Overview of methods for subgroup and biomarker identification from clinical data

Ilya Lipkovich, PhD
Eli Lilly and Company
OUTLINE

- Data mining/machine learning vs classical statistics
- Why data mining/machine learning for clinical data?
- Predictive versus prognosis effects
- Data-driven versus “guidance driven” subgroup analysis
- Taxonomy of biomarker identification methods
- Software for subgroup identification. What features to look at?
- Summary
DATA MINING/MACHINE LEARNING VS CLASSICAL STATISTICS

<table>
<thead>
<tr>
<th>Classical statistics</th>
<th>Data mining/Machine learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relatively small data sets collected from designed experiments or by sampling from well-defined populations</td>
<td>Large (sometimes dispersed and heterogeneous) data sets, often collected for purposes other than data mining</td>
</tr>
<tr>
<td>Assumes data generation mechanism: $y = f(x) + \varepsilon$, where $f(x)$ has a simple structure and the error term ($\varepsilon$) is modeled by parametric distributions</td>
<td>Aims at recovering unknown function $f(x)$ as a “black box” while the presence of the “error term” is often ignored</td>
</tr>
<tr>
<td>Objective is to estimate parameters for the entire population from available sample(s)</td>
<td>Objective is to obtain predictions for new (future) cases [supervised] or extract useful features that reveal unknown structure [unsupervised]. Analysis data often represent the entire population</td>
</tr>
<tr>
<td>Focus on hypothesis testing: a single test or a small number of pre-specified tests with clearly defined multiplicity control procedures</td>
<td>Hypothesis generation/knowledge discovery rather than formal hypothesis testing, less emphasis on statistical significance (often focusing on controlling the false discovery rate)</td>
</tr>
</tbody>
</table>
Patient diagnostics

- Example: tree-based decision rules that allowed clinicians of an emergency unit to make a quick assessment whether a patient with chest pain can be diagnosed with a myocardial infraction

Predictive models for patients’ future outcomes

- Prediction models for safety or efficacy outcomes, informed by assigned treatment, biomarkers available prior to treatment initiation, and evolving (early) patient outcomes

Modeling as part of treatment evaluation strategies

- Examples: modeling to account for selection bias due to post randomization/intercurrent events (e.g. modeling dropouts or implementing multiple imputation)
DMML FOR CLINICAL DATA: UNSUPERVISED LEARNING

- Clustering to identify patients with similar efficacy outcomes
  - Especially relevant for diseases where the patients’ well-being is described by a set of variables representing complementary and sometimes conflicting clinical criteria and scales (neuroscience)

- Identifying patients with distinct response profiles (or trajectories) over time
  - Response profiles may represent different types of patients (e.g., “early responders who later fail,” “relapsers,” “gradual responders,” “sustained responders,” etc.)

- Methods for association learning
  - Example: in pharmacovigilance to uncover drug-adverse event relationships and drug-drug interactions in spontaneous reporting systems and large healthcare databases

- Detecting outliers and unusual patterns
  - Often used in the context of fraudulent assessment of outcomes (e.g. see O’Kelly, 2004)
Clustering to identify patients with similar efficacy outcomes

- Especially relevant for diseases where the patients’ well-being is described by a set of variables representing complementary and sometimes conflicting clinical criteria and scales (neuroscience)

Identifying patients with distinct response profiles (or trajectories) over time

- Response profiles may represent different types of patients (e.g., “early responders who later fail,” “relapsers,” “gradual responders,” “sustained responders,” etc.)

Methods for association learning

- Example: in pharmacovigilance to uncover drug-adverse event relationships and drug-drug interactions in spontaneous reporting systems and large healthcare databases

Detecting outliers and unusual patterns

- Often used in the context of fraudulent assessment of outcomes (e.g. see O’Kelly, 2004)
DMML FOR CLINICAL DATA: SEMI-SUPERVISED LEARNING

- Subgroup identification
  - Identifying subgroups of patients with differential treatment effect from clinical trials data (e.g. from failed Phase 3 trials or from early phase trials with the idea of using for enrichment in subsequent stages of clinical programs)

- Estimating optimal individual treatment regimes
  - Construction of optimal dynamic treatment regimes (DTRs) utilizing information on patient’s characteristics and accumulated patient’s outcomes at each decision point

Unlike in supervises learning the outcome here is individual treatment difference/contrast that is not observable. Unless the same patient is taken all candidate treatments, only one potential treatment outcome is observed per patient.
The task of personalized medicine can be “translated” into statistical language as constructing predictive biomarker signature that would allow identifying patients with differential treatment response.

The schematic plots show four types of relationships between the outcome and a single biomarker:

- X is prognostic but not predictive
- X is prognostic and predictive
- X is predictive but not prognostic
- X is neither prognostic nor predictive
Subgroup analyses are often (rightfully) viewed as data dredging

Many authors came up with various “checklists” of principles for Subgroup Analyses

- NHS R&D HTA Programme (Brookes et al. 2001) provides a list of 25 recommendations
- Rothwell (2005) proposed a guideline with 21 rules
- Sun et al (2009) listed the existing 7 plus 4 additional criteria for assessing credibility of subgroup analysis

EMA Guideline on the Investigation of Subgroups in Confirmatory Clinical Trials (Draft, Jan 2014)

- Recognizes issues with current SA practices that “create disincentive to properly plan the investigation of subgroups”

The Guidelines encourage to “exercise caution” when conducting subgroup analyses, which is hard to operationalize...
“Guideline-driven” approach fails to encompass modern scientific approaches to statistical learning and the need for evidence-based personalized/stratified/precision medicine.

A different view: subgroup identification/analysis is framed as a special case of model selection.

This helps link subgroup identification efforts with the wealth of statistical methodology on model selection.

Pre-specified is the entire biomarker/subgroup selection strategy, not specific subgroup(s).
“Complexity control” to prevent data overfitting
- Tuning parameters controlling the search process need to be determined often in a data-driven fashion, e.g., via cross-validation
- E.g., penalized regression, a.k.a. shrinking, regularization

Evaluating the type I error rate for the entire subgroup search strategy
- E.g., using resampling under null

Obtaining “honest” estimates of treatment effect in subgroups (i.e., treatment effect expected in identified subgroups if applied to future studies)
- E.g., by using resampling methods or Bayesian model averaging/empirical Bayes
- Uncertainty associated with the entire strategy should be accounted for
GLOabal OUTCOME MODELING

- Modeling underlying outcome function \( f(x, t) = E(Y|X = x, T = t) \), where \( Y \) is an outcome, \( X \) is a collection of biomarkers and \( T=0,1 \) is a treatment indicator
  - computing individual treatment differences \( \hat{c}_i = \hat{f}(x_i, 1) - \hat{f}(x_i, 0), i = 1, ..., N \), that can be further modeled as an outcome
  - allows constructing a predictive score as a function of biomarkers, a biomarker signature: \( c(x) \)

- Some recent methods
  - Virtual Twins by Foster, Taylor and Ruberg (2011) [combining Random Forest for \( f(x, t) \) and CART for \( c(x) \)]
  - Penalized regression (FindIT) by Imai and Ratkovic (2013)
  - Bayesian hierarchical modeling (Jones et al, 2011 extending Dixon and Simon, 1991)
Directly modeling underlying treatment effect, \( c(x) \)

- Classification and regression tree methods can be adopted by incorporating treatment variable in the splitting criterion, resulting in piecewise constant fit for \( c(x) \)
- Parametric models were proposed that obviate the need for fitting in prognostic effects

Some recent methods

- Interaction trees (Su et al., 2005)
- Gi method (Loh et al., 2015) (implemented within GUIDE suite)
- Model-based recursive partitioning (Seibold et al., 2014).
- Modified covariate method by Tian et al. (2014)
Estimating optimal treatment regime $\text{sign}[c(x)]$

- Obviates the need to fit-in prognostic (main) effects, estimates optimal treatment regime by fitting a weighted classifier for treatment as a “response” with outcome-based weights $w(y, x) = \frac{y}{\text{Pr}(T=t|X=x)}$

- Weights incorporate the probabilities of treatment which are known in RCT and can be obtained by modeling propensity of treatment assignment in observational (non-randomized) studies

Some recent methods

- Outcome weighted learning (OWL) introduced by Zhao et al. (2012); ROWSi method (Xu et al. 2015)
Identifying subgroups $S$ with enhanced treatment effect $c(x) > \delta$ for $x \in S$

- Instead of estimating the response function $c(x)$ in the entire covariate space first and then carving out the interesting part where $c(x) > \delta$, these methods would directly search for such interesting regions.

Some recent methods:


- SIDES (by Lipkovich et al., 2011) and SIDEScreen (Lipkovich and Dmitrienko, 2014)
WHAT FEATURES OF A SA METHOD WE SHOULD LOOK FOR?

- What is the number of candidate predictors that can be processed in efficient manner (p=1, 20, 100, 1000)?
- What is the “model space” induced by the procedure and how model complexity is controlled to prevent overfitting?
- What outputs does the method produce?
  - Signatures of promising subgroups
  - Personalized treatment contrast
  - Optimal treatment assignment
  - Predictive biomarkers ordered by predictive strength.
- How the false discovery is controlled, if at all (type I error control, FDR)
- Does the method provide “honest” estimates (point estimates, SE, CI) of treatment effect in identified subgroups corrected for over-optimism?
  - E.g. using cross-validation, bootstrap, Bayesian model averaging
## SUMMARY OF SUBGROUP IDENTIFICATION METHODS

<table>
<thead>
<tr>
<th>Method</th>
<th>Modeling type (1)</th>
<th>Dimensionality (2)</th>
<th>Biomarker selection (3)</th>
<th>Control of false positive rate (4)</th>
<th>Complexity control (5)</th>
<th>Selection control (6)</th>
<th>Honest estimate of treatment effect (7)</th>
<th>Software implementation (8)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Virtual Twins (Foster et al. [42])</td>
<td>Freq/NP</td>
<td>High</td>
<td>P, S</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>B</td>
</tr>
<tr>
<td>Cai et al. [38]</td>
<td>Freq/NP</td>
<td>Low</td>
<td>P</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>C</td>
</tr>
<tr>
<td>Findl (Imai and Ratkovic [37])</td>
<td>Freq/P</td>
<td>High</td>
<td>S</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>C</td>
</tr>
<tr>
<td>STIMA (Dusseldorp, Concensano and Van Os, 2010)</td>
<td>Freq/NP</td>
<td>Medium</td>
<td>S</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>C</td>
</tr>
<tr>
<td>Bayesian approaches (Dixon and Simon [45]; Hodges et al. [46])</td>
<td>Bayes/P</td>
<td>Low</td>
<td>P</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>B</td>
</tr>
<tr>
<td>Interaction Trees (Su et al. [70]; Negassa et al. [68])</td>
<td>Freq/NP</td>
<td>High</td>
<td>S</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>B</td>
</tr>
<tr>
<td>Gi as part of GUIDE (Loh et al. [71])</td>
<td>Freq/NP</td>
<td>Medium</td>
<td>S</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>C</td>
</tr>
<tr>
<td>Modified covariate method (Tian et al. [75])</td>
<td>Freq/P</td>
<td>High</td>
<td>P</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>B</td>
</tr>
<tr>
<td>QUINT (Dusseldorp and Mechelen [74])</td>
<td>Freq/NP</td>
<td>Medium</td>
<td>S</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>C</td>
</tr>
<tr>
<td>Biomarker selector (Gunter et al. [21])</td>
<td>Freq/P</td>
<td>High</td>
<td>B</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>B</td>
</tr>
<tr>
<td>Qian and Murphy [78]</td>
<td>Freq/P</td>
<td>High</td>
<td>P,T</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>B</td>
</tr>
<tr>
<td>Zhao et al. [85], Xu et al. [88]</td>
<td>Freq/P</td>
<td>High</td>
<td>P,T</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>B</td>
</tr>
<tr>
<td>Zhang et al. [86]</td>
<td>Freq/SP</td>
<td>High</td>
<td>T</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>P</td>
</tr>
<tr>
<td>Adaptation of PRIM (Chen et al [93]; Kehl and Ulm [90])</td>
<td>Freq/NP</td>
<td>High</td>
<td>S</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>P</td>
</tr>
<tr>
<td>SIDES (Lipkovich et al. [91]) and SIDEScreen (Lipkovich et al. [94])</td>
<td>Freq/NP</td>
<td>Medium</td>
<td>B,S</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>B</td>
</tr>
<tr>
<td>Berger et al. [25], Sivaganesan et al. [95]</td>
<td>Bayes/NP</td>
<td>Medium</td>
<td>S</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>P</td>
</tr>
</tbody>
</table>

Lipkovich, Dmitrienko, D'Agostino. Tutorial in biostatistics... 2016
Software for subgroup identification

SIDES method

R package SIDES implementing the regular SIDES method (Subgroup Identification Based on Differential Effect Search) based on Lipkovich et al. (2011) [last update: October 04, 2016]. The package is maintained by Marie-Karelle Riviere (eldamj@gmail.com).

Download the SIDES package (an Excel add-in) which implements the regular SIDES and SIDEScreen methods [last update: March 25, 2016]. The package is maintained by Ilya Lipkovich (ilya.lipkovich@gmail.com).

Download the R functions, C++ functions (sides64.dll), and examples for the regular SIDES (Lipkovich et al. 2011), SIDEScreen (Lipkovich and Dmitrienko, 2014), and Stochastic SIDEScreen (Lipkovich et al. 2017) methods [last update: October 01, 2018]. The functions and examples are provided by Ilya Lipkovich (ilya.lipkovich@gmail.com), Alex Dmitrienko and Bohdana Ratitch.

Interaction Trees method

Download the R functions and examples for the Interaction Trees method [last update: Dec 30, 2014]. The functions and examples are provided by Xiaoqiang Su (Xiaoqiang Su’s site). Download the R code for the Interaction Trees method [last update: Dec 30, 2014].

Virtual Twins method

Download the R code for the Virtual Twins method [last update: Dec 30, 2014]. The code is provided by Jared Foster (jaredcf@umich.edu).

R package aVirtualTwins that implements an adaptation of the Virtual Twins method by Foster et al. (2011).

GUIDE package

GUIDE package for classification and regression trees now includes methods for subgroup identification. The GUIDE package is maintained by Wei-Yin Loh (Wei-Yin Loh’s site). For more information on the subgroup identification features, see Section 5.10 of the GUIDE User Manual [last update: September 25, 2018] and paper by Wei-Yin Loh, Xu He and Michael Man.

QUINT method

Quint package for Qualitative Interaction Trees. The package is maintained by Elise Dusseldorp (Elise Dusseldorp’s site) and colleagues. Reference: Dusseldorp and Mechelen (2014).

Findit method


Blasso method

Download the R code for the Bayesian two-stage Lasso strategy for biomarker selection for time-to-event endpoints [last update: December 16, 2014]. The code is provided by Xuemin Gu (xuemin.gu@bms.com). Reference: Gu, Yin and Lee (2013).

ROWSi method

Download the R code for the ROWSi method (Regularized Outcome Weighted Subgroup Identification). Reference: Yu et al. (2015).

Model-based Recursive Partitioning

R partykit package: A Toolkit for Recursive Partitioning, which can perform subgroup analyses using the functions intree(), glistree() (or more generally, melt()) and extract().

Recently a new package modeltry has been created that specializes on stratified and personalized treatment effect estimation. The package is maintained by Heidi Seibold (heidis@seibold.cc).

See examples of subgroup analysis in Seibold et al. (2015) and Seibold et al. (2016).

Other sources:

R package personalized (maintained by Jared Huling) for subgroup identification and estimation of heterogeneous treatment effects. It is a general framework that encompasses a wide range of methods including ROWSi, outcome weighted learning, and many others. See documentation and article explaining the underlying methodology.
Data mining/machine learning methods are becoming an integral part of data analysis at all stages of clinical drug development, which can be contrasted with its primary use in pre-clinical stage of “drug discovery” in the past.

We emphasize principled or disciplined use of subgroup identification (and data mining) as opposed to haphazard data-dredging and treat subgroup identification as a special case of model selection and contrast data-driven with guideline-driven approach.

Unlike standard predictive modeling methods that aim at identifying subgroups with heterogeneous outcome, using methods for tailoring/personalized medicine requires modeling individual treatment differences targeting subgroups with heterogeneous treatment effect.

Methods for subgroup identification and analysis borrow from diverse literature in machine learning, multiple testing and causal inference.

A feature of subgroup identification (and data mining in general) in drug development is the need to control the Type I error (or false discovery) rates which is a new trend in the area of machine learning.

Once subgroups have been identified, analyst is facing the challenge of obtaining “honest” estimates for associated effects that should be expected in the future data.
REFERENCES

- Sun X, Briel M, Walter SD, Guyatt GH. Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses, BMJ 2010; 340:c117doi: 10.1136/bmj.c117
THANK YOU!

Lipkovich_Ilya_A@Lilly.com