Q-Finder: An Algorithm for Credible Subgroup Discovery in Clinical Data Analysis An Application to the International Diabetes Management Practice Study (IDMPS)

Cyril Esnault, May-Line Gadonna, Maxence Queyrel, Alexandre Templier, Jean-Daniel Zucker

Q-Finder: an algorithm for credible subgroup discovery in clinical data analysis — An application to the International Diabetes Management Practice Study (IDMPS)

Cyril Esnault[1], May-Line Gadonna[1], Maxence Queyrel[1,2], Alexandre Templier[1], and Jean-Daniel Zucker[2,3]

[1]*Quinten France, 8 rue Vernier, 75017, Paris France*
[2]*Sorbonne University, IRD, UMMISCO, F-93143, Bondy, France*
[3]*Sorbonne University, INSERM, NUTRIOMICS, F-75013, Paris, France*
Correspondence:
Cyril Esnault, cyrilesnault9@gmail.com
Jean-Daniel Zucker, Jean-Daniel.zucker@ird.fr

Contents

Abstract

Addressing the heterogeneity of both the outcome of a disease and the treatment response to an intervention is a mandatory pathway for regulatory approval of medicines. In randomized clinical trials (RCT), confirmatory subgroup analyses focus on the assessment of drugs in predefined subgroups, while exploratory ones allow a posteriori the identification of subsets of patients who respond differently. Within the latter area, subgroup discovery (SD) data mining approach is widely used — particularly in precision medicine — to evaluate treatment effect across different groups of patients from various data sources (be it from clinical trials or real-world data). However, both the limited consideration by standard SD algorithms of recommended criteria to define credible subgroups and the lack of statistical power of the findings after correcting for multiple testing hinder the generation of hypothesis and their acceptance by healthcare authorities and practitioners. In this paper, we present the Q-Finder algorithm that aims to generate statistically credible subgroups to answer clinical questions, such as finding drivers of natural disease progression or treatment response. It combines an exhaustive search with a cascade of filters based on metrics assessing key credibility criteria, including relative risk reduction assessment, adjustment on confounding factors, individual feature's contribution to the subgroup's effect, interaction tests for assessing between-subgroup treatment effect interactions and tests adjustment (multiple testing). This allows Q-Finder to directly target and assess subgroups on recommended credibility criteria. The top-k credible subgroups are then selected, while accounting for subgroups' diversity and, possibly, clinical relevance. Those subgroups are tested on independent data to assess their consistency across databases, while preserving statistical power by limiting the number of tests. To illustrate this algorithm, we applied it on the database of the International Diabetes Management Practice Study (IDMPS) to better understand the drivers of improved glycemic control and rate of episodes of hypoglycemia in type 2 diabetics patients. We compared Q-Finder with state-of-the-art approaches from both Subgroup Identification and Knowledge Discovery in Databases literature. The results demonstrate its ability to identify and support a short list of highly credible and diverse data-driven subgroups for both prognostic and predictive tasks.

Keywords: Exploratory subgroup analysis, Subgroup discovery, precision medicine, Predictive factor, Prognostic factor, Credibility criteria, hypothesis generation, IDMPS

1 Introduction

Searching for subgroups of items with properties that differentiate them from others is a very general task in data analysis. There are a large number of methods for finding these subgroups that have been developed in different areas of research. Depending on the field of application, the algorithms considered differ in particular on the metrics used to qualify the groups of interest. The field of medicine is one of those where the search for subgroups has had the most applications. Indeed, the considerable heterogeneity in disease manifestation and response to treatment remains a major challenge in medicine. Understanding what drives such differences is critical to adjust treatment strategies, guide drug development, and gain insights into disease progression.

Targeting certain patient populations that would benefit from a particular treatment is becoming an important goal of precision medicine (*Loh* et al. 2019; *Korepanova* 2018)). Subgroup analysis (SA) can be used to identify the drivers of this heterogeneity. While confirmatory analyses focus on the assessment of predefined subgroups, exploratory analyses rely on identifying the most promising ones. Exploratory SA is itself divided into two types of approaches, depending on whether it is hypothesis-based or data-driven. In the latter, the analysis is called subgroup discovery (SD). It is widely used to evaluate treatment effect across different groups of patients from various data sources — be it from clinical trials, or real world data. Demonstrating a response to an intervention is a mandatory pathway for regulatory approval of medicines. However, both the limited consideration by standard algorithms of recommended criteria to assess subgroups credibility, or the findings' lack of statistical power after correcting for multiple testing, hinder the hypothesis generation process and the acceptance of such analyses by healthcare authorities and practitioners (*Mayer* et al. 2015). In this paper we present Q-Finder, which draws from two families of approaches: the first is Subgroup Identification (SI) and the second is Knowledge Discovery in Databases (KDD).

In the sequel of this section we first place SD in the context of SA used in clinical studies. We then detail the different SA tasks in clinical research. More specifically we propose a new classification of SD tasks in a wider context including both SI and KDD which supports presenting a state-of-the-art of SD approaches.

We conclude this section by presenting the limits of SD algorithms in the context of clinical research. In section 2 we describe the Q-Finder algorithm, that was designed to address the main limitations of state-of-the-art SD algorithms. In section 3 we describe the International Diabetes Management Practices Study (IDMPS) database and perform experiments to compare four different algorithms (namely SIDES (*Lipkovich* and *Dmitrienko* 2014), Virtual Twins (*Foster* et al. 2011), CN2-SD (*Lavrač* et al. 2004) and APRIORI-SD (*Kavsek* et al. 2007) on either predictive or prognostic tasks. In section 4 we discuss the results and the differences between Q-Finder and state-of-the-art algorithms. The last section is dedicated to the conclusion and perspectives.

1.1 Subgroup analysis in clinical research

Randomized Clinical Trials (RCTs) aim to test predefined hypotheses and answer specific questions in the context of clinical drug development. Essentially designed to demonstrate treatment efficacy and safety in a given indication using a limited number of patients with homogeneous characteristics, RCTs are performed in heavily controlled experimental conditions in order to maximize chances to obtain results with sufficient statistical power throughout successive trials. RCTs are the gold standard for evaluating treatment outcomes, although real life studies can reveal mismatches between efficacy and effectiveness (*Saturni* et al. 2014). Conversely, Real-World (RW) Data (electronic medical records, claims data, registries), are mainly generated for administrative purposes, going beyond what is normally collected in clinical trial programs, and represents important sources of information for healthcare decision makers.

In both RCT and RW studies, SA are used to test local effects, for instance to account for the heterogeneity in the response to treatment. In particular in RCT, SA "has become a fundamental step in the assessment of evidence from confirmatory (Phase III) clinical trials, where conclusions for the overall study population might not hold" (*Tanniou* et al. 2016). SA include both confirmatory analyses, whose purpose is to confirm predefined hypotheses, and exploratory ones, which aim to generate new knowledge and are exploratory in nature (*Lipkovich* et al. 2016). When considering a set of patients included in a database, a subgroup of patients is any subset characterized by its *extension* (all the patients in the subset, e.g. Patient's ID in {"12345", "45678"}) and its *intension* (a description that characterizes the patients in the subset: e.g. "All the adult women"). In SA, a typical type of subgroups of interest are those whose extension corresponds to patients who respond differently to a new treatment (*Zhang* et al. 2018). A formal definition of subgroups can be found in (*Lipkovich* et al. 2016).

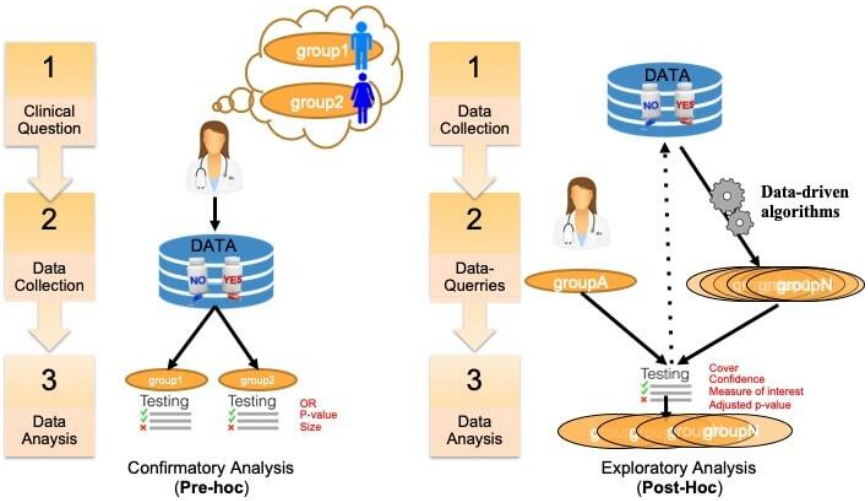

Figure 1: A classification of SA tasks distinguishing the confirmatory analyses (left) from the exploratory ones (right).

1.2 The different Subgroup Analysis tasks in clinical research

A key issue in SA in general is to assess and report its results (*Rothwell* 2005). In clinical trials, this assessment is critical and depends on the precise purpose of the study. There are different ways to distinguish the purpose of using SA in clinical research. A first distinction relates to the general purpose of the analysis that can be either aimed at studying treatment efficacy or safety, on either *a priori* defined groups or *a posteriori* groups. This dichotomous classification is depicted in Figure 1. In the literature, pre-hoc analysis is most-often called confirmatory analysis whereas post-hoc analysis is called exploratory analysis (*Lipkovich* et al. 2016).

More recently *Lipkovich* et al. (2016) have refined this classification into four different tasks:

(A) **Confirmatory subgroup analysis:** refers to statistical analysis mainly aimed at testing a medical hypothesis under optimal setting in the absence of confounding factors while strongly controlling the type 1 error rate (using the Family-Wise Error Rate) in Phase III clinical trials with a small number of prespecified subgroups.

(B) **Exploratory subgroup evaluation:** refers to statistical analysis aimed at weakly controlling the type 1 error rate (using the False Discovery Rate) of a relatively small number of prespecified subgroups that focuses mostly on "treatment-by-covariate interactions and consistency assessments".

(C) **Post-hoc subgroup evaluation:** refers to non-data-driven statistical post-hoc assessments of the treatment effect across small sets of subgroups that include responses to regulatory inquiries, analysis of safety issues, post-marketing activities in Phase IV trials, and assessment of heterogeneity in multi-regional studies.

(D) **Subgroup discovery:** refers to statistical methods aimed at selecting most promising subgroups with enhanced efficacy or desirable safety from a large pool of candidate subgroups. These post-hoc methods employ data mining/machine learning algorithms to help inform the design of future trials.

We propose a decision tree to represent this second classification where the criteria to distinguish pre-hoc analysis is the strength of type 1 error control (strong or weak respectively) while for post-hoc analysis the explicit use of the collected data (hypothesis-driven or data-driven) is considered (see Figure 2).

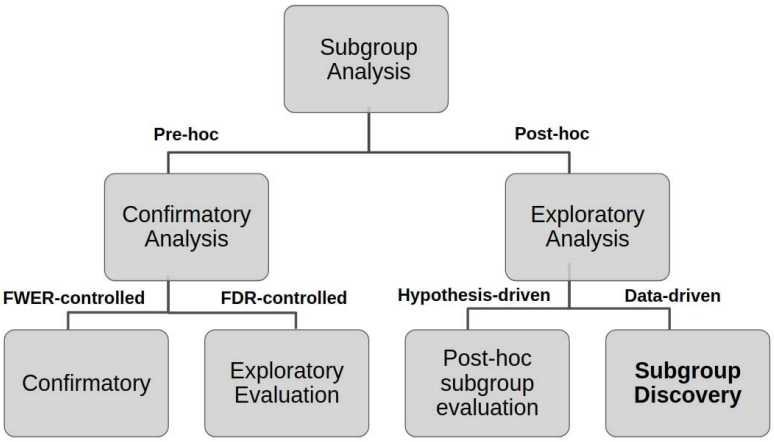

Figure 2: Hierarchical tree representing the two layers classification of SA tasks and criteria used.

The sequel of this paper is concerned with exploratory analysis that are based on Data Mining approaches and known as SD. SD has been used in a large number of applications in the medical field and data analysis of randomized clinical trials (*Sun*, *Ioannidis*, et al. 2014).

1.3 Subgroup discovery: two cultures

Two cultures related to subgroup discovery can be distinguished in the literature. The first one is deeply rooted in medical data analysis, biostatistics and more specifically in the context of drug discovery where both treatments arms and the outcome are key to the analysis. In this domain-specific context (*Lipkovich*, *Dmitrienko*, *Muysers*, et al. 2018; *Lipkovich* et al. 2016), that includes either or both candidate covariates and treatment-by-covariate interactions, SD algorithms search either for:

- a global modeling across the entire covariate space (e.g. Virtual Twins (*Foster* et al. 2011), penalized logistic regression, FindIt (*Imai* et al. 2013), Interaction Trees (*Su* et al. 2009) which extends CART to include treatment-by-covariate interactions, . . .).
- a local modeling that focuses on identifying specific regions with desirable characteristic (e.g. SIDES (*Lipkovich* and *Dmitrienko* 2014), PRIM (*Polonik* et al. 2010), TSDT (*Battioui* et al. 2014), . . .).

The second culture of SD is rooted in the Data Mining and KDD community and applies to any kind of data. The related fields include association rules, set mining, contrast sets, emerging patterns all relating to the notion of descriptive induction (*Fürnkranz* et al. 2012).

Although both cultures share common requirements and issues, their vocabulary differs and are practically mutually exclusive in the SD literature. We propose a hierarchical tree representing both cultures and their main associated algorithms (see Figure 3). Since the Q-Finder approach we propose in this paper inherits from both cultures, it is worthwhile giving an account of both of them.

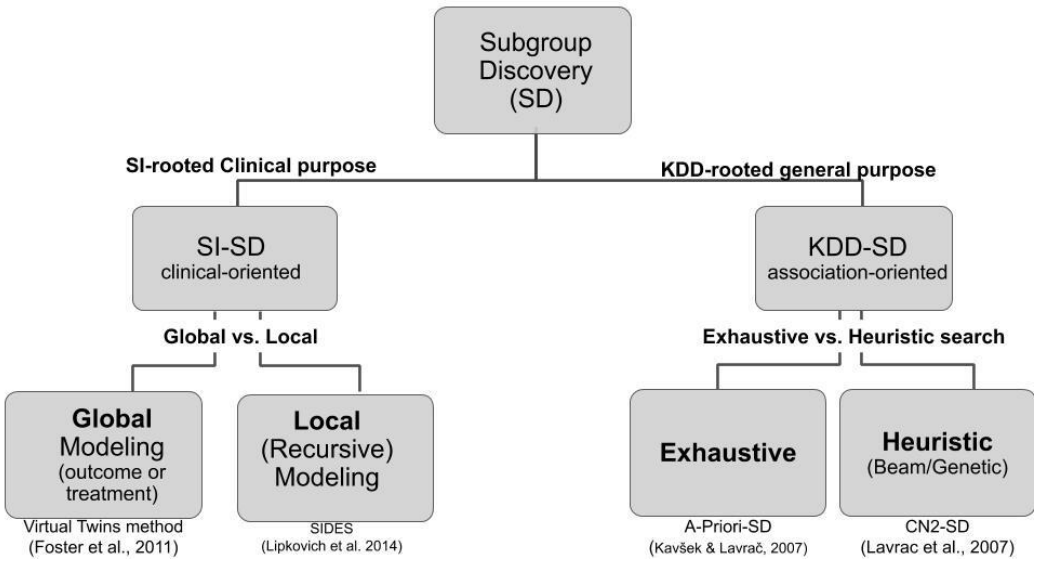

Figure 3: Hierarchical tree representing the SD approaches in both biomedical data analysis and data mining cultures. The references under the boxes correspond to representative algorithms of each kind.

In the first culture, where SD is also often referred to as SI (*Ballarini* et al. 2018; S. *Chen* et al. 2017; *Dimitrienko* et al. 2014; *Huling* et al. 2018; *Lipkovich* et al. 2016; *Lipkovich*, *Dmitrienko*, *Patra*, et al. 2017; *Xu* et al. 2015; *Zhang* et al. 2018), there is a key distinction between prognostic factors (supporting identification of patients with a good or poor outcome regardless of the treatment assignment) and predictive factors (supporting identification of patients' response to the treatment) (*Adolfsson* et al. 2000).

In this culture, SD algorithms[1] can be distinguished depending on whether they search for prognostic and/or predictive factors: the ones that can only look for predictive factors (Quint (*Dusseldorp*, L. *Doove*, et al. 2016), SIDES, Virtual Twins, Interaction trees, ...), the ones that only look for prognostic factors (PRIM, CART (*Hapfelmeier* et al. 2018), ...), and the ones that can look for both prognostic and predictive factors (STIMA (*Dusseldorp*, *Conversano*, et al. 2010), MOB (*Zeileis* et al. 2008), ...). The key measures to assess the quality of the SD results in this culture are p-value, type 1 errors, False-Discovery Rate (*Lipkovich*, *Dmitrienko*, *Muysers*, et al. 2018; *Lipkovich* et al. 2016).

In the second culture, SD is not associated with a specific sector such as clinical research. On the contrary, SD is defined as "given a population of individuals and a property of those individuals that we are interested in, [the finding of] population subgroups that are statistically the 'most interesting', e.g., are as large as possible and have the most unusual statistical (distributional) characteristics with respect to the property of interest" (*Fürnkranz* et al. 2012). More generally, SD "is a type of data mining technique that supports the identification of interesting and comprehensible associations in databases, confirming hypotheses and exploring new ones" (*Atzmueller* 2015). These associations are in the form of a set of rules represented as Subgroup $\rightarrow$ Target, where Target is the property of interest (e.g. $Hypoglycemia = Yes$) and Subgroup is a conjunction of attribute-selector-value triplets (e.g. $Age > 18$ & $Sex = F$). SD belongs to the wider domain of Association Rule mining — this explains why many algorithms bear a name formed from an association rule algorithm and an SD extension — and differs from classical supervised learning as the goal is not to find rules that best predict the target value of unknown observations but rather best support describing groups of observations that when satisfying the condition of a rule also satisfy the target (*Fürnkranz* et al. 2012).

In this second culture the SD process consists in three main phases: candidate subgroup generation, subgroups evaluation and ranking (*Helal* 2016), and subgroups pruning (e.g. top-*k* pruning). The key issues being more related to the algorithmic search for subgroups than their evaluation. This includes the search strategy (be it beam [SD, CN2-SD, Double-Beam-SD], exhaustive [APRIORI-SD, Merge-SD] or genetic [SD-IGA, SGBA-SD]), stopping criterion (minsup, minconf, maxsteps, etc.) (*Valmarska* et al. 2017), pruning technique (constraint, minimum support or coverage) and quality measures (confidence, support, usualness [CN2-SD, APRIORI-SD], etc.).

Recent theoretical and empirical analyses have elucidated different types of methods to select algorithms suitable for specific domains of application (*Helal* 2016). Applying such algorithms to SA requires considering the outcome as the variable of interest. Nevertheless, the treatment is not explicitly considered as a special variable and dozens of quality measures exist (number of rules, number of variables, support, confidence, precision, interest, novelty, significance, false positive, specificity, unusualness (WRAcc), etc.) (*Herrera* et al. 2010).

We will refer to Subgroup Discovery in the context of clinical Subgroup Identification as SI-SD and to Subgroup Discovery in the context of Knowledge Discovery in Database as KDD-SD and compare them with the Q-Finder approach. There is an extensive literature comparing algorithms belonging to each culture independently (e.g. L. L. *Doove* et al. 2013; *Zhang* et al. 2018; *Loh* et al. 2019) but, to our knowledge, they are not compared when they come from two different cultures.

1.4 Limits of current SD algorithms for clinical research

1.4.1 Lack of statistical power and hypothesis generation

As stated by *Burke* et al. (2015) "the limitations of subgroup analysis are well established —false positives due to multiple comparisons, false negatives due to inadequate power, and limited ability to inform individual treatment decisions because patients have multiple characteristics that vary simultaneously". Controlling such errors is a problem: a survey on clinical industry practices and challenges in SD quoted the lack of statistical power to test multiple subgroups as a major challenge (*Mayer* et al. 2015). As a consequence, SI-SD algorithms often fail to detect any "statistically significant" subgroups.

1. We focus here on subgroup discovery algorithms which, unlike classification algorithms, meet the objective of discovering interesting population subgroups rather than maximizing the accuracy of the classification of the induced set of rules (*Lavrač* et al. 2004).

To control for multiple testing errors SI-SD algorithms often rely on approaches that drastically restrict the number of explored candidate subgroups at the expense of hypotheses generation, usually by using recursive partitioning (L. L. *Doove* et al. 2013). Recursive partitioning approaches could miss emerging synergistic effects, defined as subgroups associated to the outcome, whose individual effects (related to each attribute-selector-value triplet) are independent from the outcome (*Hanczar* et al. 2010). As such, individual effects combinations would not be selected in tree nodes. Equally, recursive partitioning may also miss optimal combinations of attribute-selector-value triplets, as an optimal selector-value for a given attribute is only defined with relation to previously defined attribute-selector-value triplets[2] (*Hanczar* et al. 2010). Therefore, subgroups in output are defined by a combination of variables for which thresholds are not necessarily the optimal ones (with respect to the metrics of interest to be optimized). Furthermore, search space restriction strategies favor the detection of the strongest signals in the dataset, that are often already known and/or redundant from each other

Finally, pure beam search strategies could miss relevant subgroups as they try to optimize the joint, i.e. global, accuracy of all leaves, that is a tree with the most heterogeneous leaves. Consequently, when limiting the complexity (i.e. subgroups length), we can miss interesting local structures in favor of the global picture[3] (see section 6.1 in supplementary materials that shows an example where beam search strategy using a decision tree misses relevant subgroups).

On the contrary, KDD-SD approaches support the exploration of much wider search spaces at the expense of accuracy, as they do not in general control for type 1 errors (be it strong or weak).

1.4.2 Insufficient credibility and acceptance of subgroups

The "Achille's heel" of SD is the question of credibility of its results. Several meta-analyses have demonstrated that discovered subgroups rarely lead to expected results and have proposed criteria to assess the credibility of findings (*Rothwell* 2005). Such credibility metrics are key to support confidence in subgroups and their acceptance by regulatory agencies and publication journals. Several credibility metrics have been provided and recommended (*Rothwell* 2005; *Sun*, *Briel*, *Walter*, et al. 2010; *Dijkman* et al. 2009) such as the type of measures of association (relative risk, odds-ratio, ...), correction for confounders, correction for multiple testing, as well as treatment-covariate interaction tests.

SI-SD approaches use credibility metrics suited to clinical analyses. However, most of them only provide and consider in their exploration a limited number of credibility metrics (e.g. hypothesis testing p-value), compared to what is recommended in the literature. Moreover, such metrics are rarely consensual. Equally, the subgroups' generation process (that defines optimal attribute-selector-value triplets combination) mostly relies on the optimization of a limited number of criteria, and is thus not directly driven by all credibility metrics that will be used for the clinical assessment of the subgroups at the end.

On the other hand, KDD-SD can provide a considerable range of credibility metrics as there is no consensus about which quality measures to use (*Herrera* et al. 2010), such as WRAcc, Lift, Conviction, Mutual information (*Hahsler* et al. 2011). However these metrics are seldom used in clinical analyses, hindering their use in the medical field.

Another issue hindering the adoption of SD approaches lies in the comprehensibility of the algorithm itself. This often-underestimated issue is an obstacle for convincing clinical teams and regulatory agencies of the relevance and reliability of SD approaches.

2. Let's assume that a recursive partitioning algorithm has defined $BMI > 25$ as the optimal attribute-selector-value triplet on an objective function to be optimized for patients with $Age > 18$ (the latter being the first triplet to be identified by the algorithm). One can assume that better selector-values could have been obtained for this combination of attributes, to generate the optimal combination of these attributes on the objective function (e.g. $Age > 21$ & $BMI > 20$).
3. Further explanation here: http://www.realkd.org/subgroup-discovery/the-power-of-saying-i-dont-know-an-introduction-to-subgroup-discovery-and-local-modeling/

2 Q-Finder's pipeline to increase credible findings generation

In this section we present an approach that aims at combining some of the advantages of both SI-SD and KDD-SD cultures, while dealing with limitations observed in current SD algorithms (see section 1.4). To this end, we introduce Q-Finder, which relies on a four-steps approach (summarized in Figure 4): exhaustive subgroup candidates generation, candidate subgroups assessment on a set of credibility metrics, selection of a limited number of most promising subgroups that are then tested during the final step.

For further details, an in-depth discussion of Q-Finder is also proposed in section 6.9 of supplementary materials. This approach has been applied in several therapeutic areas, with published examples available (*Alves* et al. 2020; *Mornet* et al. 2020; *Ibald-Mulli* et al. 2019; *Zhou* et al. 2019; *Zhou* et al. 2018; *Rollot* et al. 2018; *Dumontet* et al. 2018; *Gaston-Mathe* et al. 2017; *Dumontet* et al. 2016; *Adam* et al. 2016; *Amrane* et al. 2015; *Eveno* et al. 2014; *Nabholtz* et al. 2012).

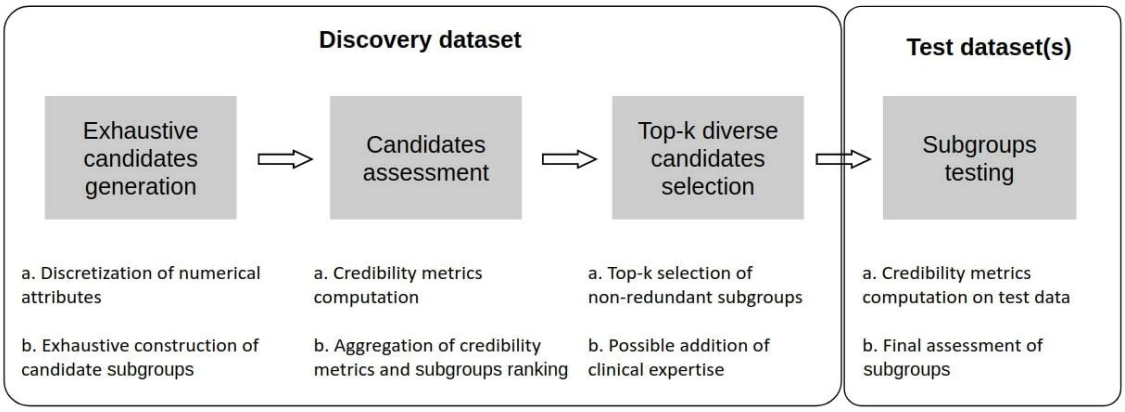

Figure 4: Q-Finder works in 4 main stages: an exhaustive generation of candidate subgroups, a ranking of candidate subgroups via an evaluation of their empirical credibility, a selection of the best candidates (taking into account the redundancy between subgroups) then an assessment of subgroups' credibility on one or more test datasets

2.1 Basic definitions : patterns, predictive and prognostic rules

Numerous formalizations of KDD-SD have been given in the literature. We will briefly introduce some basic definitions of database, individuals, basic patterns, complex patterns, subgroup complexity and subgroup description related to the ones introduced by *Atzmueller* (2015). A database is formally defined as $D = (I, A)$, a set I of N individuals and a set A of K attributes. We will only distinguish nominal and numerical attributes. For nominal attributes, a basic pattern $(a_i = v_{i,j})$ is a Boolean function that is true if the value of attribute $a_i \in A$ is equal to $v_{i,j}$ in the domain of a_i for a given individual of I. For a numerical attribute (be it real or integer) a_i, both basic patterns $(a_i \geq v_{i,j})$ and $(a_i \leq v_{i,j})$ can be defined for each value $v_{i,j}$ in the domain of a_i. The associated Boolean function is defined similarly. The set of all basic patterns is denoted by Σ.

A conjunctive language is classically considered to describe subgroups. An association rule $(X \rightarrow Y)$ is composed of a complex pattern (also called itemset) X and a basic pattern Y, where X is called antecedent (or left-hand-side (LHS) or Subgroup) and Y the consequent (or right-hand-side (RHS) or Target). A complex pattern CP is described by a set of basic patterns $CP = \{BP_1, \ldots, BP_k, \ldots BP_C\}, BP_k \in \Sigma$. It is logically interpreted as a conjunction of basic patterns. In other words a complex pattern CP represents the body of a rule $BP_1 \wedge ... \wedge BP_C$. In Q-Finder, its length C corresponds to the *complexity* of the associated rule. The set of observations covered by a complex pattern CP is called the *extension* of the subgroup, i.e. the individuals for which CP is true $\{x \in I;$ CP is true for x$\}$. In this formalism, the set of all possible association rules

is included in the powerset of Σ although many subsets are not considered because their extension is by construction empty (e.g. $a_i \geq v_{i,j} \wedge a_i \leq v_{i,k}$ when $v_{i,j} > v_{i,k}$). Moreover, this set of all subgroups can be *partially ordered* in a lattice structure (*Ganascia* 1993). We will not rely on such lattice structure because the length of subgroups (i.e. their complexity) is sufficient to partially order the set of generated candidates in subsets[4].

In SI-SD, many databases include information about treatment distinguishing different individuals grouped in arms. This notion is critical to distinguish two types of rules. The prognostic rules are not related to a treatment effect on a given outcome, unlike the predictive rules.

These two main types of rules can be summarized as follows:

Prognostic rule: SUBGROUP $\rightarrow$ TARGET

Predictive rule: SUBGROUP where TREATMENT $\rightarrow$ TARGET

2.2 Preprocessing and Candidate Subgroups generation in Q-Finder

In Q-Finder, to control the size of the set of basic patterns $|\Sigma|$, all numerical attributes are systematically discretized in bins. A hyperparameter $\#Bins$ sets the maximum number of values $v_{i,j}$ of any numeric attribute a_i (default value: 10). If this number is above $\#Bins$, the attribute a_i is quantized using a discretization method $DiscretizationMethod$ (see algorithm 1 line 8). Different methods exist for quantization, the default one being equal-frequency binning. In the same way, the number of distinct values for a given nominal attribute might be bounded by the hyperparameter $\#Cats$ (default value[5]: ∞). If the number of modalities is above this threshold, a reduction method $ReductionMethod$ may be used (by default: use the $(\#Cats - 1)$ most frequent values of a_i and a create a value "other" for all the remaining ones). Let us call Kc the number of nominal attributes and Kb the number of numerical attributes. After this preprocessing step the number of basic patterns $|\Sigma|$ is bounded and we have the relation: $|\Sigma| \leq 2 * Kb * \#Bins + Kc * \#Cats$.

Given a set of basic patterns Σ, we call "candidate generation" the search procedure that generates the subgroups (i.e. complex patterns conjunction of basic ones). The number of complex patterns of complexity C is bounded by the number of C-combination of Σ (i.e. the binomial coefficient $\binom{|\Sigma|}{C}$). There is extensive literature in KDD-SD on the type of exploration of these complex patterns (*Fürnkranz* et al. 2012). Experiments have shown that the exhaustive search based methods perform better than other methods which prune the search before evaluation (*Helal* 2016). This is particularly true when the problem size is reasonable (i.e. a few thousand individuals) which is mostly the case in SD. The Q-Finder Candidate generation is straightforward, it outputs a subset of all *C-combinations* of Σ (with $C \in [\![1; C_{max}]\!]$) as described below in Algorithm 1.

4. A methodology to further order the subgroups is introduced in section 2.3.2
5. In this way, no reduction is done by default.

Algorithm 1 Basic patterns and candidate subgroup generation of complexity $\leq C_{max}$

1: **Input**: D, $\#Bins$, $\#Cats$, C_{max} maximum complexity of generated subgroups, $ReductionMethod$, $DiscretizationMethod$

2: $\Sigma = \{\}$ # Set of basic patterns
3: **For each** nominal attribute a_i **do**
4: **If** $\#valueof(a_i) > \#Cats$ **then**
5: Reduce the number of values of a_i to $\#Cats$ using $ReductionMethod$
6: **For each** $v_{i,j}$ **do** $\Sigma = \Sigma \cup \{(a_i = v_{i,j})\}$
7: **For each** numerical attribute a_i
8: **If** $\#valueof(a_i) > \#Bins$ **then**
9: Discretize the values of a_i in $\#Bins$ using $DiscretizationMethod$
10: **For each** $v_{i,j}$ **do** $\Sigma = \Sigma \cup \{(a_i \geq v_{i,j}), (a_i \leq v_{i,j})\}$
11: $G = \{\}$ # Set of generated subgroups
12: **For each** combination s of 1 to C_{max} elements of Σ **do**
13: **If** one attribute a_i appears twice or more in s or if the extension of s is empty by construction **then skip**
14: **else** $G = G \cup \{s\}$
15: **Output**: G the set of generated candidate subgroups of length $\leq C_{max}$

In practice the Q-Finder algorithm not only supports constructing left-bounded and right-bounded intervals but also supports bounded intervals depending on the number of basic patterns (one or two) associated to a given numerical attribute. If bounded intervals are considered, step 13 of the algorithm becomes "**If** one attribute a_i appears twice or more in s with the same selector or if the extension of s is empty by construction **then skip**".

2.3 Empirical credibility of subgroups

Q-Finder's candidates generation step may potentially produce a very large number of subgroups. Because of its exhaustive strategy, it produces a number of subgroups which grows exponentially with complexity. Dealing with a massive exploration of database is the challenge of any KDD-SD algorithm be it exhaustive or heuristic, as the number of computed statistical tests may induce a high risk of false positives, that needs to be mitigated.

Q-Finder addresses this challenge by only selecting a subset of candidate subgroups and testing them on independent data, to assess the replicability of the results while controlling the number of tests (and thus the type 1 error). This strategy requires to address two issues:

- a way of evaluating the empirical credibility of subgroups, in order to rank them from most to least promising
- a top-*k* selection strategy, in order to select a set of subgroups that seem most credible and will be tested on an independent dataset.

2.3.1 Credibility metrics

The notion of credibility often appears in the literature on subgroup analysis (*Burke* et al. 2015; *Schnell* et al. 2016; *Sun*, *Briel*, *Walter*, et al. 2010; *Sun*, *Briel*, and *Jason* 2012; *Dijkman* et al. 2009) described according to different criteria. In particular *Oxman* et al. (1992) detail seven existing criteria to help clinicians assess the credibility of putative subgroup effects on a continuum from "highly plausible" to "extremely unlikely". *Sun*, *Briel*, *Walter*, et al. (2010) suggest four additional credibility criteria and re-structure a checklist of items addressing study design, analysis, and context. In the present context, credibility is related to a sequence of *a priori* ordered statistical metrics that are progressively increasing the confidence (credibility) of a given

subgroup. The seven criteria described below are aligned with the clinical domain endpoints (*Sun*, *Briel*, *Walter*, et al. 2010; *Dijkman* et al. 2009). Using these criteria when selecting the top-ranked subgroups ought to both promote the finding of credible subgroups and facilitate their acceptance by clinicians, agencies and publication journals.

Drawing from this literature, continuous metrics to measure subgroups' credibility are used in Q-Finder (more details on literature's recommendations in relation to Q-Finder metrics in section 6.9.4 of supplementary materials). Several *credibility criteria* are defined, each composed of both a continuous metric and a minimum or maximum threshold (which may be modified by the user):

1. **Coverage criterion:** The coverage metric is defined by the ratio between the subgroup's size and the dataset's size. This allows to only consider the subgroups that correspond to large enough groups of patients to be clinically relevant. It can be compared to defining a minimum SUPPORT of the antecedent of a rule in the KDD-SD literature. Default minimum threshold for coverage is 10%.

2. **Effect size criterion:** As recommended by both *Sun*, *Briel*, *Walter*, et al. (2010) and *Dijkman* et al. (2009), Q-Finder's exploration relies by default on relative risk reductions, which differ according to the probability distribution of the outcome (ODDS-RATIOS for discrete or negative binomial distributions, RISK-RATIOS for normal or Poisson distributions, HAZARD RATIOS for survival analysis). Those metrics allow to quantify the strength of the association between the antecedent (the subgroup) and consequent (the target) of the rule. Relative risk reductions remain in most situations constant across varying baseline risks, in comparison to absolute risk reductions. In the KDD-SD literature, this continuous metric is usually the CONFIDENCE (i.e. how often the target is true among the individuals that satisfy the subgroups).

   The effect size metric may vary depending on whether one is looking for predictive or prognostic factors. When searching for prognostic factors, Q-Finder only considers the effect size measuring the subgroup's effect (default minimum threshold for effect size is 1.2). When searching for predictive factors, Q-Finder considers simultaneously two effect sizes: the *treatment effect within the subgroup* and the *differential treatment effect*, defined as the difference in treatment effect for patients inside the subgroup versus outside the subgroup (see Table S1 and S2 in supplementary materials Section for an example with odds-ratios). When generating predictive factors, one can consider the *differential treatment effect* on its own, or in combination with the *treatment effect within the subgroup*. The latter case allows to identify subgroups in which the treatment effect is both positive and stronger than outside the subgroup (default thresholds are 1.0 for the *treatment effect within the subgroup* and 1.2 for the *differential treatment effect*).

3. **Effect significance criterion:** the association between each subgroup and the target is assessed using a nullity test from a generalized linear model. For the identification of predictive factors, an interaction test is performed to assess between-subgroup treatment effect interactions as recommended by *Dijkman* et al. (2009). A threshold (typically 5%) is used to define when the p-value related to each effect size metric is considered significant.

4. **Basic patterns contributions criteria:** Basic patterns contributions to the subgroup's global effect are evaluated through two sub-criteria: the absolute contribution of each basic pattern and the contributions ratio between basic patterns.

   The *absolute contribution* of a basic pattern is defined by the improvement in effect when this basic pattern is present, compared to the subgroup's effect when this basic pattern is absent. Each basic pattern contribution should be above a defined threshold (by default 0.2, 0 and 0.2 respectively for the subgroup's effect, the *treatment effect within the subgroup* and the *differential treatment effect*), thus ensuring that each increase in subgroup's complexity goes along with some gain in effect and therefore in interest.

   The *contributions ratio* between basic patterns is the ratio between the maximum *absolute contribution* and the minimum *absolute contribution*. A maximum threshold (by default 5 for the subgroup's effect

or the *differential treatment effect*) is set for this criterion, thus ensuring that basic patterns' contributions to the subgroup's effect are not too unbalanced. Indeed, if a basic pattern bears only a small portion of the global subgroup's effect, then the global effect's increase is not worth the complexity's increase due to this pattern's addition.

5. **Effect size criterion corrected for confounders:** the strength of the association is assessed through relative risk reductions (as in criterion 2) while correcting for confounding factors using a generalized linear model. Added covariates are known confounding factors of the outcome, which are susceptible to be unbalanced between patients within and without each subgroup, as well as between treatment arms for predictive factors identification tasks (*Sun*, *Briel*, *Walter*, et al. 2010; *Dijkman* et al. 2009). As for criterion 2, adjusted relative risks ought to be above a given threshold (same as for criterion 2).

6. **Effect significance criterion corrected for confounders:** as for the effect significance criterion (criterion 3) and using the same model as in criterion 5, a threshold (typically 5%) is used to define when the p-value related to each effect size metric corrected for confounders is considered significant.

7. **Effect adjusted significance criterion corrected for confounders:** the p-value computed in criterion 6 is adjusted to account for multiple testing, as recommended by *Dijkman* et al. (2009). This procedure relies on a Bonferroni or a Benjamini-Hochberg correction to control for type 1 errors. As for criterion 6, a threshold is used to determine whether the p-value remains significant after multiple testing correction (typically 5%)

These seven credibility metrics are at the core of Q-Finder. However, they can be further extended by other measures of interest to better fit each research question.

2.3.2 Aggregation rules and subgroups ranking

Aggregation rules are defined to discriminate subgroups according to a set of criteria and therefore to help select the most interesting and/or promising ones for each research question. This is a key concept of Q-Finder, as the goal is to select a set of "top" subgroups before testing them on an independent dataset, whether or not they pass all credibility criteria. In practice, ranking subgroups into aggregation ranks is helpful when no subgroup passes all credibility criteria, and we need to look into lower aggregation ranks to select the most promising subgroups. This approach contrasts with most SI-SD algorithms, where outputs are only subgroups passing all predefined indicators, hindering the generation of hypotheses if these are difficult to achieve.

To this end, a set of credibility criteria is parameterized by the user, depending on the desired properties of the searched subgroups (see section 2.3.1). Q-Finder computes each metric for each of the candidate subgroups of complexity $C \leq C_{max}$ and verifies if the associated thresholds are met. A vector of Boolean can thus be associated to each subgroup depending on which thresholds are met, and are used to order the candidate subgroups, according to prespecified aggregation rules.

By default, Q-Finder prioritizes subgroups that meet the following credibility criteria: subgroups with a minimal value of coverage (**coverage criterion**), defined by basic patterns that sufficiently contribute to the subgroup's effect (**basic patterns contribution criteria**), with a minimal level of effect size adjusted for confounding factors[6] (**effect size criterion corrected for confounders**) and adjusted p-values for multiple testing below a given level of risk (**effect adjusted significance criterion corrected for confounders**). Please note that the above-mentioned effect could either be the subgroup's effect size (for prognostic factors) or the *treatment effect within the subgroup* and/or the *differential treatment effect* (for predictive factors). Aggregation rules are the following (from least to most stringent):

6. Looking for subgroups with a predefined minimal effect size is aligned with recent recommendations from the American Statistical Association (*Wasserstein* et al. 2019): "Thoughtful research includes careful consideration of the definition of a meaningful effect size. As a researcher you should communicate this up front, before data are collected and analyzed. Then it is just too late as it is easy to justify the observed results after the fact and to over-interpret trivial effect sizes as significant. Many authors in this special issue argue that consideration of the effect size and its 'scientific meaningfulness' is essential for reliable inference (e.g., *Blume* et al. 2018; *Betensky* 2019)."

- Rank 1: subgroups that satisfy the coverage criterion
- Rank 2: subgroups of rank 1 that also satisfy the effect size criterion
- Rank 3: subgroups of rank 2 that also satisfy the basic patterns contribution criteria
- Rank 4: subgroups of rank 3 that also satisfy the effect significance criterion
- Rank 5: subgroups of rank 3 or 4 that also satisfy the effect criterion corrected for confounders
- Rank 6: subgroups of rank 5 that also satisfy the effect significance criterion corrected for confounders
- Rank 7: subgroups of rank 6 that also satisfy the effect adjusted significance criterion corrected for confounders

These aggregation rules can be visualized through a decision tree (see Figure S2 in supplementary materials). One can notice that subgroups with an odds-ratio adjusted for confounders but not significant (rank 5) are ranked before subgroups with significant odds-ratios (not adjusted for confounders, rank 4) for hypotheses generation. This ranking is consistent with favoring adjusted odds-ratios with a lack of statistical power to potential biased estimates. As well as the possibility of adjusting the list of parameters, the order of priority between parameters can also be changed to take into account different priorities.

In addition, a continuous criterion is chosen to sort subgroups of the same aggregation rank. Classically, the criterion called *Effect significance criterion corrected for confounders* is preferred. This is consistent with recommendations by *Sun*, *Briel*, *Walter*, et al. (2010) that state that the smaller the p-value, the more credible the subgroup becomes. In case of a tie, additional criteria can be used to determine the final ranking, such as the *effect size criterion corrected for confounders*, to favor subgroups with stronger effect sizes. This ranking procedure is summarized in algorithm 2.

Algorithm 2 Ranking candidate subgroups

Input: G the list of candidate subgroups of length $\leq C_{max}$,
m_c a continuous credibility metric (*e.g.* a p-value),
M the list of credibility criteria (*e.g.* $[(p\text{-}value < 5\%), (OR > 1)]$),
$AggregationRules$

G_{sorted} = sort(G, m_c) # Sort G according to m_c
$Ranks = \text{rep}(0, |G|)$ # Create a vector of $|G|$ zeros to store ranks of each $s_i \in G$
For s_i **in** G **do**
  $cred = [\,]$ # vector representing the subgroup's credibility
  For m_j **in** M **do**:
    If s_i passes credibility criteria m_j **then**
      $cred[j] = 1$
    Else
      $cred[j] = 0$
  $\lfloor Ranks[i] \rfloor = AggregationRules(cred)$ # Integer part of the rank of s_i is the aggregation rank given by $AggregationRules$ applied to $cred$
  $\{Ranks[i]\} = \text{index}(s_i, G_{sorted})$ # Fractional part of the rank of s_i is the index of s_i in G_{sorted}

Output: G_{ranked} = sort(G, $Ranks$) # The list of subgroups of G sorted according to $Ranks$

2.4 Q-Finder subgroups diversity and top-k selection

2.4.1 Subgroups diversity

Q-Finder performs a subgroups top-k selection to be tested on an independent dataset. One of the known issues in KDD-SD of top-k mining algorithms is that they are prone to output redundant subgroups as each

subgroup is considered individually. Several authors including *Leeuwen* et al. (2012) have argued to search for subgroups that offer a high diversity: diverse subgroup set discovery. Therefore, the goal is to take into account the fact that many subgroups might be redundant either extensionally (their basic patterns are very similar) or intensionally (the objects covered by the subgroup are similar). A general approach to address this issue is to define a redundancy measure. It can for example consider the number of common attributes between two subgroups, or the percentage of common examples covered by two different subgroups. The last requires more computation but results in a better diversification of subgroups as it considers possible correlations between variables.

Q-Finder proposes a definition of intensional redundancy between basic patterns, where two basic patterns (attribute-selector-value triplets, respectively $a_1 - s_1 - v_1$ and $a_2 - s_2 - v_2$) are considered redundant if:

- $a_1 = a_2$
- AND:
  - For nominal attributes: $v_1 = v_2$
  - For numerical attributes:
    - $s_1 = s_2$
    - OR considering s_1 as "$\leq$" and s_2 as "$\geq$", $v_1 \geq v_2$

Based on the basic patterns redundancy definition, two subgroups are called redundant if C_{min} basic patterns are redundant between them; C_{min} being the minimum complexity of the two subgroups.

2.4.2 Selection of top-k subgroups to be tested

Different strategies exist to identify an optimal top-k selection of non-redundant subgroups (*Xiong* et al. 2006), based on subgroups' intensions, extensions, or both. In addition to those existing strategies, Q-Finder proposes its own approach based on subgroups' intensions (see Algorithm 3) to determine an optimal set of k non-redundant subgroups S_k from the ranked set of generated subgroups G_{ranked} (output from Algorithm 2).

The best candidate subgroup is iteratively selected using 2 continuous metrics : m_c from Algorithm 2 and another continuous metric. This top-k algorithm was originally designed using a p-value metric[7] for m_c and an effect size[8] for the second metric[9]. For the sake of clarity, we will describe this algorithm using those 2 metrics:

- Subgroups should be selected from less complex to most complex (favoring less complex subgroups)
- When two subgroups of equal complexity are redundant, only the one associated with the best p-value should be retained.
- When two subgroups of different complexities are redundant
  - The most complex subgroup of the two is discarded iff its chosen effect size metric is lower than the less complex one.
  - The less complex subgroup of the two is discarded iff both its chosen p-value and effect size metric are respectively higher and lower than the more complex one[10]

7. P-value credibility metric can be chosen from metrics 3, 6 or 7 presented in 2.3.1
8. Effect size credibility metric can be chosen from metrics 2 or 5 presented in 2.3.1
9. The user can adapt this algorithm using any relevant continuous metrics' couple
10. Note that instead of discarding the less complex subgroup of the two, one might want to keep both. The algorithm will need to be revised accordingly.

This top-k selection process based on these principles is detailed in Algorithm 3.

Algorithm 3 Q-Finder's iterative top-k selection based on subgroups' intensions

Input: k maximum number of selected subgroups,
G_{ranked} set of ranked generated subgroups, with complexities ranging from C_{min} to C_{max}
δ_{ES} minimum delta to consider that a subgroup has a higher effect size[11]

G_{split} = splitByComplexity(G_{ranked}) # split G_{ranked} by subgroup complexity ($G_{split}[1]$ corresponds to complexity 1, $G_{split}[2]$ to complexity 2, ...)
$S_k = \{\}$ # Initialize S_k, the set of top candidate subgroups
For $c = C_{min}$ **to** C_{max} **do**
  For g **in** $G_{split}[c]$ **do** # g : candidate subgroup
    If p-value(g) > max(p-values(S_k)) **and** size(S_k) == k **then continue** to next c
    For s **in** S_k **do** # s : subgroup in the top-k
      If redundant(g, s) **then**
        If complexity(g) == complexity(s) **then**
          continue to next g
        If complexity(g) > complexity(s) **then**
          If EffectSize(g) $\leq$ EffectSize(s) + δ_{ES} **then**
            continue to next g
    For s **in** S_k **do**
      If redundant(g, s) **and** complexity(g) > complexity(s) **and**
      EffectSize(g) > EffectSize(s) + δ_{ES} **and** p-value(g) < p-value(s) **then**
        $S_k = S_k \backslash \{s\}$
    $S_k = S_k \cup \{g\}$
    while size(S_k) > k **do**
      $S_k = S_k \backslash$ {subgroup from S_k with the highest p-value}

Output: S_k # top-k best candidate subgroups

The result of this step is a set of most promising non-redundant subgroups, that has a maximum size of k.

2.5 Possible addition of clinical expertise

Clinical input can be used to overrule algorithm's preference during top-k selection, by removing candidate subgroups from G_{ranked} (the set of candidate subgroups cf. Algorithm 3) or force the addition of a subgroup into S_k (the set of best candidates cf. Algorithm 3). More generally, clinical experts can directly select top-k relevant subgroups among the most credible ones. This stage, that is sometimes referred to as *Interactive Machine Learning* (*Holzinger* 2016), is aligned with the American Statistical Association recommendations that encourage researchers for seeking experts judgement in any statistical analysis, including for evaluating the importance and the strength of empirical evidence (*Wasserstein* et al. 2019). By integrating experts into Q-Finder's process for subgroups selection, one allows the consideration of non-measurable properties, such as the novelty, interest or applicability of the proposed subgroups[12].

2.6 Subgroups' generalization credibility

In Q-Finder the final step consists in computing the credibility metrics of the top-k subgroups on the testing set, in order to assess their generalization credibility, that is subgroups consistency across databases (*Sun, Briel, Walter*, et al. 2010; *Dijkman* et al. 2009). However, contrary to the candidate subgroups generation

11. Above that delta value, the increase in effect size is worth enough to justify an increase in complexity.
12. *Wasserstein* et al. (2019) argue to be open in study designs and analyses: "One might say that subjectivity is not a problem; it is part of the solution."

phase previously performed, the number of tested subgroups in this phase is well-controlled (as recommended in *Sun*, *Briel*, *Walter*, et al. (2010) and *Dijkman* et al. (2009)), as it is limited by the parameter k. This allows a better control of the type 1 error that was more difficult to achieve until then. For that purpose, Q-Finder performs a correction for multiple testing during computation of the significance metrics, to account for the number of subgroups tested on independent data (default: Benjamini-Hochberg procedure). top-k subgroups satisfying the credibility criteria on the test dataset are considered highly credible.

3 Experiments and Results

This section is dedicated to compare Q-Finder with representative algorithms for predictive or prognostic SD. First, the IDMPS database on which experiments were run is described. Then, the research questions are stated and both a prognostic and a predictive task are described. Lastly, four different methods and their results are given and compared with Q-Finder.

3.1 Introduction of the IDMPS database

The International Diabetes Management Practice Study (IDMPS) database is an ongoing international, observational registry conducted in waves across multiple international centers in developing countries since 2005. Each wave consists of a yearly 2-weeks fact-finding survey, which aims to document in a standardized manner: practice environments, care processes, habits, lifestyle and disease control of patients with diabetes under real world conditions. It has recently led to new findings related to the suboptimal glycemic control in individuals with type 2 diabetes in developing countries and the need to improve organization of care (*Aschner* et al. 2020). Observational registries for patients suffering such conditions are pivotal in understanding disease management. In 2017, an estimated 425 million people were afflicted by diabetes worldwide, with Type 2 Diabetes Mellitus (T2DM) accounting for approximately 90% of cases. By 2030, diabetics should represent 7.7% of the adult population, or 439 million people; and 629 million people by 2045 (L. *Chen* et al. 2012; *Shaw* et al. 2010; *Ogurtsova* et al. 2017). The two most recent waves to date of IDMPS (wave 6 [2013-2014] and wave 7 [2016-2017]) were selected for the following experiments, including data from 24 countries from Africa, Middle East, India, Pakistan, Russia and Ukraine. Only data from patients having T2DM and taking either a Basal insulin, a combination of Basal and Prandial insulin or a Premixed insulin were included.

3.2 Research questions

3.2.1 Prognostic factors identification

One of the main goals of the IDMPS initiative is to evaluate patient's disease management. To do so, a key outcome in diabetes is the blood level of glycated hemoglobin (HbA1c). High HbA1c is a risk factor for micro- and macrovascular complications of diabetes (*Wijngaarden* et al. 2017). Patients with T2DM who reduce their HbA1c level of 1% are 19% less likely to suffer cataracts, 16% less likely to suffer heart failure and 43% less likely to suffer amputation or death due to peripheral vascular disease (*Alomar* et al. 2019, *Susan* et al. 2010).

Given the importance of HbA1c control for diabetics patients, we deemed interesting to focus our prognostic factors detection on patients meeting the recommended HbA1c threshold. This recommended threshold varies depending on several factors, such as age or history of vascular complications. For most T2DM patients, this threshold is set at 7%, which is how we define glycemic control for TD2M patients. Our research question can then be formulated as follows: "What are the prognostic factors of glycemic control in TD2M patients?". We consider the following variables as confounding factors: Patient's age (*Ma* et al. 2016), Gender (*Ma* et al. 2016), BMI (*Candler* et al. 2018), Level of education (*Tshiananga* et al. 2012) and Time since diabetes diagnosis (*Juarez* et al. 2012). Considering the geographical heterogeneity in IDMPS, we added the continent where the data was collected.

This experiment included 1857 patients from IDMPS wave 6 and 2330 patients from IDMPS wave 7, with 63 variables considered as candidate prognostic factors. In wave 6, 17.7% of patients were under the 7% HbA1c threshold, versus 18.8% in wave 7.

3.2.2 Predictive factors identification

Another key outcome in diabetes management is the occurrence of hypoglycemia events, which is one of the main complications linked to diabetes. Hypoglycemia symptoms include dizziness, sweating, shakiness; but can also lead to unconsciousness or death in severe cases. Previous studies have shown the impact of insulin treatments on the incidence of hypoglycemia, including comparing premixed insulin analogues to basal insulin analogues (with or without prandial insulin). In some cases, hypoglycemia rates were found to be slightly higher in patients population treated with premixed insulin analogues (*Petrovski* et al. 2018).

We focused our predictive factors detection on hypoglycemia risk in the past 3 months under premixed insulin versus basal insulin (alone or in combination with prandial insulin).

Our research question can then be formulated as follows: "What are the subgroups in which the treatment effect (premixed insulin versus basal insulin with or without prandial insulin) on the risk of hypoglycemia in the past 3 months is both positive and higher than outside the subgroups?" Illustrative example: "The risk ratio in experiencing hypoglycemia under premixed insulin versus basal insulin (with or without prandial insulin) is greater on male patients than on female patients".

This experiment included 2006 patients from IDMPS wave 6 and 2505 patients from IDMPS wave 7, with 62 variables considered as candidate predictive factors. In wave 6, 32.4% of patients were taking Premixed insulin with a hypoglycemia rate of 32.2%, versus 25.6% for basal insulin regimen. In wave 7, 39.0% of patients were taking Premixed insulin with a hypoglycemia rate of 33.1%, versus 28.3% for basal insulin regimen.

3.3 Analytical strategies

An objective of this paper is to compare the Q-Finder algorithm to state-of-the-art approaches for clinical SD in both SI-SD and KDD-SD. There are a vast number of approaches in both domains, we chose two state-of-the art methods from KDD-SD to address the prognostic factors research, and two methods from SI-SD to address the predictive factors research. Among SI-SD methods, we chose SIDES (Subgroup Identification Differential Effect Search method) and Virtual Twins. The first one is arguably the most well known local recursive methods while Virtual Twins is a recognized method, representative of global modelling approaches. In the domain of KDD-SD methods, we chose APRIORI-SD and CN2-SD which are well-known representative of respectively exhaustive and heuristic approaches to SD.

While these four methods do cover the spectrum of SD and identification methods, both SIDES and Virtual Twins are well adapted to predictive tasks, APRIORI-SD and CN2-SD can only address prognostic tasks. Since Q-Finder can address both tasks, it is compared with the two methods that are adapted to each of the two tasks described in section 3.2. For all the analyses, IDMPS wave 6 were used as the discovery dataset and IDMPS wave 7 as the test dataset. To allow comparison of results, only the top-10 subgroups of each algorithm are considered without any human intervention during the selection. Finally, default parameters of each algorithm were selected, except shared parameters which we kept as similar as possible.

3.3.1 Exploring prognostic subgroups

For each of the three approaches to identify prognostic subgroups (CN2-SD, APRIORI-SD, and Q-Finder) we detail the version and main parameters.

CN2-SD[13]**:** A beam search algorithm adapted from association rule learning CN2 to SD. It introduces a weighted covering method, where examples covered by a subgroup are not removed from the training set but their weights are decreased. This allows examples to appear in several subgroups and cover groups with more diversity. The version used is the one found in Orange 3.23.1. The default parameters are:

13. https://pypi.org/project/Orange3/

$WRAcc$ as the optimisation metric, $beam\_width = 20$ (the bigger the beam, the more combinations are tested), $max\_rule\_length = 3$ (parameter representing the maximum complexity of a subgroup[14]) and $min\_covered\_examples = 10\%$ (minimum coverage of a subgroup[15]).

APRIORI-SD[16]: An exhaustive search algorithm adapted from association rule learning APRIORI to SD. Compared to APRIORI it only considers subgroups that contain the target variable in the right-hand side. Like CN2-SD, it also uses the weighted covering method. The Python package pysubgroup version 0.6.1 (*Lemmerich* et al. 2018) is used, with the following parameters: $WRAcc$ as the optimisation metric, $maxdepth = 3$ [14] and $result\_set\_size\_coverage = 10\%$ [15].

Q-Finder prognostic mode: The version used is 5.4 with $C_{max} = 3$, $\#Bins$ = 10 and $\#Cats = \infty$ (see section 2.2). Only left-bounded and right-bounded intervals are considered. The thresholds for credibility criteria are the default values presented in section 2.3.1 : $minimum\ coverage = 10\%$, $minimum\ basic\ pattern\ absolute\ contribution = 0.2$, $maximum\ basic\ pattern\ contribution\ ratio = 5$, $minimum\ effect\ size = 1.2$ (with or without correction for confounders), and $maximum\ effect\acute{s}\ significance\ threshold = 0.05$ (with or without correction for confounders). Multiple testing correction is addressed using *Bonferroni correction* in the discovery dataset and *Benjamini-Hochberg procedure* in the test dataset. For the ranking steps, aggregation rules are the ones presented in section 2.3.2, m_c being the p-value for subgroup's effect when corrected for confounders. The default top-k selection is performed with the odds-ratio corrected for confounders as the second metric and $\delta_{ES} = 0.2$ (see section 2.4.2).

3.3.2 Exploring predictive subgroups

For each of the three approaches to identify predictive subgroups (Virtual Twins, SIDES and Q-Finder) we detail the version and main parameters.

Virtual Twins[17]: Following the vignette's recommendation from the R package aVirtualTwins version 1.0.1, missing values were *a priori* imputed on the discovery dataset using $rfImpute()$ from the randomForest package version 4.6.14. For this step and each of the following, the $seed$ was set to 42. After the imputation, Virtual Twin's first step consisted in using *randomForest()* from the randomForest package (version 4.6.14) with $ntree = 500$ and $threshold = 0.5$ (threshold above which the treatment effect is considered significant for a patient). The second step consisted in performing a classification tree with $maxdepth = 3$ (maximum depth of the classification tree[14]). Only the leaves for which the predicted outcome was the target were considered as outputted subgroups.

SIDES[18]: The version considered is 1.14 from the SIDES R package. The parameters considered are: $M = 5$ (maximum number of best promising subgroups selected at each step of the algorithm), $alpha = 0.05$ (overall type 1 error rate, which is compared with p-values corrected for multiple testing using a resampling-based method to protect the overall type 1 error rate), $S = 200$ (minimum subgroup size desired, set at 10% of the discovery dataset[15]), $L = 3$ (maximum depth of the tree[14]), $D = 0$ (minimum difference between the treatment and the control arm), $gamma = 1$ (relative improvement parameter), $num\_crit = 1$ (splitting criterion used: maximizing the differential effect between the two child subgroups), $H = 1$ (i.e. no random split of the discovery dataset), $ord.bin = 10$ (number of classes continuous covariates are discretized into[19]). As SIDES is a non-deterministic algorithm, the $seed$ was set to 42.

Q-Finder predictive mode: The version used is 5.4 with $C_{max} = 3$, $\#Bins$ = 10 and $\#Cats = \infty$ (see section 2.2). Only left-bounded and right-bounded intervals are considered. The thresholds[20] for credibility criteria are the default values presented in section 2.3.1 : $minimum\ coverage = 10\%$, $minimum\ basic\ pattern\ absolute\ contribution = (0, 0.2)$, $maximum\ basic\ pattern\ contribution\ ratio = (\infty, 5)$,

14. This corresponds to Q-Finder's maximum complexity parameter
15. This corresponds to Q-Finder's minimum threshold for the coverage criterion
16. https://github.com/flemmerich/pysubgroup
17. https://cran.r-project.org/web/packages/aVirtualTwins/vignettes/full-example.html
18. https://cran.r-project.org/web/packages/SIDES/index.html
19. This corresponds to Q-Finder's $\#Bins$ parameter
20. In predictive mode the user indicates 2 thresholds instead of 1 for some criteria, with relation to the treatment effect within the subgroup (first value) and the differential treatment effect (second value)

$minimum\ effect\ size = (1, 1.2)$ (with or without correction for confounders), and $maximum\ effect\acute{s}\ significance\ threshold = (0.05, 0.05)$ (with or without correction for confounders). Multiple testing correction is addressed using *Bonferroni correction* in the discovery dataset and *Benjamini-Hochberg procedure* in the test dataset. For the ranking steps, aggregation rules are the ones presented in section 2.3.2, m_c being the p-value for differential treatment effect when corrected for confounders. Nevertheless, they are additional intermediate ranks to account for criteria with 2 thresholds (one for treatment effect within the subgroup, the other for differential treatment effect):

- **Rank i:** threshold met for treatment effect within the subgroup only
- **Rank i+1:** threshold met for differential treatment effect only
- **Rank i+2:** threshold met for both treatment effect within the subgroup and differential treatment effect

The default top-k selection is performed with the odds-ratio for differential treatment effect corrected for confounders as the second metric and $\delta_{ES} = 0.2$ (see section 2.4.2).

3.4 Results

3.4.1 Prognostic factors identification

Q-Finder results on the prognostic task: Q-Finder generated 203 subgroups satisfying all the credibility criteria. Among the top-10 subgroups selected while accounting for diversity, 2 are of complexity 1, none are of complexity 2 and 8 are of complexity 3. The results are presented below in Table 1 along with the main metrics of interest computed on both the discovery and the test datasets (see Table S8 in supplementary materials for the additional metrics computed and outputted from Q-Finder). The two first-ranked subgroups S1 and S2 are both of complexity 1 and state that patients whose last postprandial glucose (PPG) level was below 172.0 mg/dl (resp. whose last fasting blood glucose (FBG) level was below 129.6 mg/dl) do have a better glycemic control than the others. Both subgroups are very close to the glycemic control targets established by the American Diabetes Association (resp. 180 mg/dl for PPG and 130 mg/dl for FBG (*American Diabetes Association* 2017)). The coverage (or support) of the first subgroup S1 is 30% of the discovery dataset, its adjusted odds-ratio is 4.8 ([3.5; 6.5]) and its p-value is $1.81E-23$ on the discovery dataset. All selected subgroups were successfully reapplied on the test dataset, with odds-ratios corrected for confounders above 1.81 and p-values below 0.05 when adjusted for multiple testing by Benjamini-Hochberg procedure. It is worth noticing that all the subgroups were significant using the more conservative Bonferroni correction in the discovery dataset.

Table 1: Q-Finder results on the detection of prognostic factors describing patients with better glycemic control

Subgroup Ranking*	Subgroup description	Coverage Discovery / Test	Adjusted odds-ratios (IC95%) Discovery**	p-value Discovery	Adjusted p-value Discovery***	Adjusted odds-ratios (IC95%) Test**	p-value Test	Adjusted p-value Test***
S1	Last postprandial glucose measurement (mg/dL) $\leq$ 172.0	30% / 27%	4.78 [3.5; 6.5]	1.81E-23	1.15E-18	4.28 [3.2; 5.7]	2.09E-24	1.04E-23
S2	Last fasting blood glucose measurement (mg/dL) $\leq$ 129.6	38% / 36%	3.60 [2.8; 4.7]	9.86E-21	6.28E-16	5.06 [4.0; 6.5]	9.82E-37	9.82E-36
S3	Follow healthy diet and exercise plan = Yes AND Device used for insulin: Vials and syringes = No AND Cumulated # of individual therapies taken by the patient $\leq$ 3	14% / 16%	2.57 [1.9; 3.5]	7.08E-9	4.50E-4	2.50 [1.9; 3.3]	1.78E-11	3.84E-11
S4	Follow healthy diet and exercise plan = Yes AND Device used for insulin: Vials and syringes = No AND # of different cardiovascular treatments $\leq$ 2	22% / 17%	2.26 [1.7; 3.0]	9.96E-9	6.34E-4	2.36 [1.8; 3.0]	5.24E-11	7.48E-11
S5	Follow healthy diet and exercise plan = Yes AND # of OGLD $\leq$ 1 AND Type of health insurance = Public	16% / 24%	2.47 [1.8; 3.4]	1.05E-8	6.69E-4	2.44 [1.9; 3.1]	2.20E-13	7.33E-13
S6	Follow healthy diet and exercise plan = Yes AND Covered by a health insurance = Yes AND # of different cardiovascular treatments $\leq$ 2	17% / 11%	2.34 [1.7; 3.1]	1.28E-8	8.15E-4	1.81 [1.3; 2.5]	1.42E-4	1.58E-4
S7	Follow healthy diet and exercise plan = Yes AND Covered by a health insurance = Yes AND Cumulated # of individual therapies taken by the patient $\leq$ 4	17% / 16%	2.33 [1.7; 3.1]	1.30E-8	8.27E-4	2.44 [1.9; 3.2]	1.92E-11	3.84E-11
S8	Follow healthy diet and exercise plan = Yes AND Times seen by a diabetologist in the past 3 months = 0 AND Cumulated # of individual therapies taken by the patient $\leq$ 4	16% / 17%	2.43 [1.8; 3.3]	1.85E-8	1.18E-3	2.25 [1.7; 3.0]	5.46E-9	6.82E-9
S9	Follow healthy diet and exercise plan = Yes AND Covered by a health insurance = Yes AND Treated for other form of dyslipidemia = Yes	22% / 19%	2.33 [1.7; 3.2]	1.94E-7	1.24E-2	2.64 [2.0; 3.5]	4.87E-11	7.48E-11
S10	Follow healthy diet and exercise plan = Yes AND Covered by a health insurance = Yes AND Received biguanides = No	12% / 8%	2.40 [1.7; 3.3]	2.66E-7	1.70E-2	1.86 [1.3; 2.6]	3.07E-4	3.07E-4

** Subgroup ranking is based on p-values on discovery dataset*
*** Odds-ratios are adjusted for confounding factors through multiple regression model*
**** Adjusted p-values for multiple testing are based on a Bonferroni correction (resp. Benjamini-Hochberg procedure) on the discovery dataset (resp. on the test dataset)*

Results for CN2-SD and APRIORI-SD: Results for both CN2-SD and APRIORI-SD are given below. For CN2-SD, no subgroups were outputted using the default parameters, described in 3.3.1. An analysis of the sensitivity is presented in the discussion of the results (see section 4.2). For APRIORI-SD, 186 subgroups were outputted. Among the top-10 subgroups based on the WRAcc measure, 1 is of complexity 1, 2 are of complexity 2 and 7 are of complexity 3. The complexity 1 subgroup (S4 in Table 2) is defined by a *last postprandial glucose measurement below 144 mg/dl* (WRAcc on discovery dataset: 0.0329). All complexity 2 and 3 subgroups, except S10, are also defined by this basic pattern, combined with other patterns such as $Receives\ GLP{-}1\ analogues = No$ or $Self{-}monitoring\ testing\ performed\ at\ bed\ time = No$. The results are presented below in Table 2 with the WRAcc measure, both on the discovery and the test datasets:

Table 2: APRIORI-SD results on the detection of prognostic factors describing patients with better glycemic control

Subgroup Ranking*	Subgroup description	WRAcc Discovery	WRAcc Test
S1	Last postprandial glucose measurement (mg/dL) $\leq$ 144 AND Self-monitoring testing performed at bed time = No	3.30E-2	2.52E-2
S2	Last postprandial glucose measurement (mg/dL) $\leq$ 144 AND Self-monitoring testing performed at bed time = No AND # of sorts of required hospitalization (macro/microvascular, hypo) = 0	3.30E-2	2.47E-2
S3	Last postprandial glucose measurement (mg/dL) $\leq$ 144 AND # of sorts of required hospitalization (macro/microvascular, hypo) = 0	3.29E-2	2.82E-2
S4	Last postprandial glucose measurement (mg/dL) $\leq$ 144	3.29E-2	2.91E-2
S5	Last postprandial glucose measurement (mg/dL) $\leq$ 144 AND Self-monitoring testing performed at bed time = No AND Receives GLP-1 analogues = No	3.28E-2	2.38E-2
S6	Last postprandial glucose measurement (mg/dL) $\leq$ 144 AND Receives amylin agonist = No AND Receives GLP-1 analogues = No	3.27E-2	2.77E-2
S7	Last postprandial glucose measurement (mg/dL) $\leq$ 144 AND Receives GLP-1 analogues = No AND # of sorts of required hospitalization (macro/microvascular, hypo) = 0	3.27E-2	2.68E-2
S8	Last postprandial glucose measurement (mg/dL) $\leq$ 144 AND Receives GLP-1 analogues = Yes	3.27E-2	2.77E-2
S9	Last postprandial glucose measurement (mg/dL) $\leq$ 144 AND Self-monitoring testing performed at bed time = No AND Receives amylin agonist = No	3.27E-2	2.46E-2
S10	Follow healthy diet and exercise plan = Yes AND Receives more than 2 OGLD = No AND Patient living in = Urban area	3.08E-2	3.43E-2

** Subgroup ranking is based on WRAcc measure in discovery dataset.*

3.4.2 Predictive factors identification

Q-Finder results on the predictive task: Q-Finder generated 2775 subgroups in the discovery dataset that pass all the criteria of credibility on the predictive task. Among the top-10 subgroups selected while accounting for diversity, all are of complexity 3 except one. The results are presented below in Table 3 with main criteria of interest computed on both the discovery and the test datasets (see Table S9 and S10 in supplementary materials for the additional metrics computed and outputted from Q-Finder).

Subgroup S2 states that patients who use a disposable pen, don't smoke and are not heavily treated for diabetes, have a higher risk than the others in experiencing hypoglycemia under Premixed insulin than under Basal insulin (coverage = 25%, adjusted odds-ratio for differential treatment effect = 3.31 [2.0 ; 5.6], p-value = 7.13E-6).

The seven first selected subgroups were successfully reapplied on the test dataset, with adjusted odds-ratios related to differential treatment effect above 1.86. Indeed, these subgroups have a p-value below 0.05 adjusted for multiple testing using Benjamini-Hochberg procedure, despite the fact that no subgroups were "statistically significant" after Bonferroni correction in the discovery dataset. It is worth noticing that all subgroups have adjusted odds-ratios above 1.0 in the test dataset.

Table 3: Q-Finder results on the detection of predictive factors describing patients with a higher risk than the others in experiencing hypoglycemia under Premixed insulin than under Basal insulin (with or without Prandial insulin).

Subgroup Ranking*	**Subgroup description**	**Coverage Discovery / Test**	**Adjusted odds-ratios for differential treatment effect (IC95%) Discovery****	**p-value for differential treatment effect Discovery**	**Adjusted odds-ratios for differential treatment effect (IC95%) Test****	**p-value for differential treatment effect Test**	**Adjusted p-value for differential treatment effect Test*****
S1	Statins for dyslipidemia = Yes AND Device used for insulin: Vials and syringes = No AND Total # of anti-diabetics agents $\leq$ 1	28% / 31%	3.04 [1.9; 5.0]	7.02E-6	2.12 [1.4; 3.2]	2.36E-4	1.18E-3
S2	Device used for insulin: Disposable pen = Yes AND Smoking habits = Never AND Total # of anti-diabetics agents $\leq$ 1	25% / 26%	3.31 [2.0; 5.6]	7.13E-6	1.93 [1.3; 2.9]	2.04E-3	4.28E-3
S3	Total # of anti-diabetics agents $\leq$ 1 AND # of different devices used by the patient $\geq$ 1	48% / 61%	2.71 [1.8; 4.2]	9.55E-6	2.59 [1.7; 4.0]	1.92E-5	1.92E-4
S4	Treated for other form of dyslipidemia = Yes AND Times seen by a diabetologist in the past 3 months $\leq$ 1 AND Device used for insulin: Vials and syringes = No	33% / 38%	3.55 [2.0; 6.3]	1.26E-5	1.93 [1.2; 3.0]	5.02E-3	7.17E-3
S5	Receives oral glycaemic lowering drugs = Yes AND Times seen by a diabetologist in the past 3 months = 0 AND Device used for insulin: Vials and syringes = No	29% / 34%	2.98 [1.8; 4.9]	2.40E-5	1.86 [1.2; 2.8]	2.14E-3	4.28E-3
S6	Statins for dyslipidemia = Yes AND Total # of anti-diabetics agents $\leq$ 1 AND Age at diagnosis (year) $\leq$ 56	30% / 33%	2.74 [1.7; 4.4]	2.64E-5	2.04 [1.4; 3.0]	4.08E-4	1.34E-3
S7	Treated for other form of dyslipidemia = Yes AND Times seen by a diabetologist in the past 3 months $\leq$ 1 AND # of different devices used by the patient $\geq$ 1	33% / 44%	3.37 [1.9; 6.0]	2.79E-5	2.05 [1.2; 3.4]	4.58E-3	7.17E-3
S8	Statins for dyslepidemia = Yes AND Device used for insulin: Vials and syringes = No AND HDL serum cholesterol (mg/dL) $\leq$ 58.0	27% / 30%	3.22 [1.9; 5.6]	2.82E-5	1.05 [0.7; 1.7]	8.21E-1	8.21E-1
S9	Statins for dyslipidemia = Yes AND Visits diabetes websites = No AND Duration of insulin therapy (year) $\geq$ 4	34% / 32%	2.59 [1.7; 4.1]	3.12E-5	1.14 [0.8; 1.7]	5.09E-1	5.65E-1
S10	Other form of dyslipidemia = Yes AND Visits diabetes websites = No AND Duration of insulin therapy (year) $\geq$ 4	40% / 37%	2.56 [1.6; 4.0]	3.22E-5	1.25 [0.9; 1.8]	2.48E-1	3.10E-1

** Subgroup ranking is based on p-value for differential treatment effect on discovery dataset*
*** Odds-ratios are adjusted for confounding factors through multiple regression model*
**** Adjusted p-values for multiple testing are based on a Benjamini-Hochberg procedure on the test dataset*

Results for SIDES and Virtual Twins on the predictive task: Results for both SIDES and Virtual Twins are given below. For SIDES, no subgroups were outputted using the default parameters, described in section 3.3.2. An analysis of the sensitivity is presented in the discussion of the results (see section 4.2). For Virtual Twins, only three subgroups were obtained, 1 of complexity 2 and 2 of complexity 3. The results are presented below in Table 4 with the metrics that are outputted from the algorithm, both on the discovery and the test datasets. All subgroups are defined by a same attribute, the "number of different lipid-lowering agents for dyslipidemia".

Table 4: Virtual Twins results on the detection of predictive factors describing patients with a higher risk than the others in experiencing hypoglycemia under Premixed insulin than under Basal insulin (with or without Prandial insulin).

Subgroup Ranking*	Subgroup description	Treatment event rate Discovery / Test	Control event rate Discovery / Test	Treatment sample size Discovery / Test	Control sample size Discovery / Test	Risk Ratio Discovery / Test
S1	# of OGLD ≥ 2 AND # of different lipid-lowering agents for dyslipidemia ≥ 1	33% / 34%	16% / 26%	72 / 408	340 / 755	2.06 / 1.36
S2	# of OGLD ≤ 2 AND Duration of insulin therapy (year) ≥ 3 AND # of different lipid-lowering agents for dyslipidemia ≥ 1	38% / 39%	29% / 32%	238 / 339	432 / 433	1.31 / 1.21
S3	Receives oral glycaemic lowering drugs = Yes AND Total serum triglycerides (mg/dL) ≥ 169.7 AND # of different lipid-lowering agents for dyslipidemia $= 0$	24% / 33%	23% / 27%	57 / 9	110 / 18	1.08 / 1.20

** Subgroup ranking is based on risk ratios in discovery dataset*
In the discovery dataset, risk ratios were computed after missing values imputation. In the test dataset, risk ratios were computed on the original dataset.

4 Discussion

4.1 Discussion of the results

For clarity we discuss the results in relation to Q-Finder for both the search of prognostic factors and predictive factors.

4.1.1 Q-Finder generates the top-k hypotheses

Q-Finder has proposed 10 prognostic factors and 10 predictive factors. This is more than the set of subgroups generated by Virtual Twins and conversely to SIDES and CN2-SD that did not generate any subgroups with their default parameters This illustrates that with default parameters Q-Finder systematically gives results whose credibility are assessed.

As for SIDES, the lack of results may well be explained by the strategy it uses to generate hypotheses. Indeed, SIDES filtering strategy, in which subgroups have to pass all predefined criteria in the learning phase (including the p-value corrected for multiple testing, a very conservative step), strongly limits hypotheses generation. The absence of results is therefore not uncommon with SIDES. On the contrary, the top-k selection strategy of Q-Finder favors the generation of hypotheses since the k best-ranked subgroups of the discovery dataset will be considered as hypotheses to be tested on independent data. This both allows to assess Q-Finder's results robustness, while preserving the statistical power (as only k tests are performed in the test dataset). Therefore, conversely to SIDES, the correction for multiple testing that is performed in the discovery dataset (that both gives more credibility to the results from the learning phase and increases the subgroups discrimination in the ranking phase) does not hinder the most promising subgroups to be tested and possibly validated on an independent dataset. Q-Finder is thus aligned with the notion of "statistical thoughtfulness"[21] recently promoted by the American Statistical Association (*Wasserstein* et al. 2019).

For CN2-SD, the lack of results may be due to the beam search, which does not cover the entirety of the search space and may thus miss relevant subgroups. Indeed, in a pure beam search strategy, the search for subgroups of higher complexity is based on the ones of lower complexity. This can therefore lead to missing subgroups, notably to favor the overall accuracy at the expense of local structures[22]. Equally, beam search strategies could miss subgroups with optimal thresholds, as stated in section 1.4. Indeed, the ability

21. *Wasserstein* et al. (2019) support the view that thoughtful researchers should "recognize when they are doing exploratory studies and when they are doing more rigidly pre-planned studies". They argue that "Most scientific research is exploratory in nature" and "the design, conduct, and analysis of a study are necessarily flexible, and must be open to the discovery of unexpected patterns that prompt new questions and hypotheses".
22. This topic is in particular discussed in http://www.realkd.org/subgroup-discovery/the-power-of-saying-i-dont-know-an-introduction-to-subgroup-discovery-and-local-modeling/

to perform an exhaustive search allows Q-Finder to find the optimal selector-values for each combination of attributes that meet as much as possible the set of credibility criteria (as defined in section 2.3. This point is illustrated in Table 1 with subgroups S3 and S8, where the attribute-selector pair*"Cumulated number of individual therapies taken by the patient* $\leq$*"* is associated with the value 3 or 4 depending on the context of the other basic patterns). Finally, non-exhaustive searches can also miss the detection of emerging synergistic effects, that have probably also been ruled out by SIDES, since the null (or very small) individual effects of each basic pattern would not be selected in a node of a tree. Nevertheless, one of the major advantages of beam search is related to its memory consumption. Since the algorithm stores only a limited number of basic patterns at each level of the search tree, the size of the memory in the worst case is $O(Bm)$, where B is the beam width, and m is the complexity of the subgroup. It is also faster as only the B most promising subgroups of complexity m are considered to explore the subgroups of complexity $m + 1$.

4.1.2 Credibility of the generated subgroups: Q-Finder favors the generation of credible subgroups

By searching for subgroups that meet the recommended and standard credibility criteria for clinical research, Q-Finder makes it possible to directly target promising and credible subgroups for their final clinical evaluation. More precisely, subgroups are assessed on their coverage and effect sizes adjusted for confounding factors, on their adjusted p-values for multiple testing, and the contribution of each basic pattern to the overall relationship with the outcome. Like most SI-SD algorithms for the search of predictive factors (e.g.: MOB, Interaction Trees, STIMA, ...), SIDES and Virtual Twins only cover a limited number of these credibility criteria (see Table S4 in supplementary materials). SIDES and Virtual Twins for example do not drive the subgroups generation on risk ratios corrected for known biases (i.e. the "confounding factors", which are already known as being associated with the outcome). Therefore, the results generated by SIDES and Virtual Twins are possibly biased and have thus a higher risk of being ruled out afterwards during their clinical assessment. Similarly, SIDES and Virtual Twins may have ruled out subgroups that could have held after correcting for confounding factors.

This is even more obvious for CN2-SD and APRIORI-SD, whose detection of prognostic factors are based on a main criterion: the WRAcc. This criterion represents a trade-off between coverage and effect. Although widely used in the KDD-SD community, it is neither conform with the standards in clinical research, nor corrected for confounding factors (see Table S3 in supplementary materials).

For all these algorithms, the identified subgroups may thus be ruled out during their posterior evaluation by the metrics of interest. Moreover, although the adjusted effect sizes of all APRIORI-SD subgroups appear high in both discovery and test datasets they are redundant. In fact, several subgroups sharing the same basic patterns are associated with the very same extension (as suggested by the identical results on credibility measures in the discovery dataset), which masks the fact that they are the same subgroups (e.g. S1 and S2 as well S3 and S4 in Table 5). The fact that an increase in complexity is not always accompanied by an increase in effect is due to the fact that APRIORI-SD does not include any parameters evaluating the contribution of basic patterns (such as the Q-Finder *basic patterns contribution criteria*), which leads to unnecessarily more complex subgroups. Finally, one can see that adjusted effect sizes of Virtual Twins subgroups are mostly smaller than those of the Q-Finder subgroups, and that none of the three subgroups generated by Virtual Twins have p-values below 0.05 once confusion biases have been corrected in the test dataset (Table 6). Based on the credibility criteria used in clinical research, Virtual Twins has therefore generated less convincing results than Q-Finder.

Moreover, whether Virtual Twins, APRIORI-SD or CN2-SD, the robustness of the results is not meant to be evaluated on independent data. Similarly, they do not seek to control and assess the risk of false positives, regardless of their presence in the results. This seriously undermines the credibility of the results.

Table 5: Credibility metrics from literature (used in Q-Finder) computed on APRIORI-SD results

Subgroup Ranking*	**Subgroup description**	**Coverage Discovery / Test**	**Adjusted odds-ratios (IC95%) Discovery****	**p-value Discovery**	**Adjusted odds-ratios (IC95%) Test****	**p-value Test**	**Adjusted p-value Test*****
S1	Last postprandial glucose measurement (mg/dL) $\leq$ 144 AND Self monitoring testing performed at bed time = No	15% / 12%	4.73 [3.5; 6.5]	2.17E-22	3.80 [2.8; 5.1]	6.12E-19	1.03E-18
S2	Last postprandial glucose measurement (mg/dL) $\leq$ 144 AND Self monitoring testing performed at bed time = No AND # of sorts of required hospitalization (macro/microvascular, hypo) = 0	15% / 12%	4.73 [3.5; 6.5]	2.17E-22	3.76 [2.8; 5.1]	2.39E-18	3.42E-18
S3	Last postprandial glucose measurement (mg/dL) $\leq$ 144 AND # of sorts of required hospitalization (macro/microvascular, hypo) = 0	16% / 13%	4.65 [3.4; 6.4]	4.82E-22	4.21 [3.1; 5.6]	3.97E-22	1.99E-21
S4	Last postprandial glucose measurement (mg/dL) $\leq$ 144	16% / 13%	4.65 [3.4; 6.4]	4.82E-22	4.31 [3.2; 5.8]	3.24E-23	3.24E-22
S5	Last postprandial glucose measurement (mg/dL) $\leq$ 144 AND Self testing performed at bed time = No AND Receives GLP-1 analogues = No	15% / 12%	4.81 [3.5; 6.6]	1.78E-22	3.67 [2.7; 5.0]	2.14E-17	2.37E-17
S6	Last postprandial glucose measurement (mg/dL) $\leq$ 144 AND Receives amylin agonist = No AND Receives GLP-1 analogues = No	15% / 13%	4.73 [3.5; 6.5]	4.15E-22	4.19 [3.1; 5.6]	1.14E-21	3.52E-21
S7	Last postprandial glucose measurement (mg/dL) $\leq$ 144 AND Receives GLP-1 analogues = No AND # of sorts of required hospitalization (macro/microvascular, hypo) = 0	15% / 13%	4.73 [3.5; 6.5]	3.95E-22	4.07 [3.0; 5.5]	1.67E-20	3.34E-20
S8	Follow healthy diet and exercise plan = Yes AND Times seen by a diabetologist in the past 3 months = 0 AND Cumulated # of individual therapies taken by the patient $\leq$ 4	15% / 13%	4.73 [3.5; 6.5]	3.95E-22	4.17 [3.1; 5.6]	1.41E-21	3.52E-21
S9	Last postprandial glucose measurement (mg/dL) $\leq$ 144 AND Self monitoring testing performed at bed time = No AND Receives amylin agonist = No	15% / 12%	4.7 [3.4; 6.4]	3.86E-22	3.72 [2.8; 5.0]	3.74E-18	4.68E-18
S10	Follow healthy diet and exercise plan = Yes AND Receives more than 2 OGLD = No AND Patient living in = Urban area	42% / 25%	2.6 [2.0; 3.4]	1.01E-12	2.41 [1.9; 3.0]	1.10E-14	1.10E-14

** Subgroup ranking is the same as in Table 2*
*** Odds-ratios are adjusted for confounding factors through multiple regression model*
**** Adjusted p-values for multiple testing are based on Benjamini-Hochberg procedure on the test dataset*

Table 6: Credibility metrics from literature (used in Q-Finder) computed on Virtual Twins results

Subgroup Ranking*	**Subgroup description**	**Coverage Discovery / Test**	**Adjusted odds-ratios for differential treatment effect (IC95%) Discovery****	**p-value for differential treatment effect Discovery**	**Adjusted odds-ratios for differential treatment effect (IC95%) Test****	**p-value for differential treatment effect Test**	**Adjusted p-value for differential treatment effect Test*****
S1	# of OGLD $\geq$ 2AND # of different lipid-lowering agents for dyslipidemia $\geq$ 1	21% / 21%	2.00 [1.0; 3.8]	3.27E-2	1.37 [0.8; 2.2]	2.21E-1	3.32E-1
S2	# of OGLD $\leq$ 2 AND Duration of insulin therapy (year) $\geq$ 3 AND # of different lipid-lowering agents for dyslipidemia $\geq$ 1	47% / 44%	1.77 [1.2; 2.7]	9.01E-3	1.37 [0.9; 2.0]	8.80E-2	2.64E-1
S3	Receives oral glycaemic lowering drugs = Yes AND Total serum triglycerides (mg/dL) $\geq$ 169.7 AND # of different lipid-lowering agents for dyslipidemia = 0	4% / 1%	0.53 [0.2; 1.5]	2.51E-1	0.68 [0.1; 4.5]	7.03E-1	7.03E-1

** Subgroup ranking is the same as in Table 4*
*** Odds-ratios are adjusted for confounding factors through multiple regression model*
**** Adjusted p-values for multiple testing are based on a Benjamini-Hochberg procedure on the test dataset*

4.1.3 Better supporting subgroups

Q-Finder supports the set of subgroups with standard and recommended criteria of credibility in clinical research. Therefore, all metrics used in Q-Finder for generating hypotheses are given as outputs for transparency. As strongly recommended by the American Statistical Association (*Wasserstein* et al. 2019):

- p-values are reported in continuous. This should allow experts to better interpret them, and avoid basing any decision on a p-value threshold that would misrepresent what "worthy" and "unworthy" results are[23].
- p-values can be "interpreted in lights of its context of sample size and meaningful effect size". This set of metrics is key for scientific inference of results.

Q-Finder also provides confidence intervals of effect sizes, to help experts to assess results.

This is to be contrasted with most packages, including the ones used in this paper to compare Q-Finder. Indeed, Virtual Twins package only gives information about size and risk ratios (not adjusted for confounding factors). The SIDES package would only output continuous p-values below an arbitrary threshold. As for CN2-SD and APRIORI-SD, we are far from the standards for the publication of prognostic factors (see Table S5 and S6 in supplementary materials for comparison of output metrics from standard packages). As a result, using only a subset of the recommended credibility metrics to both generate and evaluate the subgroups leads to less well-supported results and a higher risk of a posteriori discarding them.

4.1.4 Diversity: Q-Finder favors the generation of various subgroups and limits redundancy

By combining an exhaustive search and an innovative selection algorithm, Q-Finder has made it possible to promote the generation of subgroups whose descriptions differ, for both prognostic and predictive factors (see Table 7 and Table 8 that compare the subgroups diversity level between algorithms). Overall, diversity on subgroups description is less present in subgroups from Virtual Twins which only generated 3 subgroups all defined by the *"number of different lipid-lowering agents for dyslipidemia"* attribute. For APRIORI-SD, which generated a large number of subgroups, 9 out of 10 subgroups are defined by the same basic pattern (*"last postprandial glucose measurement* $\leq$ *144 mg/dl"*). In addition, note that those of the APRIORI-SD subgroups which are excessively complex (as indicated in section 4.1.2) would have been avoided by Q-Finder's top-k selection algorithm, for which an increase in complexity requires an increase in effect.

However, we observe that 8 out of the top-10 Q-Finder prognostic subgroups share a basic pattern (i.e. *"Follow healthy diet and exercise plan = Yes"*). The results could be further improved by using other types of diversity algorithms based on the subgroups' extensions (see section 6.9.6 in supplementary materials regarding redundancy). The known draw-back of searching for extensional redundancy is related to its higher computational cost. One could also note that several basic patterns although not syntactically redundant do share a similar clinical meaning (e.g. *"Statins for dyslipidemia = Yes"* and *"Treated for other form of dyslipidemia = Yes"* are both about taking a dyslipidemia treatment). This is explained by the fact that several attributes in the dataset contain related information. Stricter pre-selection of attributes, based on both correlation analysis and clinical expertise before performing the analysis, is a classic approach to reduce this type of redundancy.

23. As mentioned by the American Statistical Association (*Wasserstein* et al. 2019), arbitrary p-value thresholds could lead to biased conclusions and published results, and are only acceptable for "automated tools" and "automated decision rule". In that respect, Q-Finder does use p-value thresholds for the automatic ranking of subgroups, but no filter on p-value thresholds is done whether to select the top-k subgroups (some of them could have p-values above 0.05) or to report their results on both discovery and test datasets. "Completeness in reporting" is therefore allowed in Q-Finder by presenting the k findings obtained "without regard to statistical significance or any such criterion."

Table 7: Prognostic subgroups diversity visualization per attribute and selector-value pairs

		Attributes	Last post prendial glucose measurement (mg/dL)	Self-monitoring testing performed at bed time	Receives GLP-1 analogues	# of sorts of required hospitalization (macro/microvascular, hypo)	Receives amylin agonist	Receives more than 2 OGLD	Patient living in	Follow healthy diet and exercise plan	Covered by a health insurance	Device used for insulin: Vials and syringes	# of different cardiovascular treatments	Received biguanides	Type of health insurance	Treated for other form of dyslipidemia	Cumulated # of individual therapies taken by the patient		Times seen by a diabetologist in the past 3 months	# of OGLD	Last fasting blood glucose measurement (mg/dL)	Last postprandial glucose measurement (mg/dL)
Algorithm	**# of distinct attributes**	**Subgroups Ranking**	≤ 144	No	No	0	No	No	Urban area	Yes	Yes	No	≤ 2	No	Public	Yes	≤ 4	≤ 3	0	1	≤ 129.6	≤ 172
APRIORI-SD	8	S1	X	X																		
		S2	X	X		X																
		S3	X			X																
		S4	X																			
		S5	X	X	X																	
		S6	X		X		X															
		S7	X		X	X																
		S8	X		X																	
		S9	X	X			X															
		S10						X	X	X												
Q-Finder prognostic mode	12	S1																				X
		S2																			X	
		S3								X		X						X				
		S4								X		X	X									
		S5								X					X					X		
		S6								X	X		X									
		S7								X	X						X					
		S8								X							X		X			
		S9								X	X					X						
		S10								X	X			X								

Table 8: Predictive subgroups diversity visualization per attribute and selector-value pairs

		Attributes	# of different lipidlowering agents for dyslipidemia		Total serum triglycerides (mgdL)	Duration of insulin therapy (year)	# of OGLD		Receives oral glycaemic lowering drugs	Device used for insulin Vials and syringes	Total # of antidiabetics agents	Statins for dyslipidemia	Visits diabetes websites	Treated for other form of dyslipidemia	# of different devices used by the patient	Smocking habits	Times seen by a diabetologist in the past 3 months		Other form of dyslipidemia	HDL serum cholesterol (mgdL)	Device used for insulin Disposable pen	Age at diagnosis (year)		Duration of insulin therapy (year)
Algorithm	**# of distinct attributes**	**Subgroups Ranking**	0	≥ 1	≥ 169.7	≥ 3	≤ 2	≥ 2	Yes	No	≤ 1	Yes	No	Yes	≥ 1	Never	≤ 1	0	Yes	≤ 58.0	Yes	≤ 56	≥ 4	
Virtual Twins	5	S1		X				X																
		S2		X		X	X																	
		S3	X		X				X															
Q-Finder predictive mode	14	S1								X	X	X												
		S2									X					X					X			
		S3									X				X									
		S4								X				X			X							
		S5							X	X								X						
		S6									X	X										X		
		S7												X	X		X							
		S8								X		X								X				
		S9										X	X										X	
		S10											X						X				X	

4.2 Limits of the experiments

4.2.1 Algorithms used for benchmarking

We only considered two algorithms for the detection of prognostic factors (CN2-SD and APRIORI-SD) and two algorithms for the detection of predictive factors (Virtual Twins and SIDES) for the experiments. These algorithms have been chosen because they are representative of SI-SD and KDD-SD algorithms. Although it would be interesting to compare with other algorithms (such as MOB, STIMA, Interaction Trees, ...) to strengthen the key messages delivered in this paper, a simple review of the literature on these algorithms allows to generalize some of these messages, whether on the ability to target suited hypotheses or on the ability to support them with recommended credibility metrics.

Equally, we only used default thresholds of the algorithms, except when it was relevant in view of the comparison between algorithms (e.g. we used the same coverage value of 10%). One can argue that other thresholds could have been tested to improve the algorithms' outputs. However, the goal is not here to prove the deficiencies of other algorithms through these experiments, but to generate elements of discussion that shows Q-Finder specificity and the source of its power (e.g. regarding the optimized metrics, the outputs metrics, etc). Nevertheless, a limited analysis of parameters sensitivity was performed for both SIDES and CN2-SD which did not generate any subgroups with the default set of parameters. For SIDES, we explored an increase of the threshold of significance to 0.2 and a decreased maximum number of best promising subgroups selected at each step to 1. Only the first case produced a single candidate subgroup (see below, p-value = 0.066 corrected for multiple testing using a resampling-based method to address the overall type 1 error rate). This subgroup is "close" to some of the top 10 predictive subgroups in Q-Finder, which supports the results obtained with Q-Finder:

```
Receives oral glycaemic lowering drugs = Yes AND
Treated for other form of dyslipidemia = Yes AND
          Visits diabetes websites = No
```

For CN2-SD, we explored a slight increase of the beam parameter to 50 and a decreased coverage parameter to 5%. No results were obtained in the first case, and the second case did generate 2 subgroups of complexity 3 that share two attributes (see Table S7 in supplementary materials for the results). More generally, sensitivity analyses are recommended in any SD tasks, by marginally modifying algorithms parameters or the outcome definition (e.g.: $HbA1c < 7.5\%$ instead of 7%).

4.2.2 Limits of the IDMPS databases: surveys

As the IDMPS databases are derived from surveys spread over time, they each reflect an image at a given time. As a result, treatment initiation may have occurred before data recording. In this situation, the data studied are not necessarily the baseline of the study, which gives the results a purely descriptive character. Indeed, a variable by which a subgroup is defined should not be affected by treatment response (*Dijkman* et al. 2009). The most common use case of SD is rather the retrospective study of prospective data (e.g. RCT) or real world data, in which temporal information is collected, in order to only consider the information before treatment's intake (i.e. the "baseline" period).

4.3 Generalization to other pathologies or research questions

Q-Finder was applied in the field of diabetes to many other research questions, such as the detection of patient profiles that benefit the most of SGLT2i compared to DDP4i in terms of renal function preservation, using Electronic Health Record data (*Zhou* et al. 2019; *Zhou* et al. 2018); the identification of profiles of patients who better control their blood sugar, using data from pooled observational studies (*Rollot* et al. 2018, "Reali project"); and the discovery of new predictors of diabetic ketoacidosis (DKA), a serious complication of type 1 diabetes, using data from a national diabetes registry (*Ibald-Mulli* et al. 2019). Q-Finder was also successfully applied in the context of several other pathologies such as hypophosphatasia, using SNPs data (*Mornet* et al. 2020), dry eye disease using prospective clinical trials data (*Amrane* et al. 2015), and cancer using clinical data from RCTs (*Alves* et al. 2020; *Dumontet* et al. 2018; *Dumontet* et al. 2016; *Nabholtz* et al. 2012) or transcriptomic data from a research cohort (*Adam* et al. 2016).

The Q-Finder approach is indeed generic by design and can be applied to any pathology and research questions, as can many SD algorithms, provided that the data can be represented in tabular form. In each case, the aggregation rules and metrics of interest are defined according to each research question to align with the needs and generate relevant and useful hypotheses. In this respect, the Q-Finder's methodology can be adapted to more complex situations, where the nal assessment by clinicians must also rely on clinical metrics. For example, in the case of the search for treatment responders subgroups, the search may be motivated by other criteria such as "not being associated with a specific adverse effect", or "having an equally good treatment effect regardless of patient age". Regarding the experiment that was done in this article, one could have searched for subgroups predictive of low rate of hypoglycemia (outcome) while focusing on subgroups of patients with strong glycaemic control, to identify subgroups of interest associated with both higher treatment efcacy and better safety than average.

Q-Finder can easily be adapted to any other research questions, including non-clinical ones, as its parameters can be set to directly target subgroups of interest in relation to any types of objective. Extracting the best hypotheses possible from a dataset, based on multiple criteria, using both statistical and business metrics is a common need in many sectors. For example, in the banking and insurance sectors, a common need is to identify the subgroups of customers most likely to churn (outcome) with a specific focus on those associated with high levels of prot (business metrics).

5 Conclusion

Subgroup Discovery has become an important task in the field of Subgroup Analysis. Q-Finder inherits both SI-SD and KDD-SD culture, borrowing metrics and evaluation from the first one and hypothesis generation from the second. As such, Q-Finder is a SD algorithm dedicated to the identification of either prognostic or predictive factors in clinical research. The generated subgroups are driven on a set of recommended

criteria in clinical studies to directly target promising and credible subgroups for their final clinical evaluation. This contrasts with most standard algorithms that rely only partially on these credibility metrics, and for which the risk of being ruled out afterwards by a clinical assessment is greater. Q-Finder also favors the hypothesis generation thanks to (1) an exhaustive dataset exploration that allows for emerging synergistic effects, optimally-defined subgroups and new insights to come out, and (2) its top-k selection strategy that selects credible and diverse subgroups to be tested on independent datasets. The latter step both allows the assessment of subgroups robustness while preserving the statistical power by testing a limited number of highly credible subgroups. Final results are then assessed by providing (1) a list of standard credibility metrics for both experts' adherence and publication purposes, as well as (2) the criteria used during the exploration for the full transparency of the algorithm.

In many aspects, Q-Finder thus tends to comply with the recent recommendations of the American Statistical Association (*Wasserstein* et al. 2019) that amongst others encourage hypothesis generation in exploratory studies, the prior definition of meaningful effect sizes, reporting continuous p-values in their context of sample size and effect size. They also insist that researchers should be open "to the role of Expert judgement" and involve them at every stage of the inquiry. As a matter of fact, beyond its fully automatic mode, the Q-Finder approach also supports selecting subgroups based on clinical expertise to both increase subgroups relevance to the research question and reduce false positives.

Applied on the IDMPS database to benchmark it against state-of-the-art algorithms, Q-Finder results were best in jointly satisfying the empirical credibility of subgroups (e.g., higher effect sizes adjusted for confounders and lower p-values adjusted for multiple testing), and their diversity. These subgroups are also those that are supported by the largest number of credibility measures. Q-Finder has already proved its value on real-life use cases by successfully addressing high-stake research questions in relation to a specific pathology and/or drug such as efficacy and safety questions and by dealing with main limits of standard algorithms (e.g. the lack of results or the low subgroups credibility). Its high comprehensibility did favor the acceptance by clinical teams of the identified subgroups. Finally, Q-Finder could straightforwardly be extended to other research questions (including non-clinical ones), notably by tailoring the metrics used in the exploration to directly target the subgroups of interest in relation to the objective.

Acknowledgement

We deeply thanks Sanofi medical, Jean-Marc Chantelot and the IDMPS Steering Committee for their medical expertise, financial support and proofreading. We also express thanks to Martin Montmerle, Mélissa Rollot, Margot Blanchon and Alexandre Civet for their remarks and invaluable feedbacks. Finally we thank the whole Quinten team for their dedication to the Q-Finder development during the past twelve years.

Author contributions

CE, MG, AT and JZ conceived the idea for this paper. CE, MG, MQ and JZ implemented the analysis. CE, MG and JZ wrote sections of the manuscript. All authors contributed to manuscript revision, read and approved the submitted version. CE, MG and JZ equally contributed to this work.

Competing interests

The authors declare the following competing interests. Employment: CE, MG, MQ, and AT are employed by Quinten. Financial support: The development of Q-Finder has been fully funded by Quinten. This article has been funded by Quinten with the help of Sanofi who provided the dataset, and contributed to the revisions.

Data availability statement

Qualified researchers may request access to patient level data and related study documents including the clinical study report, study protocol with any amendments, blank case report form, statistical analysis plan,

and dataset specifications. Patient level data will be anonymized and study documents will be redacted to protect the privacy of trial participants. Further details on Sanofi's data sharing criteria, eligible studies, and process for requesting access can be found at: https://www.clinicalstudydatarequest.com

References

Adam, J., *Sourisseau*, T., *Olaussen*, K. A., *Robin*, A., *Zhu*, C. Q., *Templier*, A., *Civet*, A., *Girard*, P., *Lazar*, V., *Validire*, P., *Tsao*, M. S., *Soria*, J.-C., and *Besse*, B. "MMS19 as a potential predictive marker of adjuvant chemotherapy benefit in resected non-small cell lung cancer". In: *Cancer Biomarkers* 17.3 (Sept. 26, 2016), pp. 323–333. ISSN: 15740153, 18758592. DOI: 10.3233/CBM-160644.

Adolfsson, J. and *Steineck*, G. "Prognostic and treatment-predictive factors-is there a difference?" In: *Prostate cancer and prostatic diseases* 3.4 (2000), pp. 265–268. DOI: 10.1038/sj.pcan.4500490.

Alomar, M. J., *Al-Ansari*, K. R., and *Hassan*, N. A. "Comparison of awareness of diabetes mellitus type II with treatment's outcome in term of direct cost in a hospital in Saudi Arabia". In: *World Journal of Diabetes* 10.8 (Aug. 2019), pp. 463–472. DOI: 10.4239/wjd.v10.i8.463.

Alves, A., *Civet*, A., *Laurent*, A., *Parc*, Y., *Penna*, C., *Msika*, S., *Hirsch*, M., *Pocard*, M., and *COINCIDE*, G. "Social deprivation aggravates post-operative morbidity in carcinologic colorectal surgery: Results of the COINCIDE multicenter study". In: *Journal of Visceral Surgery* (July 2020), pp. 1–9.

American Diabetes Association. "6. Glycemic Targets". In: *Diabetes Care* 40 (Supplement 1 Jan. 2017), S48–S56. ISSN: 0149-5992, 1935-5548. DOI: 10.2337/dc17-S009.

Amrane, M., *Civet*, A., *Templier*, A., *Kang*, D., and *Figueiredo*, F. C. "Patients with Moderate to Severe Dry Eye Disease in Routine Clinical Practice in the UK - Physician and Patient's Assessments". In: *Investigative Ophthalmology Visual Science* 56.7 (2015), pp. 4443–4443.

Aschner, P., *Gagliardino*, J. J., *Ilkova*, H., *Lavalle*, F., *Ramachandran*, A., *Mbanya*, J. C., *Shestakova*, M., *Chantelot*, J.-M., and *Chan*, J. C. N. "Persistent poor glycaemic control in individuals with type 2 diabetes in developing countries: 12 years of real-world evidence of the International Diabetes Management Practices Study (IDMPS)". In: *Diabetologia* 63.4 (2020), pp. 711–721. DOI: 10.1007/s00125-019-05078-3.

Atzmueller, M. "Subgroup discovery". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 5.1 (2015), pp. 35–49. DOI: doi:10.1002/widm.1144.

Ballarini, N. M., *Rosenkranz*, G. K., *Jaki*, T., *König*, F., and *Posch*, M. "Subgroup identification in clinical trials via the predicted individual treatment effect". In: *PLoS ONE* 13.10 (2018), e0205971–22. DOI: doi:10.1371/journal.pone.0205971.

Battioui, C., *Shen*, L., and *Ruberg*, S. "A resampling-based ensemble tree method to identify patient subgroups with enhanced treatment effect". In: *Proceedings of the Joint Statistical Meetings*. 2014.

Betensky, R. A. "The *p* -Value Requires Context, Not a Threshold". In: *The American Statistician* 73 (sup1 Mar. 29, 2019), pp. 115–117. ISSN: 0003-1305, 1537-2731. DOI: 10.1080/00031305.2018.1529624.

Blume, J. D., *D'Agostino McGowan*, L., *Dupont*, W. D., and *Greevy*, R. A. "Second-generation p-values: Improved rigor, reproducibility, & transparency in statistical analyses". In: *PLOS ONE* 13.3 (Mar. 22, 2018). Ed. by N. R. *Smalheiser*, e0188299. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0188299.

Burke, J. F., *Sussman*, J. B., *Kent*, D. M., and *Hayward*, R. A. "Three simple rules to ensure reasonably credible subgroup analyses." In: *BMJ* 351 (2015), h5651. DOI: doi:10.1136/bmj.h5651.

Candler, T. P., *Mahmoud*, O., *Lynn*, R. M., *Majbar*, A. A., *Barrett*, T. G., and *Shield*, J. P. H. "Treatment adherence and BMI reduction are key predictors of HbA1c 1 year after diagnosis of childhood type 2 diabetes in the United Kingdom". In: *Pediatric Diabetes* 19.8 (Dec. 2018), pp. 1393–1399. ISSN: 1399-543X, 1399-5448. DOI: 10.1111/pedi.12761.

Chen, L., *Magliano*, D. J., and *Zimmet*, P. Z. "The worldwide epidemiology of type 2 diabetes mellitus—present and future perspectives". In: *Nature Reviews Endocrinology* 8.4 (Apr. 2012), pp. 228–236. ISSN: 1759-5029, 1759-5037. DOI: 10.1038/nrendo.2011.183.

Chen, S., *Tian*, L., *Cai*, T., and *Yu*, M. "A general statistical framework for subgroup identification and comparative treatment scoring". In: *Biometrics* 73.4 (2017), pp. 1199–1209. DOI: 10.1111/biom.12676.

Dijkman, B., *Kooistra*, B., *Bhandari*, M., and *Evidence-Based Surgery Working Group*. "How to work with a subgroup analysis." In: *Canadian journal of surgery. Journal canadien de chirurgie* 52.6 (Dec. 2009), pp. 515–522.

Dimitrienko, A. and *Lipkovitch*, I. "SLIDES: Exploratory subgroup analysis: Post-hoc subgroup identification in clinical trials". In: (2014), pp. 1–28.

Doove, L. L., *Dusseldorp*, E., *Van Deun*, K., and *Van Mechelen*, I. "A comparison of five recursive partitioning methods to find person subgroups involved in meaningful treatment–subgroup interactions". In: *Advances in Data Analysis and Classification* 8.4 (2013), pp. 403–425.

Dumontet, C., *Hulin*, C., *Dimopoulos*, M., *Belch*, A., *Dispenzieri*, A., *Ludwig*, H., *Rodon*, P., *Van Droogenbroeck*, J., *Qiu*, L., *Cavo*, M., *Van de Velde*, A., *Lahuerta*, J., *Allangba*, O., *Lee*, J., *Boyle*, E., *Perrot*, A., *Moreau*, P., *Manier*, S., *Attal*, M., *Roussel*, M., *Mohty*, M., *Mary*, J., *Civet*, A., *Costa*, B., *Tinel*, A., *Gaston-Mathé*, Y., and *Facon*, T. "Develpment of a predictive model to identify patients with multiple myeloma not eligible for autologous transplant at risk for severe infections using data from the first trial". In: *Haematologica* 101 (2016).

Dusseldorp, E., *Conversano*, C., and *Van Os*, B. J. "Combining an Additive and Tree-Based Regression Model Simultaneously: STIMA". In: *Journal of Computational and Graphical Statistics* 19.3 (Jan. 2010), pp. 514–530. ISSN: 1061-8600, 1537-2715. DOI: 10.1198/jcgs.2010.06089.

Dusseldorp, E., *Doove*, L., and *Mechelen*, I. van. "Quint: An R package for the identification of subgroups of clients who differ in which treatment alternative is best for them". In: *Behavior Research Methods* (May 2016), pp. 1–14. DOI: 10.3758/s13428-015-0594-z.

Eveno, C., *Parc*, Y., *Laurent*, A., *Tresallet*, C., *Vaillant*, J.-C., *Civet*, A., *Ducreux*, M., and *Emile*, J.-F. "An abnormal body mass index of is associated with an increased risk of rectosigmoid cancer risk: interest a short recto-sigmoidoscopy for early detection." In: Vienna, Austria: United European Gastroenterology Journal, 2014.

Foster, J. C., *Taylor*, J. M. G., and *Ruberg*, S. J. "Subgroup identification from randomized clinical trial data". In: *Statistics in Medicine* 30.24 (Aug. 2011), pp. 2867–2880.

Ganascia, J.-G. "TDIS - an Algebraic Formalization." In: *IJCAI* (1993).

Gaston-Mathe, Y., *Fan*, T., *Shaunik*, A., *Brulle-Wohlhueter*, C., and *A*, C. "Using machine learning algorithms to identify predictive factors of clinical outcomes with iGlarLixi or iGlar in the LixiLan-L trial". In: *Diabetologia* 60.1 (2017), pp. 1–608. DOI: 10.1007/s00125-017-4350-z.

Hahsler, M., *Chelluboina*, S., *Hornik*, K., and *Buchta*, C. "The arules R-Package Ecosystem: Analyzing Interesting Patterns from Large Transaction Data Sets". In: *Journal of Machine Learning Research* 12 (June 2011), pp. 2021–2025.

Hapfelmeier, A., *Kurt*, U., and *Haller*, B. "Subgroup identification by recursive segmentation". In: *Journal of Applied Statistics* 0.0 (Mar. 2018), pp. 1–24. DOI: 10.1080/02664763.2018.1444152.

Herrera, F., *Carmona*, C. J., *González*, P., and *Jesus*, M. J. del. "An overview on subgroup discovery: foundations and applications". In: *Knowledge and Information Systems* 29.3 (2010), pp. 495–525.

Holzinger, A. "Interactive machine learning for health informatics: when do we need the human-in-the-loop?" In: *Brain Informatics* 3.2 (Feb. 2016), pp. 119–131. DOI: 10.1007/s40708-016-0042-6.

Huling, J. D. and *Yu*, M. "Subgroup Identification Using the personalized Package". In: *arXiv.org* (2018). DOI: arXiv:1809.07905.

Ibald-Mulli, A., *Margot*, B., *Alexandre*, C., *Julia*, H., *Simon*, G., *Oussama*, B., and *Dieter*, P. *Identification of predictive factors of DKA using a subgroup discovery algorithm*. EASD. Sept. 18, 2019.

Imai, K. and *Ratkovic*, M. "Estimating treatment effect heterogeneity in randomized program evaluation". In: *The Annals of Applied Statistics* 7.1 (Mar. 2013), pp. 443–470. ISSN: 1932-6157. DOI: 10.1214/12-AOAS593.

Juarez, D. T., *Sentell*, T., *Tokumaru*, S., *Goo*, R., *Davis*, J. W., and *Mau*, M. M. "Factors Associated With Poor Glycemic Control or Wide Glycemic Variability Among Diabetes Patients in Hawaii, 2006–2009".

In: *Preventing Chronic Disease* 9 (Sept. 27, 2012), p. 120065. ISSN: 1545-1151. DOI: 10.5888/pcd9.120065.

Kavsek, B. and *Lavrač*, N. "APRIORI-SD: Adapting Association Rule Learning to Subgroup Discovery". In: *Applied Artificial Intelligence* 20.7 (2007), pp. 543–583. DOI: 10.1080/08839510600779688.

Korepanova, N. V. "Subgroup Discovery for Treatment Optimization". In: (2018), pp. 1–6.

Lavrač, N., *Kavsek*, B., *Flach*, P. A., and *Todorovski*, L. "Subgroup Discovery with CN2-SD." In: *Journal of Machine Learning Research* (2004). DOI: 1005332.1005338.

Leeuwen, M. van and *Knobbe*, A. "Diverse subgroup set discovery". In: *Data Mining and Knowledge Discovery* 25.2 (2012), pp. 208–242. DOI: 10.1007/s10618-012-0273-y.

Lemmerich, F. and *Becker*, M. "pysubgroup: Easy-to-use subgroup discovery in python". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. 2018, pp. 658–662.

Lipkovich, I. et al. "Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials". In: *Statistics in Medicine* 36.1 (2016), pp. 136–196. DOI: 10.1002/sim.7064.

Lipkovich, I. and *Dmitrienko*, A. "Strategies for Identifying Predictive Biomarkers and Subgroups with Enhanced Treatment Effect in Clinical Trials Using SIDES". In: *Journal of Biopharmaceutical Statistics* 24.1 (2014), pp. 130–153. DOI: 10.1080/10543406.2013.856024.

Lipkovich, I., *Dmitrienko*, A., *Muysers*, C., and *Ratitch*, B. "Multiplicity issues in exploratory subgroup analysis". In: *Journal of Biopharmaceutical Statistics* 28.1 (2018), pp. 63–81. DOI: 10.1080/10543406.2017.1397009.

Lipkovich, I., *Dmitrienko*, A., *Patra*, K., *Ratitch*, B., and *Pulkstenis*, E. "Subgroup Identification in Clinical Trials by Stochastic SIDEScreen Methods". In: *Statistics in Biopharmaceutical Research* 9.4 (2017), pp. 368–378. DOI: https://doi.org/10.1080/19466315.2017.1371069.

Loh, W.-Y., *Cao*, L., and *Zhou*, P. "Subgroup identification for precision medicine: A comparative review of 13 methods". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 9.5 (2019), pp. 604–21. DOI: 10.1002/widm.1326.

Ma, Q., *Liu*, H., *Xiang*, G., *Shan*, W., and *Xing*, W. "Association between glycated hemoglobin A1c levels with age and gender in Chinese adults with no prior diagnosis of diabetes mellitus". In: *Biomedical Reports* 4.6 (2016), pp. 737–740. ISSN: 2049-9434, 2049-9442. DOI: 10.3892/br.2016.643.

Mornet, E., *Mélissa*, R., *Cyrine*, H., *Christelle*, D., *Agnès*, T., *Alexandre*, T., *Brigitte*, S.-B., and *Alexandre*, C. "Recherche de SNP modulateurs du phénotype hypophosphatasique par un algorithme d'identification de règles d'association (subgroup discovery)". In: Tours, France: Assises de Génétique Humaine et Médicale, 2020.

Nabholtz, J.-M., *Dauplat*, M.-M., *Abrial*, C., *Weber*, B., *Mouret-Reynier*, M.-A., *Gligorov*, J., *Tredan*, O., *Vanlemmens*, L., *Petit*, T., *Guiu*, S., *Jouannaud*, C., *Tubiana-Mathieu*, N., *Kwiatkowski*, F., *Cayre*, A., *Uhrhammer*, N., *Privat*, M., *Desrichard*, A., *Chollet*, P., *Chalabi*, N., and *Penault-Llorca*, F. "Abstract P3-06-20: Is it possible to predict the efficacy of a combination of Panitumumab plus FEC 100 followed by docetaxel (T) for patients with triple negative breast cancer (TNBC)? Final biomarker results from a phase II neoadjuvant trial." In: *Cancer Research* 72.24 Supplement (2012), P3-06-20–P3-06-20. DOI: 10.1158/0008-5472.SABCS12-P3-06-20. eprint: https://cancerres.aacrjournals.org/content/72/24_Supplement/P3-06-20.

Ogurtsova, K., *Rocha Fernandes*, J. D. da, *HUANG*, Y., *Linnenkamp*, U., *Guariguata*, L., *Cho*, N. H., *Cavan*, D., *Shaw*, J. E., and *Makaroff*, L. E. "IDF Diabetes Atlas: Global estimates for the prevalence of diabetes for 2015 and 2040." In: *Diabetes research and clinical practice* 128 (June 2017), pp. 40–50. DOI: 10.1016/j.diabres.2017.03.024.

Oxman, A. D. and *Guyatt*, G. H. "A consumer's guide to subgroup analyses." In: *Annals of Internal Medicine* 116.1 (Jan. 1992), pp. 78–84. DOI: 10.7326/0003-4819-116-1-78.

Petrovski, G., *Gjergji*, D., *Grbic*, A., *Vukovic*, B., *Krajnc*, M., and *Grulovic*, N. "Switching From Pre-mixed Insulin to Regimens with Insulin Glargine in Type 2 Diabetes: A Prospective, Observational Study of Data From Adriatic Countries". In: *Diabetes Therapy* 9.4 (Aug. 2018), pp. 1657–1668. ISSN: 1869-6953, 1869-6961. DOI: 10.1007/s13300-018-0467-4.

Polonik, W. and *Wang*, Z. "PRIM analysis". In: *Journal of Multivariate Analysis* 101.3 (Mar. 2010), pp. 525–540. ISSN: 0047259X. DOI: 10.1016/j.jmva.2009.08.010.

Rollot, M., *Bonnemaire*, M., *Brulle-Wohlhueter*, C., *Pedrazzini*, L., *Boelle-Le Corfec*, E., *Bigot*, G., *Didac*, M., *Bonadonna*, R., *Gourdy*, P., *Müller-Wieland*, D., *Hacman*, O., *Chiorean*, A., and *Freemantle*, N. "A machine learning algorithm can identify clusters of patients with favourable glycaemic outcomes in a pooled European Gla-300 studies (REALI): Novel signposts for clinicians?" In: *Diabetologia : journal of the European Association for the Study of Diabetes (EASD)* 61.Supplement 1 (Oct. 1, 2018), p. 876. ISSN: 0012-186X. URL: https://publications.rwth-aachen.de/record/740010 (visited on 12/05/2019).

Saturni, S., *Bellini*, F., *Braido*, F., *Paggiaro*, P., *Sanduzzi*, A., *Scichilone*, N., *Santus*, P. A., *Morandi*, L., and *Papi*, A. "Randomized controlled trials and real life studies. Approaches and methodologies: a clinical point of view." In: *Pulmonary Pharmacology amp; Therapeutics* 27.2 (2014), pp. 129–138. DOI: 10.1016/j.pupt.2014.01.005.

Schnell, P. M., *Tang*, Q., *Offen*, W. W., and *Carlin*, B. P. "A Bayesian credible subgroups approach to identifying patient subgroups with positive treatment effects". In: *Biometrics* 72.4 (2016), pp. 1026–1036. DOI: 10.1111/biom.12522.

Shaw, J., *Sicree*, R., and *Zimmet*, P. "Global estimates of the prevalence of diabetes for 2010 and 2030". In: *Diabetes Research and Clinical Practice* 87.1 (Jan. 2010), pp. 4–14. ISSN: 01688227. DOI: :10.1016/j.diabres.2009.10.007.

Su, X., *Tsai*, C.-L., *Wang*, H., *Nickerson*, D. M., and *Li*, B. "Subgroup Analysis via Recursive Partitioning". In: *SSRN Electronic Journal* (2009). ISSN: 1556-5068. DOI: 10.2139/ssrn.1341380.

Sun, X., *Briel*, M., and *Jason*, B. "Credibility of claims of subgroup effects in randomised controlled trials: systematic review". In: *BMJ* 344.mar15 1 (2012), e1553–e1553. DOI: 10.1016/j.spinee.2012.07.029.

Sun, X., *Ioannidis*, J., *Agoritsas*, T., *Alba*, A., and *Guyatt*, G. "How to Use a Subgroup Analysis". In: *JAMA* 311.4 (2014), pp. 405–7. DOI: https://doi.org/10.1001/jama.2013.285063.

Susan, L. D., *Kristina*, S. B., and *Nicole*, R. Y. "The Impact of Body Weight on Patient Utilities with or without Type 2 Diabetes: A Review of the Medical Literature". In: *Value in Health* 11.3 (Oct. 2010), pp. 478–486. DOI: https://doi.org/10.1111/j.1524-4733.2007.00260.x.

Tanniou, J., *Tweel*, I. van der, *Teerenstra*, S., and *Roes*, K. C. B. "Subgroup analyses in confirmatory clinical trials: time to be specific about their purposes". In: *BMC Medical Research Methodology* (2016), pp. 1–15. DOI: 10.1186/s12874-016-0122-6.

Tshiananga, J. K. T., *Kocher*, S., *Weber*, C., *Erny-Albrecht*, K., *Berndt*, K., and *Neeser*, K. "The Effect of Nurse-led Diabetes Self-management Education on Glycosylated Hemoglobin and Cardiovascular Risk Factors: A Meta-analysis". In: *The Diabetes Educator* 38.1 (Jan. 2012), pp. 108–123. ISSN: 0145-7217, 1554-6063. DOI: 10.1177/0145721711423978.

Wasserstein, R. L., *Schirm*, A. L., and *Lazar*, N. A. "Moving to a World Beyond "$p < 0.05$"". In: *The American Statistician* 73.sup1 (2019), pp. 1–19. DOI: 10.1080/00031305.2019.1583913. eprint: https://doi.org/10.1080/00031305.2019.1583913.

Wijngaarden, R. P. T. van, *Overbeek*, J. A., *Heintjes*, E. M., *Schubert*, A., *Diels*, J., *Straatman*, H., *Steyerberg*, E. W., and *Herings*, R. M. C. "Relation Between Different Measures of Glycemic Exposure and Microvascular and Macrovascular Complications in Patients with Type 2 Diabetes Mellitus: An Observational Cohort Study". In: *Diabetes Therapy* 8.5 (2017), pp. 1097–1109.

Xiong, H., *Brodie*, M., and *Ma*, S. "TOP-COP - Mining TOP-K Strongly Correlated Pairs in Large Databases." In: *ICDM* (2006), pp. 1162–1166.

Xu, Y., *Yu*, M., *Zhao*, Y.-Q., *Li*, Q., *Wang*, S., and *Shao*, J. "Regularized outcome weighted subgroup identification for differential treatment effects". In: *Biometrics* 71.3 (2015), pp. 645–653. DOI: 10.1111/biom.12322.

Zeileis, A., *Hothorn*, T., and *Hornik*, K. "Model-Based Recursive Partitioning". In: *Journal of Computational and Graphical Statistics* 17.2 (2008), pp. 492–514. DOI: 10.1198/106186008X319331.

Zhang, Z., *Seibold*, H., *Vettore*, M. V., *Song*, W.-J., and *François*, V. "Subgroup identification in clinical trials: an overview of available methods and their implementations with R". In: *Annals of Translational Medicine* 6.7 (2018), pp. 122–122. DOI: 10.21037/atm.2018.03.07.

Zhou, F. L., *Watada*, H., *Tajima*, Y., *Berthelot*, M., *Kang*, D., *Esnault*, C., *Shuto*, Y., *Maegawa*, H., and *Koya*, D. "Identification of subgroups of patients with type 2 diabetes with differences in renal function preservation, comparing patients receiving sodium-glucose co-transporter-2 inhibitors with those receiving dipeptidyl peptidase-4 inhibitors, using a supervised machine-learning algorithm (PROFILE study): A retrospective analysis of a Japanese commercial medical database". In: *Diabetes Obes Metab* 21.8 (2019), pp. 1925–1934. DOI: 10.1111/dom.13753.

— "PDB16 - Compare renal functional preservation outcome of SGLT2 inhibitor in patients with type 2 diabetes: a retrospective cohort study of japanese commercial database with advanced analytics approach". In: *Value in Health* 21 (2018), S121. DOI: 10.1016/j.jval.2018.09.722.

6 Supplementary Material

6.1 Beam search strategy using decision tree versus exhaustive algorithm

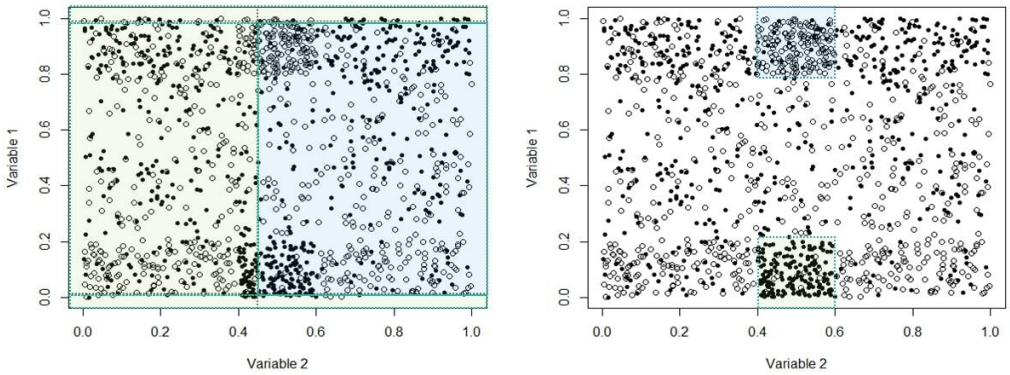

Figure S1: beam search strategy using decision tree versus SD algorithm

A well-chosen example with a categorical target variable (700 empty circles = "No" and 700 filled circles = "Yes") and two numeric description variables. On the left-side, colored areas show decision surface of a three level deep decision tree (green areas related to the "Yes" target, blue area to the "No" target). 4 subgroups are identified:

- Variable 1 $<$ 0.015 (1st split): 76% of Yes, representing 3% of the population
- Variable 1 $\geq$ 0.99 (2nd split): 69% of Yes, representing 2% of the population
- Variable 1 $\geq$ 0.015 and $<$ 0.99 & Variable 2 $<$ 0.45 (3rd split): 52% of Yes, representing 41% of the population
- Variable 1 $\geq$ 0.015 and $<$ 0.99 & Variable 2 $\geq$ 0.45 (3rd split) $\geq$ 54% of No, representing 54% of the population

The 2 last subgroups are of low accuracy in comparison to the target distribution (50%/50% of Yes/No), while the 2 firsts are of low density (less than or equal to 3%). The decision tree did not manage in finding the two subgroups we can easily see with our bare eyes on the right-side (one in the lower center and one in the upper center of the data space), defined as:

- Variable 1 $\geq$ 0.8 & variable 2 $\geq$ 0.4 and $\leq$ 0.6: 91% of No, representing 13% of the population (164 "No" versus 17 "Yes")
- Variable 1 $\leq$ 0.2 & variable 2 $\geq$ 0.4 and $\leq$ 0.6: 92% of Yes, representing 13% of the population (14 "No" versus 166 "Yes").

Both subgroups have higher accuracies than any subgroup from the decision tree. Driven both by a recursive partitioning process and by the interest of overall performance, the decision tree did not capture these regions. For more details, Mario Boley[24] further explores this topic.

24. http://www.realkd.org/subgroup-discovery/the-power-of-saying-i-dont-know-an-introduction-to-subgroup-discovery-and-local-modeling/

6.2 Figures related to Odds-Ratio

Table S1: Odds-Ratios formula for the search of prognostic factors

	Patients in the target	Patients not in the target
Patients in the subgroup's extension	a	b
Patients not in the subgroup's extension	c	d

Given a, b, c, d the number of patients corresponding to each condition:

$$\text{OR subgroup effect: } \frac{(a \times d)}{(b \times c)}$$

Table S2: Odds-Ratios formula for the search of predictive factors

Patients within the subgroup	Patients in the target	Patients not in the target
Patients treated by the treatment	a	b
Patients treated by the comparator	c	d
Patients outside the subgroup	**Patients in the target**	**Patients not in the target**
Patients treated by the treatment	a'	b'
Patients treated by the comparator	c'	d'

Given a, b, c, d, a', b', c', d' the number of patients corresponding to each condition:

$$\text{OR treatment effect in subgroup: } \frac{(a \times d)}{(b \times c)}$$

$$\text{OR differential treatment effect: } \frac{(a \times d)}{(b \times c)} \div \frac{(a' \times d')}{(b' \times c')}$$

6.3 Tables related to the metrics optimized by each algorithm for generating subgroups

Table S3: Metrics optimized by each algorithm for generating prognostic subgroups

	Search strategy	**Size**	**Basic patterns contribution**	**Effect size / Quality measure**	**Inference**	**Adjustment for confounding factors**
Q-Finder	Exhaustive	Coverage	Absolute contribution for each basic pattern & contributions ratio	Risk-Ratio	p-value corrected for multiple testing	Yes
Apriori-SD	Exhaustive	Coverage	-	WRAcc	-	No
CN2-SD	Beam	Coverage	-	WRAcc	-	No

Table S4: Metrics optimized by each algorithm for generating predictive subgroups

	Search strategy	**Size**	**Basic patterns contribution**	**Effect size / Quality measure**	**Inference**	**Adjustment for confounding factors**
Q-Finder	Exhaustive	Coverage	Absolute contribution for each basic pattern & contributions ratio	Risk-Ratio of both treatment effect within the subgroup and/or differential treatment effect	p-value corrected for multiple testing	Yes
Virtual Twins	Beam	-	-	Difference between response of active treatment compared to control treatment	-	No
SIDES	Beam	Count	The relative improvement parameter	Minimum difference between the treatment and the control	p-value corrected for multiple testing	No

6.4 Tables related to packages' output metrics

Table S5: Output metrics from Q-Finder, Apriori-SD and CN2-SD to support prognostic subgroups

	Subgroup Description	**Size**	**Basic patterns contribution**	**Effect size / Quality measure**	**Inference**	**Same in independant dataset for robustness assessment**
Q-Finder	Yes	Coverage	Absolute contribution for each basic pattern & contributions ratio	Risk-Ratio adjusted and not adjusted for confounding factors	Raw p-values and p-values corrected for multiple testing	Yes
Apriori-SD	Yes	Count	-	WRAcc	-	No
CN2-SD	Yes	Count	-	-	-	No

Table S6: Output metrics from Q-Finder, Virtual Twins and SIDES to support predictive subgroups

	Subgroup Description	**Size**	**Basic patterns contribution**	**Effect size / Quality measure**	**Inference**	**Same in independant dataset for robustness assessment**
Q-Finder	Yes	Coverage	Absolute contribution for each basic pattern & contributions ratio	Risk-Ratio adjusted and not adjusted for confounding factors (for both treatment effect and differential treatment effect)	Raw p-values and p-values corrected for multiple testing	Yes
Virtual Twins	Yes	Count	-	Risk-Ratio, treatment event rate, control event rate	-	No
SIDES	Yes	-	-	-	p-values corrected for multiple testing	Yes

6.5 Table related to CN2-SD sensitivity analysis

Table S7: CN2-SD results by decreasing the coverage parameter to 5%

Subgroup Ranking*	**Subgroup description**	**WRAcc Discovery**	**WRAcc Test**
S1	Times seen by Health Care Provider in the past 3 months $\geq$ 1 AND Last postprandial glucose measurement (mg/dL) $\leq$ 188.0 AND Receives more than 2 OGLD = No	3.87E-2	3.42E-2
S2	Times seen by Health Care Provider in the past 3 months $\geq$ 1 AND Last fasting blood glucose measurement (mg/dL) $\leq$ 144.0 AND Last postprandial glucose measurement (mg/dL) $\leq$ 140.4	2.73E-2	2.31E-2

** Subgroup ranking is based on WRAcc measure in discovery dataset.*

6.6 Aggregation rules visualization

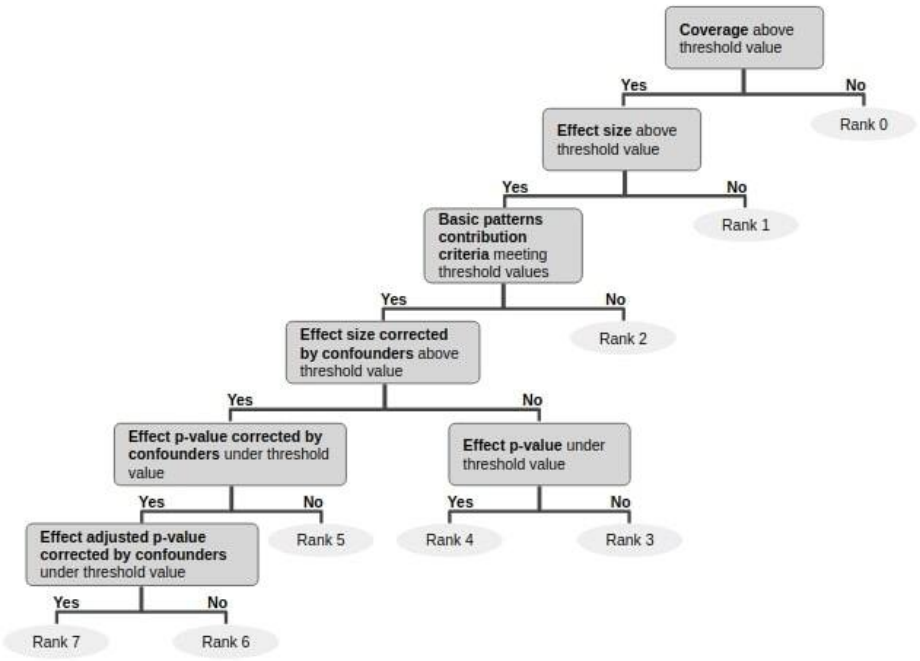

Figure S2: Aggregation rules represented by a decision tree

This figure represents the default aggregation rules associated with the search for prognostic factors through a decision tree. For the search of predictive factors, intermediate ranks should be added to both distinguish the *treatment effect within the subgroup* and the *differential treatment effect*, such as:

- **Rank i:** threshold met for treatment effect only
- **Rank i+1:** threshold met for differential treatment effect only
- **Rank i+2:** threshold met for both treatment effect and differential treatment effect

6.7 Additional metrics of Q-Finder's results in prognostic factors identification

Table S8: Q-Finder's additional output metrics in prognostic factors identification

Subgroup Ranking	Subgroup description	Basic pattern absolute contribution Discovery / Test	Basic pattern contributions ratio* Discovery / Test	Odds-ratios (IC95%) Discovery**	p-value Discovery	Odds-ratios (IC95%) Test**	p-value Test
S1	Last postprandial glucose measurement (mg/dL) $\leq$ 172.0			4.44 [3.3; 6.0]	1.89E-23	4.12 [3.2; 5.4]	2.09E-25
S2	Last fasting blood glucose measurement (mg/dL) $\leq$ 129.6			3.66 [2.8; 4.8]	1.95E-22	5.17 [4.1; 6.6]	4.62E-40
S3	Follow healthy diet and exercise plan = Yes AND Device used for insulin: Vials and syringes = No AND Cumulated # of individual therapies taken by the patient $\leq$ 3	0.78 / 0.91 0.51 / -0.28 0.54 / -0.09	1.53 / NEG	2.45 [1.8; 3.3]	2.16E-9	2.40 [1.9; 3.1]	1.25E-11
S4	Follow healthy diet and exercise plan = Yes AND Device used for insulin: Vials and syringes = No AND # of different cardiovascular treatments $\leq$ 2	0.78 /1.38 0.21 / 0.03 0.29 / -0.12	3.67 / NEG	2.24 [1.7; 2.9]	2.55E-9	2.52 [2.0; 3.2]	1.28E-13
S5	Follow healthy diet and exercise plan = Yes AND # of OGLD $\leq$ 1 AND Type of health insurance = Public	0.57 / 1.38 0.29 / -0.05 0.25 / -0.01	2.29 / NEG	2.17 [1.6; 2.9]	1.55E-7	2.63 [2.1; 3.4]	2.48E-14
S6	Follow healthy diet and exercise plan = Yes AND Covered by a health insurance = Yes AND # of different cardiovascular treatments $\leq$ 2	0.89 / 1.78 0.32 / -0.08 0.21 / 0.12	4.34 / NEG	2.34 [1.8; 3.1]	2.85E-9	2.76 [2.1; 3.6]	4.55E-13
S7	Follow healthy diet and exercise plan = Yes AND Covered by a health insurance = Yes AND Cumulated # of individual therapies taken by the patient $\leq$ 4	0.72 / 0.90 0.35 / -0.09 0.22 / -0.5	3.34 / NEG	2.36 [1.8; 3.1]	2.06E-9	2.13 [1.5; 2.9]	4.72E-6
S8	Follow healthy diet and exercise plan = Yes AND Times seen by a diabetologist in the past 3 months = 0 AND Cumulated # of individual therapies taken by the patient $\leq$ 4	0.87 /1.33 0.56 / 0.01 0.21 / -0.46	4.24 / NEG	2.56 [1.9; 3.4]	1.92E-10	2.50 [2.0; 3.1]	8.04E-16
S9	Follow healthy diet and exercise plan = Yes AND Covered by a health insurance = Yes AND Treated for other form of dyslipidemia = Yes	1.14 / 1.32 0.52 / 0.19 0.33 / -0.35	3.36 / NEG	2.48 [1.8; 3.4]	7.92E-9	2.62 [2.0; 3.4]	3.21E-14
S10	Follow healthy diet and exercise plan = Yes AND Covered by a health insurance = Yes AND Received biguanides = No	0.81 / 1.00 0.78 / 0.08 0.35 / -0.72	2.34 / NEG	2.48 [1.8; 3.4]	2.05E-8	1.92 [1.4; 2.6]	1.31E-5

** "NEG" are for negative contributions ratio values due to negative basic pattern absolute contribution.*

*** Odds-Ratios are not adjusted for confounding factors*

6.8 Additional output metrics of Q-Finder's results in predictive factors identification

Table S9: Q-Finder's additional output metrics in predictive factors identification (part 1)

Subgroup Ranking	**Subgroup description**	**Basic pattern absolute contribution to treatment effect within the subgroup (TEWTS) Discovery / Test**	**Basic pattern contribution ratio to TEWTS* Discovery / Test**	**Basic pattern absolute contribution to differential treatment effect (DTE) Discovery / Test**	**Basic pattern contribution ratio to DTE* Discovery / Test**
S1	Statins for dyslipidemia = Yes AND Device used for insulin: Vials and syringes = No AND Total # of anti-diabetics agents $\leq$ 1	0.68 / 0.15 0.53 / -0.16 0.96 / 0.09	1.80 / NEG	0.37 / 0.17 0.73 / -0.31 1.37 / 0.05	2.37 / NEG
S2	Device used for insulin: Disposable pen = Yes AND Smoking habits = Never AND Total # of anti-diabetics agents $\leq$ 1	0.90 / 0.25 0.54 / 0.02 1.00 / 0.65	1.82 / 40.18	0.97 / 0.14 0.73 / -0.14 1.37 / 0.71	1.89 / NEG
S3	Total # of anti-diabetics agents $\leq$ 1 AND # of different devices used by the patient $\geq$ 1	0.71 / 0.69 0.46 / 0.11	1.53 / 6.16	0.60 / 1.92 1.00 / 0.68	1.68 / 2.85
S4	Treated for other form of dyslipidemia = Yes AND Times seen by a diabetologist in the past 3 months $\leq$ 1 AND Device used for insulin: Vials and syringes = No	0.70 / 0.21 0.53 / 0.19 0.67 / 0.14	1.30 / 1.47	0.91 / 0.27 0.49 / 0.06 0.89 / 0.21	1.85 / 3.63
S5	Receives oral glycaemic lowering drugs = Yes AND Times seen by a diabetologist in the past 3 months = 0 AND Device used for insulin: Vials and syringes = No	0.59 / 0.27 0.81 / 0.32 0.73 / 0.33	1.39 / 1.21	0.56 / 0.31 0.55 / 0.00 0.80 / 0.41	1.45 / NEG
S6	Statins for dyslipidemia = Yes AND Total # of anti-diabetics agents $\leq$ 1 AND Age at diagnosis (year) $\leq$ 56	0.80 / 0.00 0.84 / 0.06 0.25 / 0.06	3.36 / NEG	0.60 / -0.39 0.86 / -0.26 0.22 / -0.38	3.95 / NEG
S7	Treated for other form of dyslipidemia = Yes AND Times seen by a diabetologist in the past 3 months $\leq$ 1 AND # of different devices used by the patient $\geq$ 1	0.61 / 0.20 0.48 / 0.19 0.54 / 0.03	1.27 / 6.40	0.71 / 0.23 0.35 / 0.18 0.64 / 0.18	2.01 / 1.34
S8	Statins for dyslepidemia = Yes AND Device used for insulin: Vials and syringes = No AND HDL serum cholesterol (mg/dL) $\leq$ 58.00	0.71 / 0.26 0.53 / 0.16 0.80 / 0.75	1.49 / 4.56	0.53 / 0.06 0.69 / -0.03 1.08 / 0.87	2.05 / NEG
S9	Statins for dyslipidemia = Yes AND Visits diabetes websites = No AND Duration of insulin therapy (year) $\geq$ 4	0.72 / 0.27 0.37 / 0.67 0.52 / -0.02	1.90 / NEG	0.64 / 0.03 0.39 / 0.91 0.56 / -0.13	1.66 / NEG
S10	Other form of dyslipidemia = Yes AND Visits diabetes websites = No AND Duration of insulin therapy (year) $\geq$ 4	0.55 / 0.19 0.39 / -0.14 0.54 / 0.16	1.40 / NEG	0.61 / 0.27 0.48 / -0.35 0.69 / 0.17	1.42 / NEG

** "NEG" are for negative contributions ratio values due to negative basic pattern absolute contribution.*

Table S10: Q-Finder's additional output metrics in predictive factors identification (part 2)

Subgroup Ranking	Adjusted odds ratios treatment effect within the subgroup (TEWTS) (IC 95%)* Discovery	P-value	Odds ratios TEWTS (IC 95%)** Discovery	P-value	Odds ratios for differential treatment effect (DTE) (IC 95%)** Discovery	P-value	Adjusted odds ratios TEWTS (IC 95%)* Test	P-value	Odds ratios TEWTS (IC 95%)** Test	P-value	Odds ratios for DTE (IC 95%)** Test	P-value
S1	3.04 [2.0; 4.6]	1.13E-7	3.06 [2.1; 4.6]	3.93E-8	2.92 [1.8; 4.7]	7.13E-6	1.4 [1.0; 1.9]	3.60E-2	1.41 [1.0; 1.9]	2.83E-2	1.17 [0.8; 1.7]	4.20E-1
S2	3.13 [2.0; 4.9]	3.18E-7	2.98 [2.0; 4.5]	3.13E-7	3.06 [1.9; 5.0]	1.12E-5	2.08 [1.5; 3.0]	7.16E-5	2.14 [1.5; 3.1]	2.36E-5	2.01 [1.3; 3.0]	7.84E-4
S3	2.3 [1.7; 3.1]	1.35E-7	2.36 [1.7; 3.2]	2.00E-8	2.61 [1.7; 4.0]	8.95E-6	1.94 [1.5; 2.5]	4.49E-8	2.0 [1.6; 2.5]	3.46E-9	2.84 [1.9; 4.4]	1.26E-6
S4	2.72 [1.8; 4.0]	8.34E-7	2.62 [1.8; 3.8]	5.40E-7	2.97 [1.7; 5.1]	8.20E-5	1.7 [1.3; 2.3]	4.33E-4	1.77 [1.3; 2.4]	1.06E-4	1.9 [1.2; 3.0]	4.53E-3
S5	2.79 [1.8; 4.3]	2.66E-6	2.79 [1.8; 4.2]	1.21E-6	2.75 [1.7; 4.5]	4.72E-5	1.87 [1.4; 2.6]	1.30E-4	1.87 [1.4; 2.6]	9.46E-5	1.84 [1.3; 2.7]	1.94E-3
S6	2.71 [1.8; 4.0]	3.60E-7	2.78 [1.9; 4.1]	7.96E-8	2.79 [1.8; 4.4]	1.04E-5	1.62 [1.2; 2.2]	3.82E-3	1.61 [1.2; 2.2]	3.43E-3	1.08 [0.7; 1.7]	7.30E-1
S7	2.61 [1.8; 3.9]	1.76E-6	2.49 [1.7; 3.6]	1.75E-6	2.71 [1.6; 4.7]	3.03E-4	1.66 [1.3; 2.2]	2.19E-4	1.66 [1.3; 2.1]	1.43E-4	1.88 [1.2; 3.1]	1.08E-2
S8	3.02 [2.0; 4.6]	3.58E-7	2.9 [1.9; 4.4]	3.59E-7	3.16 [1.9; 5.4]	2.03E-5	2.17 [1.6; 3.0]	4.03E-6	2.29 [1.7; 3.2]	3.82E-7	2.33 [1.6; 3.4]	1.68E-5
S9	2.47 [1.7; 3.5]	6.75E-7	2.5 [1.8; 3.5]	2.14E-7	2.51 [1.6; 3.9]	3.46E-5	2.03 [1.5; 2.8]	1.12E-5	2.11 [1.6; 2.9]	2.17E-6	2.23 [1.5; 3.3]	3.93E-5
S10	2.28 [1.6; 3.2]	8.88E-7	2.33 [1.7; 3.2]	2.29E-7	2.47 [1.6; 3.8]	3.58E-5	1.45 [1.1; 1.9]	1.38E-2	1.45 [1.1; 1.9]	1.15E-2	1.27 [0.9; 1.8]	2.12E-1

** Odds-ratios are adjusted for confounding factors through multiple regression model*
*** Odds-Ratios are not adjusted for confounding factors*

6.9 In-depth discussion on Q-Finder

6.9.1 Discovery and test datasets

It is strongly recommended to use Q-Finder with two independent datasets: a discovery dataset and a test dataset. Moreover, as it is often the case in statistical learning, a third dataset can be used as a validation dataset in order to apply on the test dataset the subgroups that did pass all criteria in the validation dataset. This allows to reinforce the confidence in the results, while assessing robustness of metrics on the test dataset. Similarly, it may be relevant to consider several test datasets for robustness assessment.

In the case where only one dataset is available, it is common to randomly split the original dataset in a discovery and a test dataset. Attention must be paid to the proportions between the two datasets, in order to maintain a sufficient number of patients in the test dataset. The decrease in the number of patients by splitting the whole dataset must be compensated by the considerable reduction in the number of tests performed in the test dataset (i.e. k tests) to either preserve or gain statistical power. Similarly, it is worth noting that in such situation, the two datasets are not perfectly independent. Therefore, robustness assessment of risk-ratios as well as adjusted p-values computations have to be interpreted more cautiously. Finally, if the dataset is too small to be split, Q-Finder can still work. However, as a general rule it is highly recommended to *a priori* select as few features and as little discretization bins as possible to limit the number of tests. In any case, whenever there is no reapplication on independent dataset, final results have to be interpreted with caution, even if they are ranked as the most credible ones. In such a situation, we recommend bootstrapping to assess the robustness of credibility metrics for each selected subgroup in the discovery dataset.

6.9.2 Management of missing values and outliers

Dealing with missing values is a critical problem in data analysis and is beyond the scope of the Q-Finder algorithm. Missing value imputation strategies are highly dependent on the underlying mechanism of missing values (be it MCAR, MAR or MNAR (*Acock* 2005)) and depend on each project and/or each variable (e.g. strategies based on clinical knowledge, Bayesian approaches, multiple imputation, ...). For all these reasons, we would recommend to let the user manage the missing data upstream, and downstream (e.g. through a sensitivity analysis), of Q-Finder. Nevertheless, if the user decides to keep missing data, the Q-Finder can still work (a contrario of many algorithms), by considering a patient in the subgroup (respectively outside) if all the basic patterns of a subgroup are satisfied and not missing (respectively if at least 1 basic pattern is not satisfied and not missing). Regarding outliers, the Q-Finder algorithm is based on statistical methods that are widely used and discussed in the literature, namely credibility criteria such as odds-ratio, regression models and p-values. Thus, the sensitivity of Q-Finder to outliers is directly related to the sensitivity of these methods, particularly for linear and logistic regression models. We recommend managing outliers upstream of the use of Q-Finder, in order to distinguish the data preparation phase itself (before Q-Finder) from the subgroup research phase (Q-Finder).

6.9.3 Variables discretization and grouping

Discretization of continuous variables is performed in Q-Finder to both reduce the number of tests and the risk of finding variable cuts that overfit the data. The default discretization method is based on an equal-frequency quantization procedure, allowing the generation of groups of similar sizes. However, other approaches could be used to better reflect variables magnitudes such as using equal-width methods (*Garcia* et al. 2013).

6.9.4 Credibility metrics

Q-Finder promotes the identification of both credible prognostic or predictive factors, by directly targeting and assessing subgroups on recommended credibility criteria (in bold and italics hereafter), as described by *Sun, Briel, Walter* et al. (2010) and *Dijkman* et al. (2009). Indeed, effects are both adjusted for confounders to check for ***comparability of known risk factors*** and are assessed using ***relative risks reduction*** which in most situations remains constant across varying baseline risks. ***Clinical importance*** of subgroups effects

or of treatment-subgroup interaction effects can also be checked and promoted by including clinical experts directly into the selection step of the top-k credible subgroups to be tested on independent dataset (see section 6.9.6). This last step also allows to both ***limit the number of tests*** and check for ***subgroups consistency across datasets***. Tests are also ***adjusted for multiplicity***, and ***interaction tests*** are used for assessing between-subgroup treatment effect interactions. Q-Finder also considers additional metrics such as filtering on a minimal subgroup's size and checking for true contributions of subgroups patterns on the overall effect. The latter could be reinforced by assessing the synergistical level of each subgroup to target the ones for which a combination of basic patterns is associated with a true gain in effect (i.e. an effect that is higher than an additive and/or multiplicative effect, or that does not arise from an interaction of basic patterns at a lower complexity).

Other credibility metrics are currently being implemented in order to further improve subgroups credibility as recommended in *Sun, Briel, Walter* et al. (2010) and *Dijkman* et al. (2009), such as assessing the treatment-subgroup interaction ***consistency across closely related outcomes*** within the study and better assessing both the ***comparability*** of prognostic or predictive factors and the significant subgroup effect ***independence*** with the other discovered subgroups. Similarly, some of the existing measures in Q-Finder could be made more powerful, such as the default corrections for multiple tests, i.e. the Benjamini-Hochberg procedure in the test dataset, and the Bonferroni correction in the discovery dataset (which makes the calculation easier given the massive number of tests performed). Indeed, both are too conservative because they do not take into account the correlations between the tests. This is all the more the case for the Bonferroni correction, which protects against type 1 error. However, this over-conservative character is attenuated in the case of the Benjamini-Hochberg procedure, which seems quite robust and remains valid for a wide variety of common dependency structures (*Goeman* et al. 2014; *Benjamini* et al. 2001). It is worth noting that these procedures do not currently hinder the hypothesis generation in Q-Finder (see section 4.1.1 in main text).

6.9.5 Aggregation rules

Aggregation rules (see section 2.3.2 in main text) are rules used to rank subgroups within groups of equal level of credibility or interest for the user. By defining such rules, users can define what the most interesting subgroups in a given specific context are and target them directly. In Q-Finder, the default aggregation rules are defined for most SD tasks in clinical research, and can be modified according to the users needs. One example of modification would be to not require the top-ranked subgroups to meet both the **effect size criterion** and the **effect size criterion corrected for confounders** as defined by default. Indeed, the latter is an unbiased effect size estimate that makes the former unnecessarily from a statistical point of view. However, such a modification would be at the expense of the computing time as the **effect size criterion** is faster to compute than the one corrected for confounders. It thus avoids to compute the latter if the threshold of the former is not met.

More generally, aggregation rules and metrics of interest should be refined according to each research question to better meet the needs and generate both relevant and useful hypotheses (see discussion in section 4.3. in main text).

6.9.6 The top-k "clinician-augmented" selection

After the ranking of subgroups, the top-k selection consists of selecting the most promising subgroups to be tested on an independent dataset. The presented top-k algorithm includes a "diversity" parameter in order to favor the generation of subgroups defined by diverse attributes. Other strategies could be considered, like selecting the most credible subgroups (i.e. regarding the subgroup's ranking) while maximizing the coverage of dataset (i.e. obtaining a set of subgroups that covers the maximum of patients) or even maximizing the coverage of targeted patients (i.e. obtaining a set of subgroups that explains as much as possible the phenomenon of interest). This can be generalized to any clinical strategy, like selecting the top-k subgroups that covers as much as possible any patients characteristic (e.g. obtaining a set of subgroups that covers all countries within a study, or all subtypes of a given disease, ...).

A drawback in promoting diversity is that we go deeper through the subgroups ranking to select various subgroups, which therefore favors the selection of generally lowest quality results, that is with lowest sizes or/and risks-ratios, and consequently with a higher risk that both the overall subgroup's effect and the basic patterns contribution are weaker and random. For example, one can notice that several Q-Finder prognostic subgroups replicated in the test dataset have non robust basic patterns (i.e. with negative absolute contribution values, see Table S8). As such, Q-Finder allows the fine-assessment of the robustness for each subgroup's basic pattern, and let the possibility to retrospectively simplify final subgroups by removing the non robust patterns that unnecessary impair subgroups.

Another caveat of such procedures that promote diversity is the risk in ruling out a very interesting subgroup (from a clinical expert's point of view) because of its redundancy with a better ranked subgroup. Therefore, in practice and in contrast to fully automated algorithms, Q-Finder supports including clinical expertise directly into the hypothesis generation process, in order to further increase the chances of generating subgroups that are not only statistically but also clinically credible. As mentioned in *Rueping* 2009, "the interestingness of a subgroup to a user is not directly dependent on its statistical significance". Therefore, integrating experts into the subgroups selection step can significantly increase the quality of the subgroups, whether to strengthen confidence in already known hypotheses or to generate innovative ones. Similarly, selecting subgroups effect or treatment-subgroup interactions that are clinically important increase subgroups credibility as stated by *Dijkman* et al. (2009). In addition, clinicians may rule out hypotheses that are clinically absurd as false positives from the discovery dataset, increasing thus the chances of selecting true ones. All choices made by clinical experts must be noted as an integral part of the hypothesis generation process.

6.9.7 Set and select Q-Finder parameters

Q-Finder proposes to define the exploration strategy(ies) upstream of obtaining the results and thus to set the hyperparameters on the basis of what the user "wishes" to obtain first. Nevertheless, the user can always restart the Q-Finder on the basis of the results, by defining a more conservative exploration (e.g. if too many good results were obtained) or more permissive (e.g. if too few results were obtained). Thus, depending on the level of the signal in the database, the user can vary his level of requirement and the level of credibility of the results.

Special attention must be paid to the hyperparameter C_{max}, whose value has a significant impact on the algorithm calculation times. For example, one may want to increase the level of complexity to obtain subgroups with larger effect sizes. However, this is at the expense of the size of the subgroups (smaller groups), as well as the simplicity of interpretation of the results by the physicians (more basic patterns). In addition, this is accompanied by a drastic increase in the number of subgroups to be tested and thus the risk of finding false positives: it then becomes more difficult to obtain low p-values adjusted for multiple testing. For all these reasons, we recommend in practice not to go beyond $C_{max} = 3$, unless the research question explicitly requires looking for groups of high complexity.

6.9.8 Management of voluminous data and calculation times

The Q-Finder algorithm that was used in the experiments is implemented in both a parallelized and optimized manner so that time computation is reduced. Overall, time computation strongly relies on machine capacities (number of CPUs, RAM capacity, ...) and code optimization. As an illustration, the identification of prognostic factors for glycaemic control (Experiment 1), from the exploration phase to the phase of re-application on test data, took 3 hours considering only 1 CPU. With 8 CPUs, this time was reduced to 30 minutes.

Computation times and data volume are generally not an issue in the clinical field composed by cohorts of a few thousand patients. Nevertheless, if the user is confronted with voluminous data (e.g. several tens of thousands) and calculation times are unreasonable, we can recommend certain options, such as making the thresholds of the credibility measures more conservative and/or slightly modifying the analytical pipeline. For example, increasing the minimal coverage or effect size thresholds will reduce the number of subgroups in the remainder of the pipeline. Moreover, removing or performing the adjustment step on confounding factors after (and not before) the selection of top-k subgroups is a way to strongly reduce computation times. The

confounding bias correction step is indeed the most time-consuming. Applying the algorithm on a random sample of the database or constraining the exploration so that it is not exhaustive are two other possible approaches.

6.9.9 General comprehensibility of the approach

The confidence in the results is based on the trust one has of the algorithm that has generated them. We think that the comprehensibility of the approach proposed by Q-Finder makes it possible to bring this level of confidence. Indeed, Q-Finder mimics the human process of hypothesis generation, where the physician generates hypotheses that are then tested on a dataset by computing the presented credibility metrics. With Q-Finder, this generation is driven by the discovery dataset and controlled by the test dataset. In a recent paper, *Murdoch* et al. (2019) introduced the important concept of *descriptive accuracy* in Data Analysis as "the degree to which an interpretation method objectively captures the relationships learned by machine-learning models". The algorithms to generate subgroups (see Algo. 1) and ranking them (see Algo. 2) are directly interpretable which give Q-Finder a high descriptive accuracy.

In our opinion, this is less true for algorithms such as SIDES or Virtual Twins, which rely on massive trees generation (a multi-nodes tree for SIDES and Random Forest for Virtual Twins) and although statistically sound may require more cognitive load to be understood by the end-user.

6.9.10 In a nutshell, why Q-Finder is an algorithm for credible SD?

Both in the title of this article and in the main text, we argue that the Q-Finder algorithm allows the generation of credible subgroups. By way of summary, we group below all the arguments that support this assertion. Q-Finder's subgroups are:

- the result of an exploration driven by a large set of credibility criteria recommended in the literature, and therefore satisfying many criteria,
- well supported by credibility metrics, which promotes their evaluation and acceptance by medical experts, while reducing the risk of being discarded *a posteriori*,
- the result of exhaustive research and not of a partial exploration of the research space, which would miss attributes-selector-value triplets and hinder the detection of emerging synergistic phenomen,
- directly defined by the optimal attribute-selector-value triplets that maximize the set of credibility criteria,
- derived from an analysis where the meaningful effect size for the research question is defined at the outset of the analysis, not after observing the results,
- both subject to an assessment of the diversity and contribution of individual effects, to avoid the risk of duplication of results or unnecessarily more complex subgroups,
- selected by medical experts (when available) prior to testing on independent data, thus supporting the selection of subgroups that are credible and relevant to the research question,
- tested on independent data, which allows both limiting the number of tests and assessing the robustness of credibility metrics,
- the result of an exploratory analysis fully assumed and therefore realized in conscience, where the level of credibility of the results must be assessed a posteriori by medical experts and not by an arbitrary p-value threshold falsely informative of what is worthy or unworthy,
- derived from an algorithm that is both interpretable by non-experts and transparent on all metrics calculated and provided as outputs.

References

Acock, A. C. "Working With Missing Values". In: *Journal of Marriage and Family* 67.4 (Nov. 2005), pp. 1012–1028. DOI: 10.1111/j.1741-3737.2005.00191.x.

Benjamini, Y. and *Yekutieli*, D. "The control of the false discovery rate in multiple testing under dependency". In: *The annals of statistics* 29.4 (2001), pp. 1165–1188.

Garcia, S., *Luengo*, J., *Sáez*, J. A., *López*, V., and *Herrera*, F. "A Survey of Discretization Techniques: Taxonomy and Empirical Analysis in Supervised Learning". In: *IEEE transactions on knowledge and data engineering* 25.4 (2013), pp. 734–750. DOI: 10.1109/TKDE.2012.35.

Goeman, J. J. and *Solari*, A. "Multiple hypothesis testing in genomics". In: *Statistics in Medicine* 33.11 (2014), pp. 1946–1978. DOI: 10.1002/sim.6082.

Murdoch, W. J., *Singh*, C., *Kumbier*, K., *Abbasi-Asl*, R., and *Yu*, B. "Definitions, methods, and applications in interpretable machine learning." In: *PNAS* 116.44 (Oct. 2019), pp. 22071–22080. DOI: 10.1073/pnas.1900654116.

Rueping, S. "Ranking interesting subgroups". In: *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*. the 26th Annual International Conference. Montreal, Quebec, Canada: ACM Press, 2009, pp. 1–8. ISBN: 978-1-60558-516-1. DOI: 10.1145/1553374.1553491.