RegressionExplorer

Citation for published version (APA):

Dingen, D., van 't Veer, M., Houthuizen, P., Mestrom, E. H. J., Korsten, E. H. H. M., Bouwman, R. A., & van Wijk, J. (2019). RegressionExplorer: Interactive exploration of logistic regression models with subgroup analysis. *IEEE Transactions on Visualization and Computer Graphics*, 25(1), 246-255. [8464305].
<https://doi.org/10.1109/TVCG.2018.2865043>

Document license:

TAVERNE

DOI:

[10.1109/TVCG.2018.2865043](https://doi.org/10.1109/TVCG.2018.2865043)

Document status and date:

Published: 01/01/2019

Document Version:

Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.tue.nl/taverne

Take down policy

If you believe that this document breaches copyright please contact us at:

openaccess@tue.nl

providing details and we will investigate your claim.

RegressionExplorer: Interactive Exploration of Logistic Regression Models with Subgroup Analysis

Dennis Dingen, Marcel van 't Veer, Patrick Houthuizen, Eveline H. J. Mestrom,
Erik H.H.M. Korsten, Arthur R.A. Bouwman, and Jarke van Wijk

Abstract—We present RegressionExplorer, a Visual Analytics tool for the interactive exploration of logistic regression models. Our application domain is Clinical Biostatistics, where models are derived from patient data with the aim to obtain clinically meaningful insights and consequences. Development and interpretation of a proper model requires domain expertise and insight into model characteristics. Because of time constraints, often a limited number of candidate models is evaluated. RegressionExplorer enables experts to quickly generate, evaluate, and compare many different models, taking the workflow for model development as starting point. Global patterns in parameter values of candidate models can be explored effectively. In addition, experts are enabled to compare candidate models across multiple subpopulations. The insights obtained can be used to formulate new hypotheses or to steer model development. The effectiveness of the tool is demonstrated for two use cases: prediction of a cardiac conduction disorder in patients after receiving a heart valve implant and prediction of hyponatremia in critically ill patients.

Index Terms—Visual analytics, Predictive visual analytics, Exploratory data analysis, Multivariate statistics, Regression analysis, Variable selection, Subgroup analysis

1 INTRODUCTION

The logistic regression model is the most frequently used model to describe the relationship between a discrete dependent variable Y and a set of independent predictor variables X_1 to X_n [11]. We refer to the independent predictor variables as *covariates* and to the dependent response variable as the *responder*. For a logistic regression model this responder is *dichotomous*, which means that the number of different values it can take is restricted to two. Generally, these two values are represented by a '0' and '1', where 1 indicates a condition of interest. The logistic regression model then estimates the probability of $Y = 1$ for given covariate values $x_1, x_2, \dots, x_n$. Also the effect of each individual covariate on the responder is estimated.

In Clinical Biostatistics, our application domain, dichotomous responders are commonly used to estimate the probability of a patient's condition (e.g., the patient being ill or healthy). After selecting such a responder of interest, the clinical researcher is faced with the challenge of selecting appropriate covariates. To that end, the researcher has to respect two immediate requirements [12].

The first requirement is that the model has to be clinically interpretable. In other words, the clinical researcher has to justify the inclusion of each covariate rationally and in physiological terms.

The second requirement, or common guiding principle, is that the researcher should seek for the most parsimonious model that still accurately reflects the true response observed in the dataset itself. This

means that simple models are, in principle, preferred over more complex ones. The rationale behind this principle is that the resulting model is assumed to be more numerically stable, more easily adoptable in practice and less prone to overfitting.

Yet, if there are underlying associations present between the covariates, and some of these covariates are omitted from the model, estimates may be confounded. Although putting all clinically and intuitively relevant covariates in the model provides for more control of confounding, the trade off between parsimony and confounded estimates remains. Furthermore, individual covariates may not exhibit strong confounding, but that can change considerably if these covariates are analyzed in concert.

Given the interplay between the requirements, the use of fully automated methods for the selection of covariates is limited because the researcher has to generate posteriori arguments for the generated model(s). As a result, *purposeful selection* currently dominates in practice. Following this, the researcher starts by analyzing all covariates individually and includes all clinically and / or statistically relevant ones in the model. Next, covariates are iteratively added or removed — the *stepwise method*. In this way, the researcher considers successive candidate models along a path through the model space. Such a path is defined by the initial selection of covariates and the order of addition or removal. This approach, however, does not necessarily result in the best or most balanced model and is time-consuming or practically intractable when relying on software system like *SPSS* and *R* because of the demanding user interaction required for generating and analyzing the candidate models in a large model space. Kehler et al. [13] recognized already some years ago that in such systems visualization is mainly used for presentation purposes and that a coupling with interactive visualization should be beneficial. In our opinion this is still the case. Consequently, the number of models that can be considered effectively is still limited, which yields an increased probability of just finding a local optimum.

Although hybrid strategies, which combine automated methods with iterative methods, have been proposed earlier [24], these have not been captured yet within an effective visual analytics approach. Moreover, Zhang et al. [31] observed that no fully integrated visual analytics approach exists for dealing with just the traditional stepwise method alone. In response, they developed a system that allows the researcher to trace through the model space iteratively. With our novel method *RegressionExplorer*, we aimed to improve the effectiveness of model space exploration by directly visualizing larger parts of the model space containing alternative models. On top of that, we have added

- Dennis Dingen is with Eindhoven University of Technology. E-mail: d.j.w.g.m.dingen@tue.nl.
- Marcel van 't Veer is with Catharina Hospital Eindhoven. E-mail: marcel.vh.veer@catharinaziekenhuis.nl.
- Patrick Houthuizen is with Catharina Hospital Eindhoven. E-mail: patrick.houthuizen@catharinaziekenhuis.nl.
- Eveline H. J. Mestrom is with Catharina Hospital Eindhoven. E-mail: eveline.mestrom@catharinaziekenhuis.nl.
- Erik H.H.M. Korsten is with Catharina Hospital Eindhoven. E-mail: erik.korsten@catharinaziekenhuis.nl.
- Arthur R.A. Bouwman is with Catharina Hospital Eindhoven. E-mail: arthur.bouwman@catharinaziekenhuis.nl.
- Jarke van Wijk is with Eindhoven University of Technology. E-mail: j.j.v.wijk@tue.nl.

Manuscript received 31 Mar. 2018; accepted 1 Aug. 2018.
Date of publication 16 Aug. 2018; date of current version 21 Oct. 2018.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TVCG.2018.2865043

a mechanism for the dynamic selection of subgroups (representing subsets of records in the dataset) that are directly distinguishable within the visualization of the model space. As a result, the researcher is enabled to spot patterns in model parameters that vary simultaneously over candidate models and subgroups (or *subpopulations* in the clinical context). Because existing systems offer no or very limited support with regard to the combination of dynamic subgroup selection with model space exploration, this constitutes the novelty of our system and saves the researcher much time in practice. All gathered insight can be used to formulate new hypotheses about the data, guide the development of models using classical methods for model building, form a basis for practical clinical decision making, or setup a complete clinical trial to verify an effect. Our method is applied for two clinical use cases: predicting cardiac conduction disorder in patients after receiving a heart valve implant, and hypernatremia in critically ill patients. Both use cases are supported with real-world data and analyzed in cooperation with domain experts.

This paper continues with an overview of related work in Section 2. Section 3 covers some background on logistic regression models in a clinical setting. Next, the currently prevailing model-building workflow followed by the typical clinical researcher is analyzed. Subsequently, a list of design goals and tasks for our method is formulated in Section 4 and our method is presented in Section 5. In Section 6 the use cases are discussed. A discussion is provided in Section 7. Finally, in Section 8 conclusions are drawn and future work is suggested.

2 STATE OF THE ART

With our focus on the clinical setting we observed some trends in literature. Early work on interactive visualization of patient attributes stored in Electronic Health Records (EHRs) dates back for over 20 years. As time went on patient groups (cohorts) have been visualized, with some basic descriptive statistics. Thereafter, the need for integrating more advanced statistics into these types of systems has been recognized and in some cases, was acted upon. Another trend was the emergence of *predictive visual analytics*, a blend between Visual Analytics methods and predictive modeling. Our work can be placed under that heading as well, but since we are dealing with subgroup analysis our work is also, albeit indirectly, related to cohort visualization. In the remainder of this section our observations are detailed.

In 2013 Rind et al. [22] published an extensive survey on systems for the interactive visualization of EHRs. Two main types were identified. The first type is designed for exploring EHRs of individual patients and typically shows the subjects' key characteristics and multiple associated event sequences over time. The second type is designed analogously for comparing multiple patients simultaneously and involves more descriptive statistics or clustering techniques. Rind et al. also identified interactive visual querying of health records as future work and called to integrate more statistics into the systems. Afterwards, systems designed for exploring the (temporal) attributes of cohorts in combination with basic descriptive statistics and interactive temporal (event) querying have emerged [1, 16, 19]. Since then the interest in integrating data mining techniques, machine learning and more complex statistics increased [5, 10, 19].

Meanwhile, the relatively new area of predictive visual analytics has been receiving more attention over the past years. It brings visual analytics methods into the domain of *predictive analytics*, which is concerned with predicting future outcomes and trends using techniques from data mining, machine learning and statistics. In addition to the identification of characteristic systems, efforts have been put into proposals involving transformations of the current visual analytics pipeline, such that it encompasses validation, building and exploration of models [8, 17, 18, 30].

It appears that the systems designed for the exploration of EHRs could be augmented with more sophisticated statistical methods by adopting techniques developed in predictive (visual) analytics. Considering (logistic) regression models, there are various aspects one can focus on, like the relationships between the responder and covariates, the relationships between covariates and the type thereof, assumptions and diagnostics of models, and subgroup analysis. Over the years some

of these aspects have been captured by visual analytics methods.

First of all, standard correlation matrices have been augmented by embedding glyphs in the cells [9]. The glyphs highlight the relationships between variables under consideration in more detail. Scatter plots have also been extended with animation and lines that trace out the movement of points in order to visualize variables for multiple cohorts simultaneously [23]. Next, interactive 3D scatter plots augmented with sensitivity (stream) lines have been introduced by Chan et al. [6]. The sensitivity of a relationship between variables is computed using regression and the system offers visual support for selection, brushing and clustering.

Another method for the exploration of correlations between variables derived from regression models consists of an annotated network visualization linked to a parallel coordinates plot, published by Zhang et al. [32]. More recently, Wang et al. [28] proposed network visualization in conjunction with tables showing (descriptive) statistics as a framework to uncover causal relations among variables in a multivariate dataset. The relationship between two variables and model properties has also been visualized. Mühlbacher [20] proposed visual domain space partitioning of all relevant variable pairs (shown with small multiples) in which a target quantity (model property) is color coded. Related is the *HyperMoVal* approach that renders linked 2D and 3D sub-projections of an n -dimensional function space around a user-specified focal point [21]. Zhang et al. [31] designed a system for building and evaluating logistic regression models that covered most of the steps depicted in Figure 1. It automatically shows the results of univariate analysis in a table for the main population and a fixed number of subpopulations (strata). A second panel shows the distributions of all the variables in small multiples, which are grouped based upon the results of factor analysis. On the bottom resides a third table containing the model properties of a trace through the model space that the user controls by iteratively adding variables to the model.

Since logistic regression models can be used as a basis for classification systems, we like to mention two related works: a system designed for the interactive construction of decision (classification) trees by Van den Elzen et al. [26], and, a generic workflow for diagnosing binary classifiers is proposed by Krause et al. [14]. In the latter, a model is treated as a black box for which the overall accuracy is shown by visualizing the true / false positive and true / false negatives using histograms, tables and ROC curves. Alternative sets of covariates are listed and coupled to a table showing the corresponding distributions and odds ratios. Additionally, a matrix shows the color coded relative importance of covariates per case in the dataset.

To conclude, systems developed for cohorts (or subgroup analysis) typically visualize cohort attributes directly, while integrating only basic statistics. On the other hand, systems developed for model building or exploration typically focus less on interactive dynamic subgroup selection and visualization. Additionally, either a single model was explored at a time or the model space traversed was limited. These observations led us to work on RegressionExplorer with the main goal of more thorough model space exploration with integrated subpopulation analysis.

3 BUILDING LOGISTIC REGRESSION MODELS IN A CLINICAL SETTING

In this section the logistic regression model is defined, its relevant properties are discussed and the typical workflow for developing it is described.

3.1 The logistic regression equation

In order to explain the logistic regression model equation we first consider the equation for standard (multiple) linear regression:

$$\hat{Y} = \beta_0 + \sum_{i=1}^n \beta_i X_i \quad (1)$$

in which $\hat{Y}$ is the predicted value for responder Y based upon the values of covariates X_1 to X_n , β_1 to β_n are the β -coefficients or weights given to individual covariates and β_0 is the y-intercept. Note that in this

Fig. 1: Workflow followed by a clinician performing clinical research. Orange indicates steps where direct visual analytics support is provided by RegressionExplorer; blue indicates where user interaction is required.

case Y is real-valued instead of dichotomous. In order to deal with dichotomous variables, one predicts the *probability* of Y occurring, instead of its real value, using

$$P(\hat{Y}) = 1 / (1 + e^{-(\beta_0 + \sum_{i=1}^n \beta_i X_i)}). \quad (2)$$

The determination of the β -coefficients involves a maximum-likelihood estimation that can be performed by many modern software packages. A logistic regression model is said to be highly interpretable. The reason is that the real-valued β -coefficients can be interpreted as odds ratios (after taking the logarithm), with their sign indicating either a positive or negative effect on the responder. In clinical terms, positive and negative effects are called respectively *risk factors* and *protective factors*.

In addition to the coefficients, the significance of the association between a covariate and the responder can be expressed using a p -value (or confidence interval). In general, a p -value can be influenced by random errors, selection or information bias, confounding, sample size and magnitude of effects [25]. Note, however, that the magnitude of the β -coefficient does not necessarily influence the p -value (considerably).

3.2 Workflow for regression model development

Figure 1 shows an overview of the workflow followed by a clinical researcher when developing a regression model [2]. We proceed with discussing each step separately.

Selecting responder and population With a hypothesis in mind, the clinical researcher selects an appropriate responder that captures the condition of interest. A large enough data sample of the representative population needs to be collected that includes values for the responder and possible related covariates. The researcher should check the frequency of the condition captured by the responder. In general it is harder to make statistical sound claims for a more rare condition.

Univariate analysis For a basic understanding the researcher starts with analyzing the one-to-one relationships between individual covariates and the responder. Covariates that relate similarly to the responder may be correlated or interrelated. Clinically important covariates are identified (purposeful selection) and the remaining ones are sorted on significance levels as a priority list for multivariate analysis later on (excluding covariates with $p > 0.2$).

Covariate relationships In a logistic regression model the covariates are *assumed* to be independent. If this is actually true in reality, the estimated parameters capture the unbiased contribution of individual covariates. In practice, however, covariates are often not truly independent and (inter)relationships exists between the covariates and / or responder.

There are three basic types of relationships between covariates and the responder (see Fig. 2), which are mainly distinguished by the causal direction assumed or proven for an observed association. A *confounder* influences the responder and another associated covariate that is itself

Fig. 2: Types of relationships between covariates and responder. A confounder is associated with both the covariate and responder, but does not lay on a causal pathway in between covariate and responder. The same holds for a mediator, except that it actually does lay on the causal path way. A moderator acts as a function between the covariate and responder.

also associated with the responder. If a covariate influences the responder and is in turn influenced by another associated covariate, it is called a *mediator*. A *moderator* acts as a function between a covariate and responder, invoking a synergistic or inverting effect. Please note that the diagram in Fig. 2 is a simplification, because in reality, a covariate can have multiple (types of) relationships simultaneously with different covariates and / or the responder.

The fact that an interrelationship is present can be detected using *tolerance* and *Variance Inflation Factor* (VIF), which is a measure of *multicollinearity*. Yet, the type of interrelationship remains unclear and depends on the model in question. For these reasons and in order to reduce the design complexity of our system we have not adopted those methods.

Multivariate analysis and diagnostics In multivariate analysis the relationship between the responder and a set of covariates is analyzed. Therefore, the clinical researcher needs to select a set of appropriate covariates X_1 to X_n from the set of all available covariates for inclusion in the model. For each variable the shape of its distribution needs to be checked for irregularities, both visually and numerically (e.g., skewness and kurtosis), which may suggest to exclude the variable from candidate models. Next, the clinical researcher follows a trace through the space of all possible models.

An initial model includes all clinically relevant covariates along with either one or all significant covariates identified during univariate analysis. The model is then iteratively updated by adding / removing plausible covariates in order of priority (stepwise method). A candidate model is often compared with alternative models using the *Akaike Information Criterion* (AIC) or a similar measure. Note that the AIC can only be used for direct comparison of models computed using the same population.

If the number of available covariates is small (say, below 15), it is feasible to compute all possible models in the model space. Comparing them is nevertheless an increasingly challenging task as the number of models becomes bigger. Furthermore, automated model selection methods, which use a statistical measure as the sole ranking criteria, often yield a condensed list of models that are hard to rationalize from a clinical perspective. Although a direct interpretation of β -coefficients as odds ratios is useful for the rationalization of a model, it remains necessary to rationalize why the set of covariates itself was chosen in the first place from a clinical perspective. The stepwise method therefore serves as a feedback mechanism for the clinical researcher while testing out hypotheses. During every iteration, interrelationships between covariates are considered and assumed. On this basis, the researcher may decide to include or exclude a particular covariate as well.

Finally, model diagnostics are used to ensure the statistical soundness of a model. The first requisite is that the number of patients in the study population (cases) is sufficient for that particular model. The minimum

number of patients required is at least 15 times the total degrees of freedom (DoFs) exhibited by the set of covariates [3]. A continuous variable exhibits 1 DoF and a categorical variable exhibits $k - 1$ DoFs, where k is the number of categories or *levels* of the categorical variable. Second, residuals and other fit measures can be considered. Third, covariates represented by higher order terms in equations (1) and (2) might result in a better fit. Fourth, a moderation effect between two variables can be captured in a regression equation by adding a term that is formed by the product of these variables (an *interaction term*) and a weight.

Subgroup analysis A clinician may investigate the behavior of a model on a specific subpopulation and compare it with the behavior of that model on a different subpopulation. Under some circumstances it is more appropriate to develop multiple models targeting specific subpopulations instead of developing a single model that targets a population in which these subpopulations overlap. Take for instance a model that predicts a patients' condition based upon the number of pregnancies. Since males are never pregnant, the development of single model for targeting patients regardless of gender may result in a skewed model. This is a moderation effect on pregnancy through gender. Developing two models that are tailored to males and females eliminates this effect. Therefore, spotting changes in model behavior over different subpopulations is useful. Subgroup analysis can also be used to artificially neutralize any moderation (interaction) effects between the split variable and the covariates by design.

In practice, subgroups based upon binary splits seem common and the number of recursive splits is very limited (say, below 3). Aside from the inherent statistical problems, like sample size and loss of power, there is the prominent problem of subgroup balancing on given variables. Standard approaches have yet to be established according to Wang et al. [29].

4 DESIGN GOALS AND USER REQUIREMENTS

Based on the model-building workflow in Section 3 we have formulated general design goals and user requirements for our system. The design goals are:

- *DG1: Maintain familiarity with the current workflow followed by the clinical researcher building a logistic regression model;*
- *DG2: Common operations and computations are done automatically to aid the clinical researcher;*
- *DG3: An overview of the model space is presented such that the clinical researcher can explore it quickly, while the system follows the guiding principles and monitors underlying assumptions automatically; and*
- *DG4: The clinical researcher should be enabled to steer the system using domain knowledge.*

For the user requirements we have used a direct mapping from the model-building workflow as show in Table 1.

5 REGRESSIONEXPLORER

In this section we give an overview of the RegressionExplorer system (see Fig. 3). The individual components map almost directly to the workflow depicted in Fig. 1 (*DG1*) and are discussed accordingly. For this overview we consider predicting failure for a math test as an example. The test is taken by secondary school students in Portugal and the dataset is publicly available [7]. The dataset is used as it is published, but with the nominal variables recoded as natural numbers. We have selected this case because it is easier to understand than medical cases and also illustrates that our approach is not restricted to the medical domain.

Workflow step	User Requirements
Pick responder	(UR1) Selection mechanism (UR2) Show distribution
Select population	Assumed
Univariate analysis	(UR3) Show prioritized covariates (UR4) Show distributions and DoF
Relationships covariates	(UR5) Suggest relationships (Fig. 2) across groups
Multivariate analysis	(UR6) Model comparison mechanism (UR7) Ability to tag interesting models and covariates
Model diagnostics	(UR8) Check model costs in DoF
Pick (consider) final model	(UR9) Ability to contextualize favorite model
Subgroup analysis	(UR10) Interactive selection mechanism (UR11) Sample size checking (UR12) Comparison of distributions across subgroups

Table 1: Mapping of selected workflow steps to user requirements.

5.1 Design

We came to our solution via an iterative prototyping process, closely cooperating with three clinical researchers / epidemiologists. We spend much effort in understanding the workflow, and had many brainstorm sessions about how it could be supported. We started with a parallel coordinate plot to show the variations in coefficients across models. However, encoding the significance and magnitude together formed a challenge. Additionally, we needed to use color to distinguish between a limited selection of subgroups, the model structure in relation to the associations was lost, the absence of a variable was hard to encode and correlation could only be compared for neighboring axes. Combining or replacing the parallel coordinate plot with a relationship graph resulted in similar problems.

Initially subgroups were defined and selected using a standard tree widget, which became harder to read and manage as the number of subpopulations increased. The icicle plot was the solution and partly justified by the limited number of splits and populations considered in practice. Reconsideration of a matrix approach with the icicle plot already in place yielded our final solution.

Finally, it remained a challenge throughout the design process to balance the integration of more advanced statistics with the design of a stable and uniform display. For instance, the choice to rely primarily on parameter significance allows for more uniformity, but is in some cases less effective. Because the tool is intended for exploratory purposes we have accepted this drawback.

Univariate analysis Panel (A), shown on the left in Fig. 3, provides the list of available variables, which is populated upon loading the data. In accordance with the workflow in Fig. 1, the first step is to select a responder by dragging a variable from (A) to (B), the responder drop panel in which its distribution is shown (*UR1*, *UR2*).

Based on the loaded dataset (population) the system then performs univariate analysis of which the results are shown as colored rectangles after the variable names (and DoFs) (*UR4*). Red encodes a positive effect (odds ratio > 1) on the responder with significant statistical certainty: a risk factor. Conversely, blue encodes a negative effect (odds ratio < 1): a protective factor. The level of significance of an effect is determined by binning its p -value using cut-off values and

Fig. 3: Overview of the RegressionExplorer system: (A) Univariate analysis panel; (B) Responder drop panel; (C) Fixed covariates drop panel; (D) Optional covariates drop panel; (E) Matrix containing all generated multivariate models; (F) Population control panel; (G) Legend.

subsequently used to saturate the color in the rectangle, as shown in the legend (G) in Fig. 3. The default cut-off values are: $p \leq 0.00004$, $p \leq 0.0004$, $p \leq 0.004$, $p \leq 0.2$, and $p > 0.2$. A lower p -value results in a higher level of significance and a more saturated color.

Variables are automatically sorted on significance level (UR3); also, each item of the list can be expanded to show individual levels of the variable (root items show the most significant effect) and hovered to show its distribution in a histogram (UR4). This supports step two of the workflow (DG3). In the figure we can readily observe that the first period grade *G1* is the most significant protective factor of math test failure. Because this is considered a continuous variable, it means that an increase in grade increases the effect. For covariate *failures*, time spent going out (*goout*) and *age* we observe the opposite effect. The willingness to take higher education captured by the binary covariate *higher* is encoded relative to its first level (0=yes, 1=no), meaning that no willingness increases the risk of failure. An increased level in both the education of the mother (*Medu*) and the father (*Fedu*) are protective factors with significant probability, and so on.

Multivariate analysis For the third step of the workflow, the step-wise method for multivariate modeling can be followed by dragging variables from (A) to (C), which holds the covariates included in the model (DG4, UR10). A covariate is removed from (C) by clicking on [X] in the title bar. After each change, the system reestimates the model, which is shown in a single row of the multivariate model matrix (E). The model number is shown in the left header of the matrix in front of the row.

The first column shows the AIC as a quality measure for the model (UR6) and the second column shows whether the required DoF for the model are compensated for with large enough population sizes (15 times the required DoF at minimum) (DG2, UR8). These two columns are marked “pinned” by default, which locks them in place as an overlay while scrolling the matrix horizontally. The remaining columns show the significance levels for the included covariates using the same convention as for (A), but using possibly different cut-off values. By default the cut-off values are $p \leq 0.00001$, $p \leq 0.0001$, $p \leq 0.001$, $p \leq 0.05$ and $p > 0.5$. Furthermore, categorical variables lead to

multiple columns, as each category needs to be encoded as a separate variable in the model (dummy encoding), where the first variable serves as the reference for the others. Hence, a categorical variable with k levels requires $k - 1$ DoFs and a continuous variable 1 DoF.

The researcher can generate many models by adding and removing covariates to and from (C), but this can be a lengthy and tedious process. Clinical researchers preferably compare many models and would like to have direct feedback on the effect of inclusion of variables, but by showing results sequentially, the mental requirements on the researcher can easily be too high. Also, time constraints might play a role.

To remedy this, we introduce the optional covariates drop panel (D) in which the clinical researcher can drag variables from (A). Panel (B), (C) and (D) show histograms, along with basic descriptive statistics upon hovering, for all covariate distributions to allow for basic checks and interpretation during analysis. After each change, new models are generated (DG2, DG4, UR10). The fixed covariates in (C) are used in all models and the optional covariates X_{opt} in (D) are used in half of the models in order to generate the $2^{|X_{opt}|}$ multivariate models that are combinatorially possible. Once generated, all of the models are assembled in the multivariate model matrix (E).

Models with a better (i.e., lower) AIC-score, have a more saturated purple color. Models that require more DoFs than allowed for by the underlying (sub)population fail the DoF check, as signaled by black (instead of white) cells in the DoF-column. Corresponding divergent fluctuations in covariate significance levels should be considered less reliable.

Also after each change, an automatic multicolumn table sort is performed from left to right in order to aid the clinical researcher with pattern recognition (UR5). Depending on the status of the sort direction indicator (toggable by clicking on the header) either filled or empty cells are placed on top. Moving a column to the left will increase the priority of that column in the sorting process. Conversely, moving a column to the right will decrease its priority.

Furthermore, a filter can be placed on top of a column to show only filled or empty cells, corresponding to the presence or absence of covariates in the models. Additionally, the matrix can be zoomed

Fig. 4: Predicting math test failure: (A) The subpopulations are generated by splitting on gender and active romantic involvement; (B) Univariate analysis for subpopulations; (C, D) The first and second part of the multivariate analysis. (1, 2, 3) Indication of a mediation, moderation and confounding effect, respectively.

and rows can be pinned in order to uncover patterns more easily (*UR6*, *UR7*, *UR9*).

When comparing models and analyzing covariates effects, the researcher can hover the cells to show the actual values for AIC scores, odds ratios, and p -values. Usually a researcher first compares the p -value for a covariate and then the odds ratio. Sometimes the odds ratios are very close, regardless of p -value, which means that clinically there is not much difference. If odds ratios are close to 1, the effect may be considered clinically irrelevant.

As described, our system enables a clinical researcher to consider multiple alternative models in parallel, which is in contrast with the stepwise method, where a clinical researcher considers one or two models at a time that can be defined with a set of fixed covariates and one additional optional covariate under consideration (*UR6*). As a result, a considerably larger number of models can be generated and compared in a shorter amount of time. In turn, the clinical researcher can focus on global patterns which leads to more effective gathering of insights (*DG4*).

Considering the results shown in (E) in Fig. 3, a first observation reveals the general negative impact on model quality when both the first and second period grades $G1$ and $G2$ are excluded. Also, the models of highest quality have $G2$ included. Note that univariate analysis showed $G2$ to be insignificant, which was put at the bottom of the list in (A), outside the visible area. If $G2$ was not considered important a priori, strict adherence to the stepwise method would have discarded $G2$ from the start. This shows why exploratory analysis can be beneficial in practice. By the same token, the significance for age is reduced or removed. Finally, $failures$ actually becomes significant only when both $G1$ and $G2$ are excluded from the models (a general pattern), as indicated by the isolated block with red rectangles in the center of the matrix.

5.2 Subgroup analysis

Dynamic subgroup analysis is an integral part of RegressionExplorer, instead of a final step as is the case in the typical workflow shown in Fig. 1. Therefore, it is supported throughout the univariate and multivariate analysis such that the researcher is enabled to obtain more insights into the subpopulations without picking a final model beforehand. In the remainder we explain how to select subgroups, briefly revisit the analysis steps and discuss variable relationships.

Population control Using the population control panel (F) shown in Fig. 3, the researcher is enabled to specify a subpopulation tree, which is visualized with an icicle plot. For our running example, consider the situation in Fig. 5. Dropping a variable from the univariate analysis panel (B) onto the population panel results in a partition of the population. By dropping another variable all previously generated subpopulations are partitioned recursively. Obviously, these split variables can also be removed again. In case of a categorical split variable, a population is directly partitioned using the levels of that variable. A continuous variable is transformed into a categorical one using a user-specified number of bins (two by default).

Since the number of subpopulations can increase quickly by adding split variables, we have not used more than two splits in practice. The reasons for this are the increased demand on screen space and frequent drops in the sample sizes. The (sub)population sizes are shown inside the cells of the icicle plot with, in case of low population size (< 20), a warning icon (*DG2*, *UR11*).

A researcher is allowed to select one or two (sub)populations by clicking in the icicle plot. Consequently, histograms shown in the univariate and multivariate analysis panels will be updated. If two populations are selected, two histograms are shown on top of each other, with the overlapping areas colored depending on the statistical similarity of the two distributions (*DG2*, *UR12*), see Fig. 5. Two identical distributions yield a complete overlap of the histograms, which

Fig. 5: Population control and distributions. (Top) Partitioning of the population into subpopulations, of which two have been selected: females (sex=0) and males (sex=1). (Bottom) Distribution of covariates *G1* (first period grade) and *Walc* (weekend alcohol consumption) across the two selected subpopulations. The colors used adhere to the legend shown in Fig. 3.

is colored light-green to indicate no (statistical) difference between the two distributions. For two non-identical distributions that are still considered statistically similar, the overlapping segments are also colored light-green. If two non-identical distributions are tested to be statistically dissimilar, the overlapping segments are colored gray. The similarity of two distributions of continuous and categorical covariates are tested with the Kolmogorov-Smirnov test and chi-squared test, respectively. If two distributions are not similar, bias may be introduced due to the imbalance.

Univariate- and multivariate analysis For both univariate- and multivariate analysis the basic approach remains the same when dealing with subpopulations. The main difference is that the cells used to indicate significant effects are now subdivided into subcells, reusing the structure of the icicle plot in the population selector panel, as shown in Fig. 4 (UR5). The updated univariate analysis panel (Fig. 4B) now shows the results across subpopulations, but the variables are still sorted on their significance levels in the main population (UR3).

Analogously, in the multivariate analysis matrix (see Fig. 4C, Fig. 4D) each row now reflects the model structure used to compute individual models across all subpopulations (UR5). The significance levels of a covariate now depend on the subpopulation and are encoded inside the cell in the appropriate position, as dictated by the icicle plot for the subpopulation tree. For instance, a model that is computed for the main population is always encoded on the top of each icicle plot inside the cells.

Confounding, mediation and moderation The theoretical assumption that all covariates are independent does often not hold up in practice, where associations between the recorded, unrecorded or even unknown covariates are present. If unaccounted for, this can lead to bias in the model. The diagram in Fig. 2 shows the most basic situations. Starting with the notion of a confounder, we find that the definition varies and proof for confounding may require strong assumptions [27]. In practice, it is also common to provide evidence for confounding based on causal reasoning using specific domain knowledge.

Since we are dealing with exploratory data analysis and are primarily concerned with the variable selection problem in the clinical domain, we are interested mainly in hinting at confounding. For this purpose we look at changes in significance levels across similar models. It is, however, known that significance testing has limited performance with regards to detecting confounding and common (clinical research) practice therefore relies on the use of a change in estimates effect (or odds ratio) as a basic detection mechanism for confounding [4, 15]. This means that if the removal or addition of a covariate from or to the model

causes a difference in odds ratio of more than 10% for another covariate, these two covariates are considered confounded. More recently even this method is increasingly being criticized [27].

In our approach, by generating many models we aim at finding some trends across the various models. In order to reduce the false positives though, the user is enabled to observe the change in odds ratios for highly similar models to provide evidence for confounding at least up to the current standard in clinical research. To explain this, please reconsider first Fig. 3. The changes in significance levels are color coded across the multivariate analysis matrix. Consider the first 8 models and notice change in significance of *G1* due to the removal of *G2*. This may indicate a confounding effect between *G1* and *G2*.

For more evidence we inspect the change in odds ratio for *G1* when dropping *G2* from the first four models in pairwise fashion with the next four and find that the change in odds ratio for *G1* varying between 40 and 50 percent across the pairs of models. This percentage is automatically calculated and added in the standard popup by the system when a model is selected and the user hovers over the cell of a covariate belonging to another model. Analogously, the change in significance levels can be spotted when subpopulations are defined, see (3) in Fig. 4. Inspection of the odds ratios for *higher* indeed confirmed that *higher* and *failures* are confounded.

A mediator is similar to a confounder, except for a change in causal direction (Fig. 2). A mediator can therefore act as a confounder and based on domain knowledge the distinction can be made plausible. There is, however, one particular pattern that provides evidence for mediation more strongly: after a mediator has been added to the model, the effect of a mediated covariate becomes invisible, which is the case with *schoolsup* (extra educational support) and *G1*, as indicated by (1) in Fig. 4.

Finally, a moderator changes the effect of another covariate as a function of itself. One way to spot these changes is through analysis of subpopulations. There are obviously many variations present in the cells shown in Fig. 4 suggesting some moderation through *gender*. Possibly some moderation takes place through *romantic* as well. Two cases stand out more: *reason* (for attending the school) in the univariate analysis panel (B) and *Walc* (weekend alcohol consumption) as indicated by (2) in (D). For the latter case we would like to remind the reader about imbalance in covariates. As shown in Fig. 5, the distribution of *Walc* is statistically different across *gender* which can also have caused the difference.

6 USE CASES

We demonstrate the system for two real clinical use cases. The first case is about predicting cardiac conduction disorder in patients after receiving a heart valve implant, which is a blockage in the electrical current in the heart. The second case is about predicting hyponatremia in critically ill patients, meaning an inappropriate salt / water balance in the body of patients in the intensive care unit.

6.1 Context

The setup was to invite clinical researchers who were studying potential impact factors on a given clinical condition with the goal to later develop a model that is both parsimonious and yet powerful. Such a model is easier to use in practice for clinical decision making. In some cases *risk charts* are generated that show the impact on a clinical condition if important factors are changed (by intervention). Understanding how different models perform depending on the included predictors is therefore key. An interactive tutorial (15 min) was given with some example cases to make sure the system was understood properly. Thereafter, the clinical researchers used the system independently, with the first author available for clarification when needed. The sessions took about one hour.

6.2 Cardiac conduction disorder

The dataset used for the first use case contains 1031 patients and 61 variables, including the responder. A total of 278 cases of cardiac conduction disorder have been recorded and stored in the variable *left bundle branch block (LBBB)*.

Fig. 6: Use case: Cardiac conduction disorder. (Top) Multivariate analysis. (Bottom) Subgroup analysis by splitting on *gender*.

Hypothesis The clinical researcher wants to gain a better understanding of the context in which *LBBB* occurs. It is suspected that the valve implant type, *valve type*, has an effect on *LBBB*. Also, the interrelationships between *valve size*, *valve type* and *valve delivery device* (*catheter*) were of interest. Finally, standard risk factors *age* and *gender* needed to be considered.

Univariate analysis From literature it is known that *age*, *gender*, *treated coronary artery disease* (*PCI*) and *heart function* (*LVF*) might be associated with higher risk for *LBBB*. Only *gender* and *LVF* showed to be significant during univariate analysis. It was, however, decided to include all variables that were of interest initially in the analysis. The variable indicating risk of complication during surgery, *euroscore*, showed to be statistically significant and was therefore also chosen to be included in the analysis.

Multivariate analysis The first observation made was that bigger *valve size* was likely associated with an increased risk as indicated by many red cells in that column, while *valve type2* was protective (see Fig. 6). However, since *valve type* is a categorical variable, *valve type2* is defined relative to *valve type1*, which indicates that *valve type2* is preferred regarding *LBBB*. The optional covariate *gender* seemed to fluctuate in a checkerboard type of pattern and we decided to move that column to the left, giving it a higher priority in the sorting procedure. A similar pattern was visible for *valve size*, so we moved the column right next to *gender*. There seemed to be a relationship between the two. *Valve type* exhibited the inverse pattern and was more significant when both *gender* and *valve size* were less significant and vice versa. Next, we moved *gender* to the far right to compare odds ratios of *valve size* more easily. We found that *gender* and *valve size* are indeed confounded. This indicated mediation of *valve size* through *valve type*, but it was then realized that *valve type* actually represented three physical devices of different sizes in just two levels. This should be recoded when building the actual model to avoid bias. Finally, covariate *catheter* seemed to fluctuate in conjunction with *gender* and *valve type*.

Subgroup analysis Continuing with our goal to understand the relationships between *catheter*, *valve size* and *gender*, it was decided to perform subgroup analysis by splitting on *gender*, as shown in Fig. 6. We verified that all covariates had similar distributions across *gender* using the population selector panel. We noted that excluding *valve type* tends to result in models of less quality. This was confirmed by filtering models on exclusion of *valve type*. Furthermore, we have observed that *catheter* responded significantly different, depending on *gender*, as did *LVF* and *valve type*.

In conclusion, we prefer inclusion of *valve type* over *valve size* for higher model quality. Furthermore, if *catheter* is important, consider developing separate models for men and women.

Fig. 7: Use case: Hypernatremia. (Top) Multivariate analysis. (Bottom) Subgroup analysis by splitting on *lis* and then on *gender*.

6.3 Hypernatremia

The dataset for the second case contains 6554 patients for which 70 variables have been recorded, including the responder — the occurrence of *hypernatremia* (*HN*). In total 1339 patients in the dataset have had this condition.

Hypothesis The clinical researcher wants to better understand the factors related to *HN*. A first question is whether kidney condition is of influence, measured with *kreatinine* and *ureum*. It is suspected that more ill patients with impaired blood circulation have a higher risk of developing *HN*. Variables that capture this are: *vaso*, *inotropie* and *lactaat*. Other variables of interest were a health score, calculated at the time of admission to the hospital, called *apache score* and a variable *bicarbonaat* that holds the pH level. Since *age* and *gender* are commonly known for confounding, these are to be checked. The final question was whether the administration of medication designed to make patients pee out salt, indicated by *lis*, has an influence.

Univariate analysis All variables, except *age* and *gender*, were significant. All significant variables and *age* (based on clinical relevance) were included in the analysis.

Multivariate analysis The results of multivariate analyses is shown in Fig. 7. It was immediately noticed that *age* (“leeftijd” in the figure) generally showed a protective effect when *kreatinine* was either excluded or also showed a protective effect. A possible interpretation is that when *kreatinine* and *ureum* are both increased, renal replacement therapy is started, which might protect against *HN* depending on age (in fact, we found out later that a clinical study was just starting on this specific question in the hospital). With regard to impaired blood circulation: *inotropie* typically shows a significant effect (odds ratio between 1.2 and 1.7), but not when *vaso* is included, which comes with even higher odds ratios. The clinical expert was surprised and considers it plausible that *inotropie* (the force of muscular contractions) is being mediated through *vaso* (narrowing of vessels).

Subgroup analysis We split the population using *lis* and started analyzing. Later we were curious if *gender* had significant effects when used as a second split, which turned out not to be the case (see Fig. 7). After both splits we checked that all covariates had similar distributions across all subpopulations using the population selector panel.

The first observation was that *age* (not significant during univariate analysis) structurally showed a moderation effect after the split on

lis. A second observation was that model quality tends to increase if *bicarbonaat* is included. The last observation is that after receiving *lis*, the levels of *creatinine* are less predictive (risk factor) for the development of *HN*.

Based on the observations, we recommend to include *bicarbonaat* when developing a model. In addition, we suggest to develop separate models for patient groups depending on whether medication was administered (*lis*).

6.4 Evaluation

Most time was spent on interpreting the result clinically, while the effective interaction time with the system was less than 10 minutes. Researchers tend to focus on their main question and accordingly select fixed and optional covariates. The main insights are then noted and clinical discussions start. Once these are settled for the moment, direct implications of the intermediate conclusions are explored. In the next iteration, secondary questions become the focus and finally a more playful attitude is taken to explore the dataset further.

The covariate selection panels were used with no further explanation. Also, the icicle plot seemed to be very intuitive and needed little explanation. The encoding used in univariate and multivariate analysis panel was almost straightforward if no subpopulations were defined. The reordering and filtering of the columns was the first step that required some training. Similarly, adding a single split in the population control panel and later a second split each required some familiarization as well. Note that the various colormaps used throughout the system are related to different aspects, which are focused on intentionally. Therefore, it seems, the colors did not overwhelm the researchers.

7 DISCUSSION AND LIMITATIONS

All clinical researchers who were introduced to the system noted the time-saving potential of our system and only little time was needed before it became intuitive for them to use. Currently, a clinical researcher considers perhaps 10-20 models in a study by hand. Using our system it takes a couple of minutes to obtain a basic understanding of the dataset in terms of covariate effects in hundreds to thousands of models.

Generally, we have observed structural impact on model quality and many consistent fluctuations across subgroups, but also some deviating ones. These observations are often missed in current practice, where subgroup analysis is typically performed in a very limited fashion after a final model has already been picked. Note that the generally structured nature of the fluctuations in model properties is what makes our solution feasible.

Furthermore, the clinical researchers tended to use our system to explore the interrelationships between variables of interest by marking them all as optional as a first step. This approach comes with a limitation on the number of optional variables due to the exponential growth in the number of models and associated computational costs to generate these. On a regular notebook we could interactively deal with about 12 optional covariates. If more variables are to be considered, the clinical researcher may be forced to mark some of the variables as fixed first to reduce the number of models. If subpopulations are defined, all models need to be estimated for all of these, hence the exponential running time is scaled up linearly in the number of subpopulations.

The latter leads us to another limitation of the system: the number of subpopulations that can be considered. Adding more than two splits generally leads to many subpopulations that can hardly be distinguished anymore in the icicle plots. We recommend using no more than two split variables, each yielding two to three subgroups. Statistically speaking, adding more splits increases the likelihood of generating unbalanced subgroups, even if sample sizes are still sufficient. For this reason we had to perform basic checks on the histograms to make sure that distributions of covariates are similar across subpopulations. Furthermore, in clinical practice subpopulation analysis is often time consuming to perform, especially when considering multiple models. Therefore, automatic support for two successive splits is already an advancement in our view.

When exploring the models to find patterns, the clinical researcher usually aims for something. Sometimes the researcher notes an interest-

ing pattern on the side. This suggests that the system may be improved by adding more sophisticated support for relating multiple findings, perhaps in combination with more automated methods to point out interesting areas. Also, if there are many categorical variables, many columns are needed in the matrix. In those cases, the current interaction support through reordering of columns, zooming, etc. only helps to a certain extent.

Finally, the system by Zhang et al. [31] also uses a matrix approach. In comparison, we have chosen to use rows for different models and columns for covariates, and not vice versa, to support sorting and filtering in a more familiar way. The main difference between the systems, however, is that their system is more geared towards building a model using the stepwise method and uses a separate mechanism to show similar variables. Also, both magnitude (odds ratio) and significance are encoded inside a cell, which yields an additional column per individual subpopulation in the univariate analysis panel. For multivariate analysis this is not possible because it would require an extra dimension in the matrix.

Since we wanted to focus on global patterns across multiple models we decided to drop the magnitude and go for significance alone in the cells, which allowed for the icicle plots to be embedded. We feel the system developed by Zhang et al. is better suited for model development using the stepwise method, while our system is better suited for uncovering global patterns across many alternative models in parallel, which inherently reduces the probability of selecting a model from the model space that is optimal only locally.

8 CONCLUSION

We presented RegressionExplorer, a system that augments the existing workflow that a clinical researcher follows for building logistic regression models. With RegressionExplorer relatively many models can be analyzed at once in order to identify patterns in model properties and covariate effects. This leads to insights much more quickly, which became evident during the analysis of the use cases.

The gathered insights can be used as input for the standard model building process or clinical decision making. The tight integration of subgroup analysis at every step of the analysis significantly adds to the complexity level of the gathered insights and initial verification of hypotheses.

Future work We envision the first extension of the system to be an automated subgroup balancing method that for a given set of variables balances out the effect across these groups by either subpopulation matching or adjusting the models accordingly. The selection method for selecting the variables, on which to perform the balancing, could be analogous to covariate selection method. The multivariate model matrix could then also be reused to encode the properties of different possible sets of these variables.

Another improvement could be to integrate automated methods that guide the clinical researcher during the exploration, along with better support for saving and comparing intermediate findings. Additionally, variable relationships and interpretations could be defined as feedback to the system. The system could then react by, for instance, adding an interaction term to the models that captures moderation effects at the model level. Finally, more model diagnostics could be integrated and shown on demand or used as a filtering mechanism.

ACKNOWLEDGMENTS

This work is the result of a collaboration between Eindhoven University, Philips and the Catharina Hospital Eindhoven. Our study was approved by the institutional review board. Finally, the authors wish to thank Tom Djajadiningrat (Philips), Rick Bezemer (Philips), Ashley de Bie (Catharina Hospital Eindhoven), Ineke Neutelings (Eindhoven University of Technology), Ron Ceelen (Philips) and Sara Cabré Bargalló for their support.

REFERENCES

- [1] J. Bernard, D. Sessler, T. May, T. Schlomm, D. Pehrke, and J. Kohlhammer. A Visual-interactive System for Prostate Cancer Stratifications. *Visualizing Electronic Health Record Data (Workshop)*, 35(3):44–55, 2015.

- [2] D. Bowers. *Medical Statistics from Scratch: An Introduction for Health Professionals*, pp. 288–289, 303–304. Wiley, 2014.
- [3] D. Bowers. *Medical Statistics from Scratch: An Introduction for Health Professionals*, p. 285. Wiley, third ed., 2014.
- [4] D. Bowers. *Medical Statistics from Scratch: An Introduction for Health Professionals*, p. 289. Wiley, 2014.
- [5] J. J. Caban and D. Gotz. Visual analytics in healthcare - opportunities and research challenges. *Journal of the American Medical Informatics Association*, 22(2):260–262, 2015. doi: 10.1093/jamia/ocv006
- [6] Y.-H. Chan, C. D. Correa, and K.-L. Ma. Regression Cube. *ACM Transactions on Interactive Intelligent Systems*, 4(1):1–32, 2014. doi: 10.1145/2590349
- [7] P. Cortez and A. Silva. Using data mining to predict secondary school student performance. In A. Brito and J. Teixeira Eds., *Proceedings of 5th Future Business Technology Conference (FUBUTEC 2008)*, pp. 5–12, april 2008.
- [8] M. El-Assady, W. Jentner, M. Stein, F. Fischer, T. Schreck, and D. Keim. Predictive visual analytics: Approaches for movie ratings and discussion of open research challenges. In *An IEEE VIS 2014 Workshop: Visualization for Predictive Analytics; Proceedings*, 2014.
- [9] M. Friendly. Corrgrams: Exploratory displays for correlation matrices. *American Statistician*, 56(4):316–324, 2002. doi: 10.1198/000313002533
- [10] D. Gotz and D. Borland. Data-Driven Healthcare: Challenges and Opportunities for Interactive Visualization. *IEEE Computer Graphics and Applications*, 36(3):90–96, 2016. doi: 10.1109/MCG.2016.59
- [11] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant. *Applied Logistic Regression*, p. 1. John Wiley & Sons, third ed., 2013.
- [12] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant. *Applied Logistic Regression*, p. 90. John Wiley & Sons, third ed., 2013.
- [13] J. Kehler and R. N. Boubela. A Generic Model for the Integration of Interactive Visualization and Statistical Computing Using R. *IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 133–134, 2012.
- [14] J. Krause, A. Dasgupta, J. Swartz, Y. Aphinyanaphongs, and E. Bertini. A workflow for visual diagnostics of binary classifiers using instance-level explanations. *IEEE Conference on Visual Analytics Science and Technology (VAST)*, Oct 2017.
- [15] P. H. Lee. Should we adjust for a confounder if empirical and theoretical criteria yield contradictory results? A simulation study. *Scientific Reports*, 4:1–14, 2014. doi: 10.1038/srep06085
- [16] M. H. Loorak, C. Perin, N. Kamal, M. Hill, and S. Carpendale. TimeSpan: Using Visualization to Explore Temporal Multi-dimensional Data of Stroke Patients. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):409–418, 2016. doi: 10.1109/TVCG.2015.2467325
- [17] J. Lu, W. Chen, Y. Ma, J. Ke, Z. Li, F. Zhang, and R. Maciejewski. Recent progress and trends in predictive visual analytics. *Frontiers of Computer Science*, 11(2):192–207, apr 2017. doi: 10.1007/s11704-016-6028-y
- [18] Y. Lu, R. Garcia, B. Hansen, M. Gleicher, and R. Maciejewski. The State-of-the-Art in Predictive Visual Analytics. *Computer Graphics Forum*, 36(3):539–562, 2017. doi: 10.1111/cgf.13210
- [19] S. Malik, F. Du, M. Monroe, E. Onukwugha, C. Plaisant, and B. Shneiderman. Cohort Comparison of Event Sequences with Balanced Integration of Visual Analytics and Statistics. *Proceedings of the 20th International Conference on Intelligent User Interfaces - IUI '15*, pp. 38–49, 2015. doi: 10.1145/2678025.2701407
- [20] T. Mühlbacher and H. Piringer. A partition-based framework for building and validating regression models. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):1962–1971, 2013. doi: 10.1109/TVCG.2013.125
- [21] H. Piringer, W. Berger, and J. Krasser. HyperMoVal: Interactive visual validation of regression models for real-time simulation. *Computer Graphics Forum*, 29(3):983–992, 2010. doi: 10.1111/j.1467-8659.2009.01684.x
- [22] A. Rind. Interactive Information Visualization to Explore and Query Electronic Health Records. *Foundations and Trends® in HumanComputer Interaction*, 5(3):207–298, 2013. doi: 10.1561/11000000039
- [23] A. Rind, W. Aigner, S. Miksch, S. Wiltner, M. Pohl, F. Drexler, B. Neubauer, and N. Suchy. Visually exploring multivariate trends in patient cohorts using animated scatter plots. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6779 LNCS(1):139–148, 2011. doi: 10.1007/978-3-642-21716-6_15
- [24] E. S. Shtatland, K. Kleinman, and E. M. Cain. A new strategy of model building in proc logistic with automated variable selection, validation, shrinkage and model averaging. *SAS Users Group International (SUGI)*, 29(191-29):1–10, 2002.
- [25] M. S. Thiese, B. Ronna, and U. Ott. P value interpretations and considerations. *Journal of Thoracic Disease*, 8(9):928–931, 2016. doi: 10.21037/jtd.2016.08.16
- [26] S. van den Elzen and J. J. van Wijk. BaobabView: Interactive construction and analysis of decision trees. *IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 151–160, 2011. doi: 10.1109/VAST.2011.6102453
- [27] T. J. Vander Weele and I. Shpitser. On the definition of a confounder. *Annals of Statistics*, 41(1):196–220, 2013. doi: 10.1214/12-AOS1058
- [28] J. Wang and K. Mueller. The Visual Causality Analyst: An Interactive Interface for Causal Reasoning. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):230–239, 2016. doi: 10.1109/TVCG.2015.2467931
- [29] S. V. Wang, M. He, Y. Jin, R. Wyss, H. J. Shin, Y. Ma, S. Keeton, B. Fireman, S. Karami, J. M. Major, S. Schneeweiss, and J. J. Gagne. A review of the performance of different methods for propensity score matched subgroup analyses and a summary of their application in peer-reviewed research studies. *Pharmacoepidemiology and Drug Safety*, 26(12):1507–1512, dec 2017. doi: 10.1002/pds.4328
- [30] X. M. Wang, T. Y. Zhang, Y. X. Ma, J. Xia, and W. Chen. A Survey of Visual Analytic Pipelines. *Journal of Computer Science and Technology*, 31(4):787–804, 2016. doi: 10.1007/s11390-016-1663-1
- [31] C. Zhang, J. Yang, F. B. Zhan, X. Gong, J. D. Brender, P. H. Langlois, S. Barlowe, and Y. Zhao. A visual analytics approach to high-dimensional logistic regression modeling and its application to an environmental health study. In *2016 IEEE Pacific Visualization Symposium (PacificVis)*, pp. 136–143, April 2016. doi: 10.1109/PACIFICVIS.2016.7465261
- [32] Z. Zhang, K. T. McDonnell, E. Zadok, and K. Mueller. Visual correlation analysis of numerical and categorical data on the correlation map. *IEEE Transactions on Visualization and Computer Graphics*, 21(2):289–303, 2015. doi: 10.1109/TVCG.2014.2350494