model selected is driven by the type of data collected an its expected distribution.

Model Selection

- A mechanism to select the 'best' (or 'most parsimonious') model from a set of candidate models.
- In this context 'best' is hard to define, however good model selection techniques seek a balance between a well-fitting model based on the available data, and simplicity.
- We want the model selected to identify trends in the data without being too complicated.
- Model selection techniques aim to systematically determine the set of explanatory variables that 'best' describe the response variable.



MOTIVATING EXAMPLE

Motivating Example

- We wish to create a model to estimate the weight of children aged between 11 and 16 using weight and height data for 19 subjects (10 males and 9 females) within this age range.
- Response variable:
  - Weight
- Potential explanatory variables:
  - Height
  - Gender
  - Age
- Exploratory plots showed that the clearest correlation present was between weight and height.

Motivating Example (2)

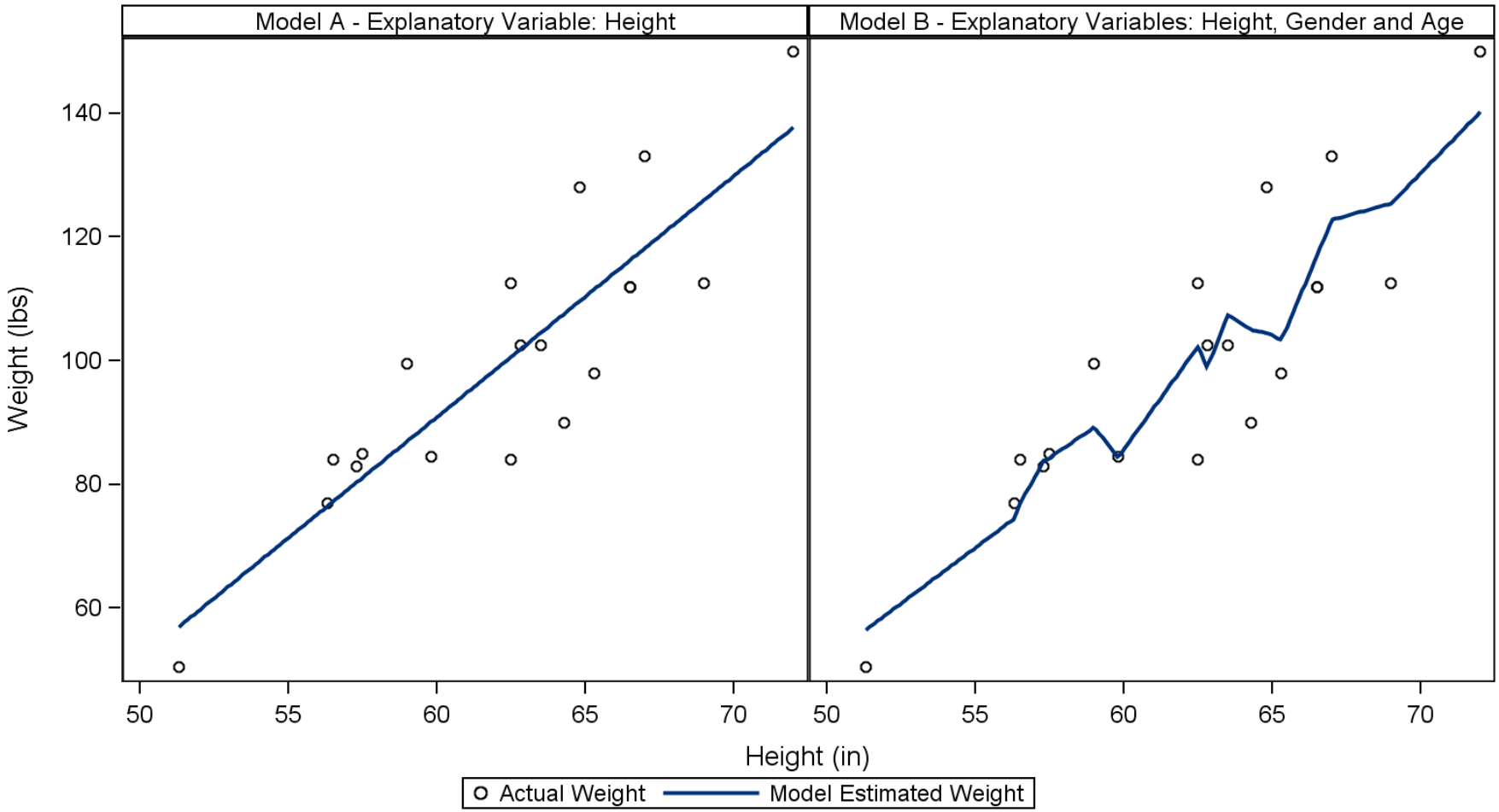
Based on the exploratory plots the following two linear regression models were fit to the data:

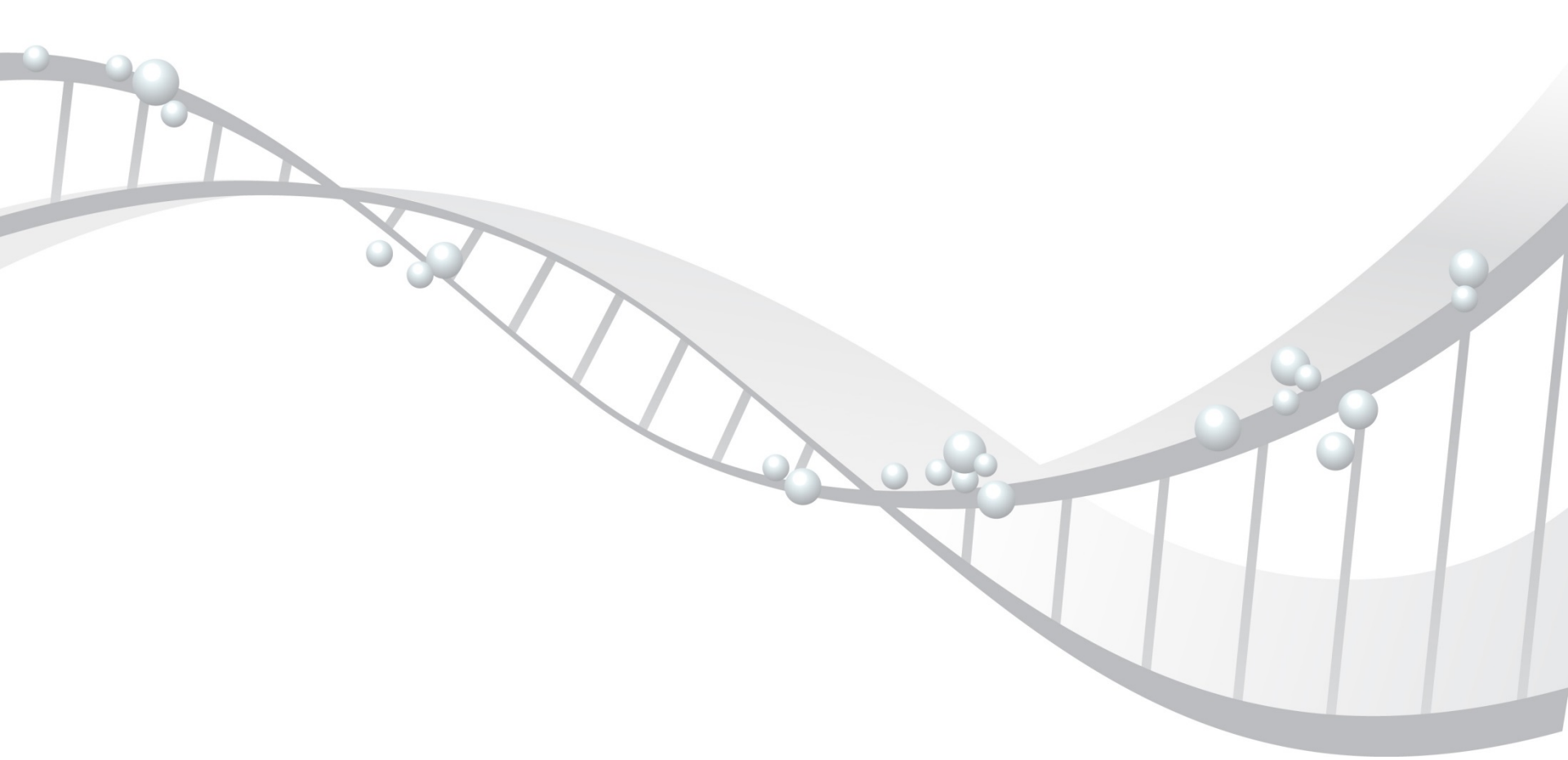
Model A: $y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i$

Model B: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$

where y_i is the response variable weight, x_{i1} and x_{i3} are the continuous covariates height and age respectively, x_{i2} is the factor gender, β_0 is the intercept, β_1, β_2 and β_3 are the parameter estimates associated with each explanatory variable and ε_i is the residual error.

Motivating Example (3)





MODEL SELECTION TECHNIQUES AVAILABLE IN SAS

Model Selection Techniques available by SAS Procedure

	PROC LOGISTIC	PROC REG	PROC PHREG	PROC GLMSELECT
Forward Selection	Y	Y	Y	Y
Backward Selection	Y	Y	Y	Y
Stepwise Selection	Y	Y	Y	Y
Branch-and-Bound Algorithm	Y		Y	
R^2 Selection (various)*		Y		
Mallows C_p Selection		Y		
Least Angle Regression				Y
LASSO				Y
Group LASSO				Y^
Elastic Net				Y^
*Including Maximum R^2 Improvement, Minimum R^2 Improvement, R^2 Selection and Adjusted R^2 Selection.				
^Technique only available in SAS v9.4.				

Brief Summary of these Procedures

- **PROC LOGISTIC:** Fits linear logistic regression models for discrete response data by the method of maximum likelihood. It can also perform conditional logistic regression.
- **PROC REG:** A general purpose procedure for fitting regression models.
- **PROC PHREG:** Performs regression analysis of survival (or 'time to event') data based on the Cox proportional hazards model.
- **PROC GLMSELECT:** Performs effect selection in the framework of general linear models. The procedure bridges the gap between PROC GLM, which doesn't include effect selection functionality, and PROC REG, which doesn't include a CLASS statement.

Stepwise Techniques

- Forward, Backward and Stepwise variable selection techniques are widely used fall under the umbrella of Stepwise techniques.
- They follow a relatively intuitive process where candidate models are explored by either adding or deleting an explanatory variable/interaction term from the model before comparing against the model from the previous stage.

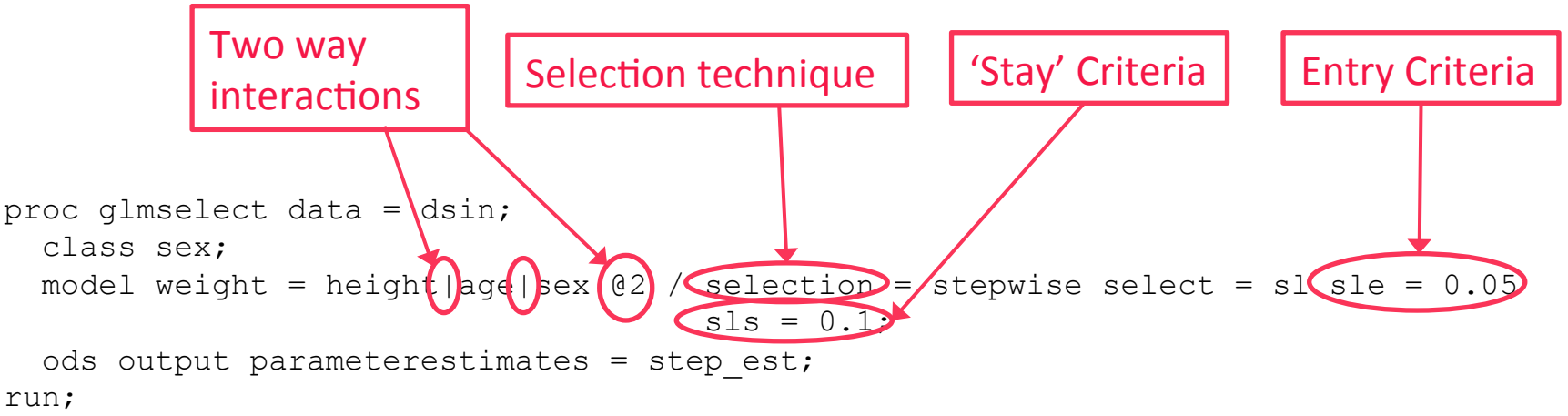
Forward Variable Selection

- Start with the null model and consider each explanatory variable individually for inclusion in the model.
- The variable which makes the greatest contribution to the model is added to the model, providing the associated p-value is less than the pre-specified value (often set at 0.05 or 0.1).
- This model is carried forward to the next step where the remaining variables are considered in turn for inclusion in the model.
- This iterative process is continued until no further explanatory variables meet the criteria for inclusion in the model.
- Once a variable is added to the model it remains in the model regardless of the p-value when further variables are added.

Backward variable selection follows a similar iterative process, but starts with the full model and variables are removed in turn.

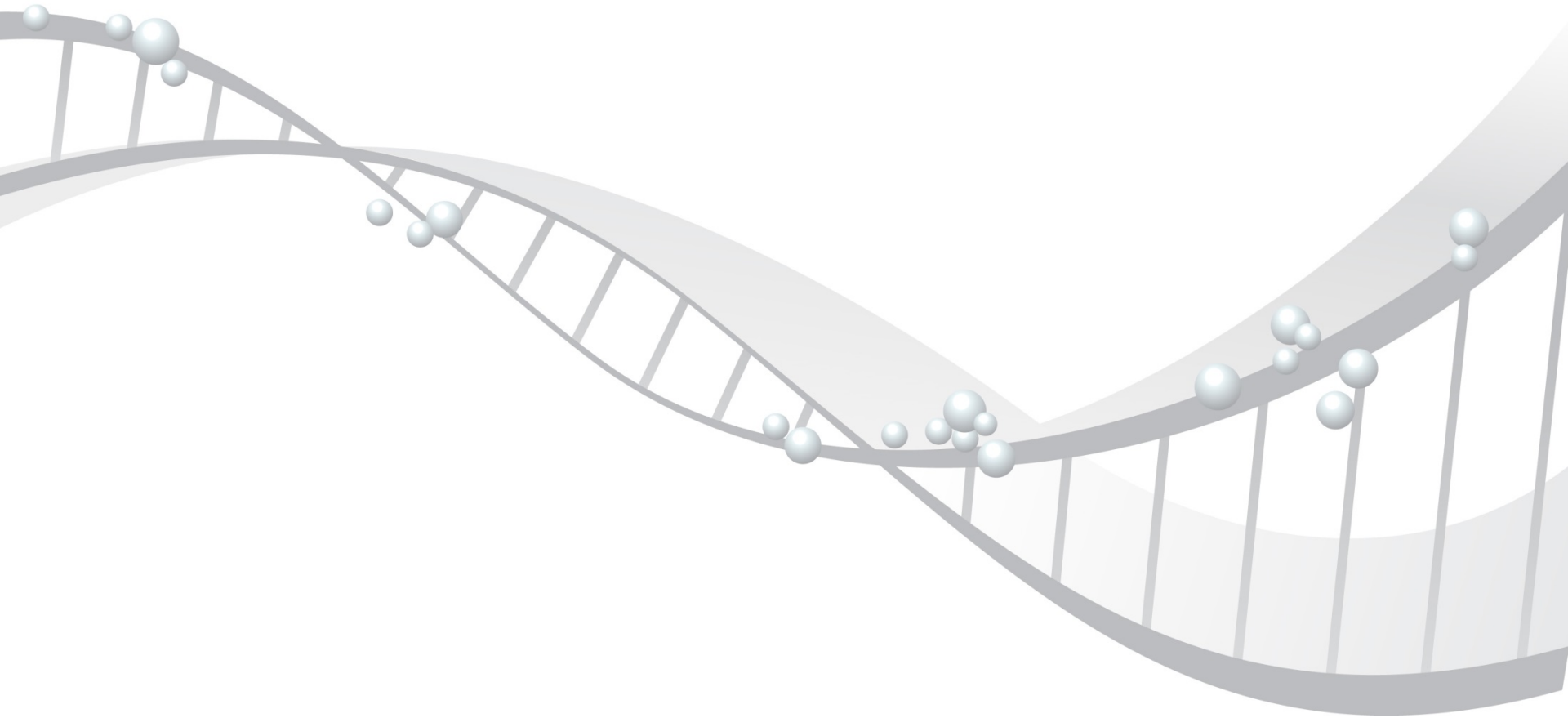
Stepwise Selection

- Similar to forward variable selection, but less restrictive in that once a variable is added to the model it does not necessarily stay there.
- A significance level for entry into, and to stay in the model must be specified for stepwise selection.



Limitations of Stepwise Techniques

- Stepwise technique require the application of a hypothesis test intended for one test to many tests. It can be shown that this assumption violation:
  - Biases p-values and standard errors towards zero;
  - Biases parameter estimates away from zero;
  - Potentially leads to the model selected being too complex.
- There is no guarantee that the final model selected truly is the best model. Applying forward and backward techniques to the same data will not necessarily yield the same final model.
- The nature of stepwise algorithms means that not all potential candidate models will be considered.



MODEL SELECTION IN CLINICAL TRIALS

Early Phase

- It is an industry requirement that factors/covariates to be included in primary analysis must be pre-specified in the protocol.
- Therefore data collected during early phase can be explored using model selection techniques to help inform these decisions.
- The nature of early phase studies means that sample sizes are likely to be small which may limit the conclusions and inferences you are able to take from the data.

Late Phase – Example Situations

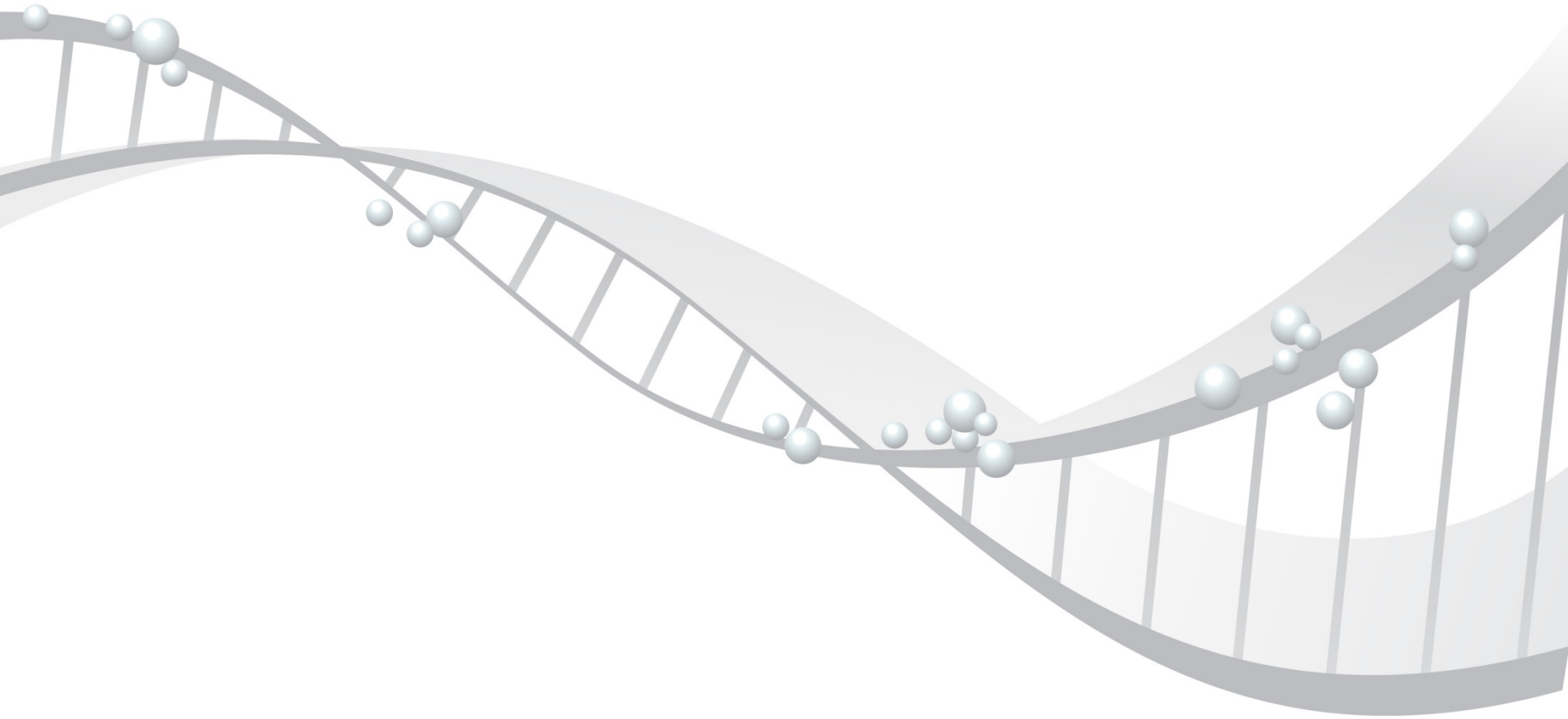
- Identifying prognostic factors/covariates which have an association with disease outcome.
- Exploring any additional/unexpected treatment effects present in the data.
- Seeking additional details about the efficacy profile of a treatment successfully taken to market.

Late Phase – Example Situations (2)

- Finding an appropriate set of factors/covariates to include in an imputation model (for example MCMC (Markov Chain Monte Carlo)) for an estimand that requires missing data points to be imputed.
- Performing exploratory analysis on data (could be pooled across multiple studies) with the view of confirming expected correlations and identifying any unexpected trends present in the data.

Late Phase – Caution!

- Is the data complete enough to draw meaningful conclusions? Particularly relevant when pooling studies.
- When performing exploratory analysis on a non-primary study endpoint there's no guarantee that known predictors will have been recorded.
- Statistical significance and clinical significance should never be confused. With larger sample sizes it is increasingly likely that a statistically significant difference will not be clinically significant.



CONCLUSION

Conclusion

- There are areas in clinical trial analysis where model selection is warranted.
- Developments in recent versions of SAS mean that numerous model selection techniques can be implemented efficiently to explore clinical data.

Conclusion – main considerations

- The following main considerations should be adhered to when employing model selection techniques:
  - The statistical assumptions of the underlying statistical model should be satisfied.
  - Data driven conclusions should always be considered in their clinical context and never trump expert knowledge in the therapeutic area.
  - Statistical significance and clinical significance should never be confused when drawing conclusions.

Questions?