frontiers

Supplementary Material

6 Beam search strategy using decision tree versus exhaustive algorithm

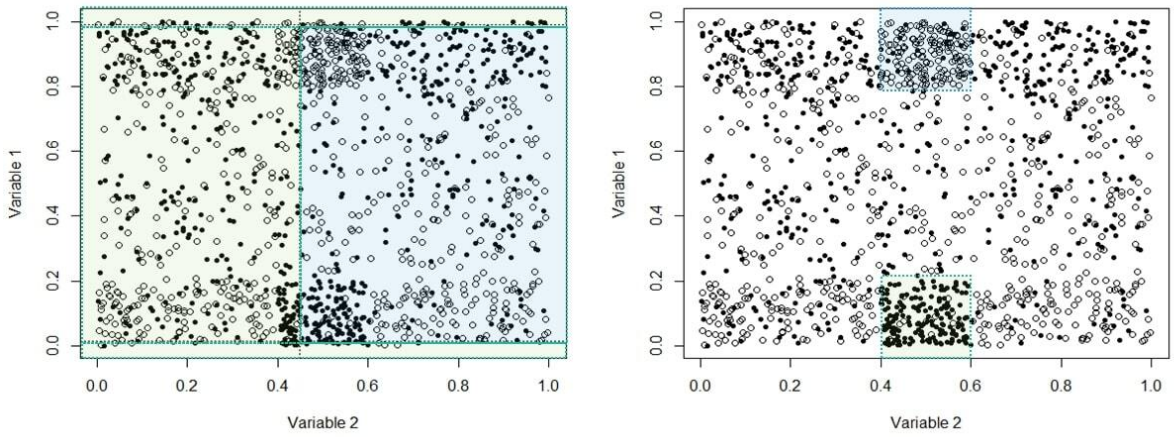

Figure S7: beam search strategy using decision tree versus SD algorithm

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 ≤ 0.2 & variable 2 ≥ 0.4 and ≤ 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[1] further explores this topic.

7 Figures related to Odds-Ratio

Table S7. 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 S8. 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:

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

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

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

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

Table S9. 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 S10. 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

9 Tables related to packages' output metrics

Table S11. 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 S12. 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

10 Table related to CN2-SD sensitivity analysis

Table S13. 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.*

11 Aggregation rules visualization

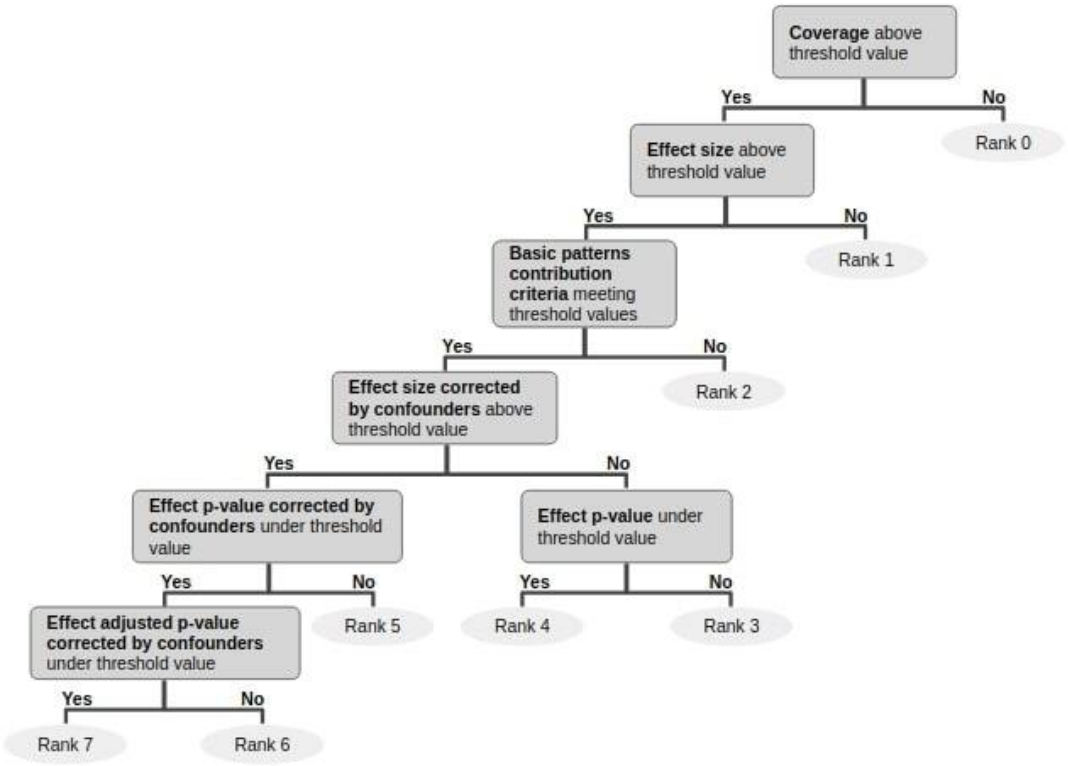

Figure S8: Aggregation rules represented by a decision tree

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

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

Table S14. 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*

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

Table S15. 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 S16. 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*

14 In-depth discussion on Q-Finder

14.1 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.

14.2 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).

14.3 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).

14.4 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 14.6). 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).

14.5 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).

14.6 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 S14). 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.

14.7 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.

14.8 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.

14.9 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.

14.10 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, Alan C "Working With Missing Values". In: *Journal of Marriage and Family* 67.4 (Nov. 2005) pp. 1012–1028.

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

Garcia, Salvador et al. "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.

Goeman, Jelle J and Aldo *Solari*. "Multiple hypothesis testing in genomics". In: *Statistics in Medicine* 33.11 (2014) pp. 1946–1978.

Murdoch, W James et al. "Definitions, methods, and applications in interpretable machine learning." In: *PNAS* 116.44 (Oct. 2019) pp. 22071–22080.

Rueping, Stefan. "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. URL: http://portal.acm.org/citation.cfm?doid=1553374.1553491 (visited on 02/25/2020)