Full Title: AI based on evolutionary computation yields algorithmic biomarker summary of a randomized rheumatoid arthritis clinical trial, accurately predicting individual patient outcomes, enabling precision medicine.

Short Title: Algorithmic clinical trial summary predicting individual patient outcomes.

Kevin Horgan[1]*, Michael F. McDermott[2], Douglas Harrington[3], Vahan Simonyan[4], Patrick Lilley[1,5]

[1] Liquid Biosciences, 6B Liberty, Suite 228, Aliso Viejo, CA 92656
[2] Leeds Institute of Rheumatic and Musculoskeletal Medicine, University of Leeds, LS9 7TF, Leeds, UK.
[3] Montana Department of Health and Human Services, Helena, MT 59601
[4] Department of Biochemistry and Molecular Medicine, George Washington School of Medicine and Health Sciences, Washington, D.C. 20052
[5] Ignite Biomedical, Medfield, MA 02052

*kevinh@liquidbiosciences.com

Abbreviations

ACR - American College of Rheumatology

CDAI - clinical disease activity index

CRP - C-reactive protein

DAS28 - disease activity score 28

DLDA - diagonal linear discriminant analysis

DMARD - disease modifying anti-rheumatic drugs

DQDA - diagonal quadratic discriminant analysis

EC: - evolutionary computation

EULAR - European League Against Rheumatism

GEO - gene expression omnibus;

LOO - leave one out

LDA -linear discriminant analysis

MTX - methotrexate

OLS - orthogonal least squares

RA - rheumatoid arthritis

RF - random forest

SVM - support vector machine

TNF - tumor necrosis factor

UHC - unsupervised hierarchical clustering

Abstract

Background: Producing transparent interpretable algorithms summarizing clinical trial outcomes to accurately predict individual patient's responses would be a significant advance. We hypothesized that software designed to analyze biomedical data, based on evolutionary computation (EC), could produce summary algorithmic biomarkers from a clinical trial, predictive of individual responses to therapy.

Methods and Findings: A previously published randomized double-blind placebo controlled clinical trial was analyzed. Patients with active rheumatoid arthritis on a stable dose of methotrexate and naive to anti-tumor necrosis factor biologic therapy, were randomized to receive infliximab or placebo. The primary endpoint was synovial disease activity assessed by magnetic resonance imaging. Secondary endpoints included the Disease Activity Score 28 (DAS28). Baseline peripheral blood gene expression variable data were available for 59 patients, plus the treatment variable, infliximab or placebo, yielding a total of 52,379 baseline variables. The binary dependent variable for analysis was DAS28 response, defined by a decrease in DAS28 score of ≥1.2, at 14 weeks. At 14 weeks, 20 of the 30 patients receiving infliximab had responded, and ten of the 29 patients receiving placebo had responded. The software derived an algorithm, with 4 gene expression variables plus treatment assignment and 12 mathematical operations, that correctly predicted responders versus non-responders for all 59 patients with available gene expression data, giving 100% accuracy, 100% sensitivity and 100% specificity. We present the algorithm to provide transparency and to enable verification. Excluding the 4 gene expression variables, we then derived similarly predictive algorithms with 4 other gene expression variables. We hypothesized that the software could derive

algorithms as predictors of treatment response to anti-tumor necrosis factor biologic therapy using just these 8 gene expression variables using previously published independent datasets from 6 rheumatoid arthritis studies. In each validation analysis the accuracy of the predictors we derived surpassed those previously reported by the original study authors.

Conclusions and Relevance: Software based on EC summarized the outcome of a clinical trial, with transparent biomarker algorithms correctly predicted the clinical outcome for all 59 RA patients. The biomarker variables were validated in 6 independent RA cohorts. This approach simplifies and expedites the development of algorithmic biomarkers accurately predicting individual treatment response, thereby enabling the deployment of precision medicine, and, in the future, providing a basis for dynamic labeling of prescription drugs. **Original Trial Registration used for analysis:** ClinicalTrials.gov registration: NCT01313520

Introduction

Randomized clinical trial outcomes are conventionally analyzed using a prespecified hypothesis and statistical tests based on the average overall response in order to assess whether treatment is effective at a population level. Much information in the clinical trial data is neglected with this approach resulting in a lack of guidance for clinicians and patients as to which individuals might benefit from the therapy. This is reflected in product labels that are devoid of individual guidance. An analytic methodology is needed that incorporates all the available trial data to yield predictors of therapeutic outcomes in individual patients.

An algorithm is a formula for processing and summarizing information enabling predictions and understanding of the underlying mechanisms. An analytic approach that utilizes all information from a clinical trial, and mathematically derives an objective transparent interpretable digital algorithmic biomarker set predicting individual response would be a significant advance. Individual treatment recommendations could then be based on a quantitative objective predictive model, easily validated, fulfilling the goal of precision medicine by enabling the right patient to receive the right therapy. The aspiration that machine learning AI would enable this transformative vision to be fulfilled has not yet been realized, hindered by lack of transparency [1, 2]. Living organisms are complex systems [3]. Biological functions, such as disease and drug responsiveness, result from complex networks of cellular and molecular interactions that can potentially be described mathematicallyreflecting mechanisms and constraints under which biological systems operate [3-6]. These interactions are typically nonlinear and multi-dimensional which complicates analyses.

120 We developed software based on EC, incorporating a comprehensive set of
121 mathematical functions, to produce transparent interpretable algorithms from
122 complex biomedical data, specifically designed to address nonlinearity and high
123 dimensionality typical of complex emergent systems. The algorithms produced are
124 verifiably accurate mathematical solutions to the outcome of interest.

125 The objective was to apply the software to a published placebo-controlled clinical
126 trial of infliximab in rheumatoid arthritis (RA), conducted by a pharmaceutical
127 company with baseline peripheral blood transcriptomic data [7,8]. Infliximab is a
128 monoclonal anti-tumor necrosis factor (TNF) alpha antibody effective in treating a
129 range of immune mediated disease including RA. The hypothesis was that the
130 software would produce transparent, interpretable, algorithms predictive of treatment
131 response that could be independently validated. We selected this trial because it
132 was randomized, and placebo controlled with tens of thousands of data points
133 available for each patient to provide a rigorous test. The software produced an
134 algorithm comprised of biomarker measurements, clinical variables, and
135 mathematical functions, highly predictive of infliximab treatment outcome containing
136 just four gene expression variables and treatment assignment. When we excluded
137 those four gene expression variables in the original algorithm from subsequent
138 analyses, we were able to derive additional algorithms, with four additional gene
139 expression variables, with similar predictivity.

140 Having identified the eight gene expression variables, we then hypothesized that
141 applying the software using only those eight variables, to previously published data
142 from six additional RA studies containing baseline gene expression data would
143 derive predictors of clinical outcomes following anti-TNF treatment. In each case,

using these eight gene expression variables, the software provided an algorithm more predictive of treatment response than reported in the original publications, thereby independently validating both the gene expression variables as response predictors, and our methodologic approach.

Methods

Discovery Cohort

The discovery cohort was from a published, four European center, randomized, double-blind placebo-controlled trial of infliximab, in RA patients naïve to biologic therapy following an inadequate response to methotrexate [7,8]. Active disease was defined as at least 6 tender and 6 swollen joints, with a rheumatoid arthritis magnetic resonance imaging score of $\geq$1 in the radio-carpal or intercarpal joints, as objective confirmation of disease activity. The patients were ~~all~~ on stable doses of methotrexate, steroids, and/or non-steroidal anti-inflammatory drugs. At weeks 0, 2, 6, and 14, the participants received either infliximab 3 mg/kg or placebo. The patients were of mean age of 50 years~~,~~ predominantly female (92%), and rheumatoid factor positive (91.5%), with a mean baseline DAS28 score of 6.2. The primary endpoint was a magnetic resonance imaging assessment of disease activity. This endpoint was not used in our analysis, because the imaging data were not available for individual patients. The trial used the European League against Rheumatism (EULAR) DAS28 score to evaluate response, defined as a decrease of $\geq$1.2, as a binary dependent variable for the analysis at 14 weeks, yes or no. Baseline peripheral blood gene expression data for 59 patients were available plus an additional variable reflecting treatment, either infliximab or placebo, resulting in a total of 52,379 potentially independent variables.

168 **Validation Cohorts**

169 To independently validate the gene expression variables in the discovery algorithms,
170 we applied these variables to 6 previously published studies with available baseline
171 gene expression data from patients with active RA, treated with anti-TNF therapies
172 [9-14]. These studies were done in different geographies, used different
173 methodologies to process the samples and gene expression data, had different
174 endpoints, and in some cases, used different anti-TNF therapies [9-14]. Summary
175 details on the discovery and validation studies are provided in Table 1, with details
176 available in the original publications [9-14].

177 **Table 1 – Discovery and validation studies of blood-based gene expression data of**
178 **anti-TNF responsiveness in RA**

First author	MacIsaac et al (2014)[7,8]	Lequerre et al (2005)[9]	Julia et al (2009)[10]	Bienkowska et al (2009)[11]	Toonen et al (2011)[12]	Nakamura et al (2016)[13]	Tanino et al (2010)[14]
GEO accession	GSE58795	GSE5392	GSE12051	GSE15258	GSE33377	GSE78068	GSE20690
	Discovery	Validation	Validation	Validation	Validation	Validation	Validation
Location	Moldova & Romania	France	Spain	USA	Netherlands	Japan	Japan
RA diagnostic criteria	RA ACR (1987)	RA ACR (1987)	RA ACR (1987)	RA	RA ACR (1987)	RA ACR (1987) or EULAR/ACR (2010)	Not reported
Treatment	infliximab or placebo	infliximab	infliximab	infliximab adalimumab, etanercept	infliximab and adalimumab	infliximab	infliximab
RA treatment population	MTX resistant. No prior TNF therapy	MTX resistant. DAS28>=5.1	MTX resistant. DAS28>3.2. No prior TNF therapy	Active RA: No TNF in past 6 months	DMARD resistant. DAS28>3.2. No prior TNF therapy	MTX resistant	MTX resistant
Platform	Rosetta/Merck custom Affymetrix 2.0 microarray	Affymetrix Human Genome U133A Array	Illumina H-6 mRNA Sentrix Human-6 Expression BeadChip	Affymetrix Human Genome U133 Plus 2.0 Array	Affymetrix GeneChip Exon 1.0 ST	Agilent-014850 Whole Human Genome Microarray 4x44K G4112F	Agilent-014850 Whole Human Genome Microarray 4x44K G4112F
Efficacy	14-week EULAR criteria	14-week EULAR criteria	14-week EULAR criteria		14-week EULAR Criteria	6-month CDAI	14-week serum CRP

179
180 GEO – gene expression omnibus; RA – rheumatoid arthritis; ACR – American College of Rheumatology; MTX –
181 methotrexate; EULAR – European League Against Rheumatism; DAS28 – disease activity score 28; TNF –
182 tumor necrosis factor; DMARD – disease modifying anti-rheumatic drugs; CRP – C-reactive protein; CDAI –
183 clinical disease activity index.

184 **Microarray data**

185 The gene expression profiles analyzed were downloaded from the GEO database.
186 Because all the data used were de-identified and publicly available, neither ethics
187 committee approval nor informed consent were required.

188 Datasets were downloaded and transposed so that gene expression values and
189 clinical variables were changed from rows to columns, and subject records were
190 changed from columns to rows. Data files were saved in CSV format and imported
191 into the software for analysis, without any processing.

192

193 **Analysis**

194 The software is a quantitative artificial intelligence platform based on EC, designed
195 as a scalable, unbiased methodology to produce transparent algorithms based on
196 mathematical relationships from complex data, without any prior assumptions other
197 than the patient selection criteria used in the studies yielding data for analysis. The
198 software fuses evolutionary principles, signal processing functions, and information
199 theory, and requires no domain expertise or prior knowledge of the nature of a
200 problem in terms of explanatory variables, dimensionality or underlying mathematical
201 relationships. A distinctive feature is that the software uses all available data to
202 derive the algorithms, without any filtering process to exclude variables based on
203 commonly used thresholds or feature selection methods. This enables the
204 identification of variables typically discarded by feature selection methods used in
205 biomarker discovery with relatively low expression or non-linear relationships to the
206 outcome, but which may be functionally important because of the non-linear and

207 binary threshold interactions pervasive in complex biologic systems. The software
208 identifies both key variables and their mathematical relationships, associated with
209 outcomes of interest. The software automatically divides the data into three distinct,
210 random subsets that are sequentially processed: a training set, a selection set, and a
211 test set. Analysis of the training subset provides an ensemble of candidate
212 algorithms, which are then evaluated on the selection subset, to select a final
213 algorithm, which is then validated on the test set. The training, selection, and test
214 data subsets are scrupulously segregated, to avoid any information leakage between
215 the discrete components of the process. In the discovery analysis of the MacIsaac et
216 al. data, there were 20 patients in the training set, 20 in the selection set and 19 in
217 the test set (Table 2).

218 **Table 2 Discovery Algorithm Metrics**

	Analysis of data from MacIsaac et al.[7]
Training set	
Errors/total	0/20
Accuracy	100%
Misclassification rate	0%
Validation set	
Errors/total	0/20
Accuracy	100%
Misclassification rate	0%
Test set	
Errors/total	0/19
Accuracy	100%
Misclassification	0%
Overall Accuracy	100%
Overall Sensitivity	100%
Overall Specificity	100%
Number of gene expression variables	4

219

220 The 8 variables in the discovery algorithms were then applied to the analyses of six
221 additional published data sets in patients with RA, using both baseline gene

222 expression data and response outcomes to anti-TNF therapy, for independent

223 validation.

224 Findings

225 Discovery Analysis: Derivation of MacIsaac et al Algorithm

226 The software initially yielded an algorithmic biomarker set with five variables and

227 twelve sequential mathematical instructions as shown in Table 3. The Gene

228 expression variables expressed quantitatively: *SPTY2D1*, *Clorf105*, *KCTD4* and

229 *UL84* and treatment assignment: infliximab or placebo.

230 **Table 3 Discovery Algorithm as Series of Operations**

231

Instruction	Input Registers	Explanation
r[03] = SINCPI(r[03])	r[03] = 19	Calculate the sine of 19, divide that by 19 and push the result (0.00788827419278696) into memory register 03
r[06] = SUB(r[13], r[02])	r[13] = KCTD4, r[02] = 7	Subtract 7 from the measurement of RNA KCTD4, then push the result into register 06
r[04] = ADD(r[15], r[01])	r[15] = Treatment_num, r[01] = 2	Add Treatment_num to 2 and push the result into register 04. Treatment_num 0 is placebo, Treatment_num 1 is infliximab.
r[05] = ADD(r[10], r[06])	r[10] = SPTY2D1	Add the value in register 06 to the measurement of RNA SPTY2D1 and push the result into register 05
r[01] = DIV(r[02], r[04])	r[02]= 7	Divide 7 by the value in register 04 and push the result into register 01
r[00] = SINCPI(r[09])	r[09] = UL84	Calculate the sine of the measurement of RNA UL84, divide that by the measurement of RNA UL84, and push the result into register 00
IF(r[00] < r[03])		If the value in register 00 is less than the value in register 03, then execute the following instruction, otherwise skip it
IF(r[11] < r[01])	r[11] = C1orf105	If the measurement of RNA C1orf105 is less than the value in register 01, then execute the following instruction, otherwise skip it
r[00] = ADD(r[04], r[05])		If both of the above IF statements are TRUE, then add the value in register 04 to the value in register 05 and push the result into register 00. If one or both of the above IF statements are false, skip this instruction.
r[05] = DIV(r[11], r[00])	r[11] = C1orf105	Divide the measurement of RNA C1orf105 by the value in register 00, and push the result into register 05
r[04] = CONSTDIV(7, 5)		Divide 7 by 5 and push the result into register 04
r[00] = ADD(r[04], r[05])		Add the value in register 04 to the value in register 05, and push the result into register 00
Final interpretation		If the contents of memory register 00 are negative, then patient is predicted to be a non-responder, otherwise a responder

232

233 The algorithm is presented in the form of an equation here.

$$\left\{ \frac{\text{C1orf105}}{\frac{7}{\boldsymbol{treatment}+2} + \boldsymbol{KCTD4} - 7 + \boldsymbol{SPTY2D1}, \begin{matrix} \left[\frac{\sin(UL84)}{UL84} < \frac{\sin(19)}{19}\right] AND \left[\text{C1orf105} < \frac{7}{treatment+2}\right] \\ \frac{\sin(\boldsymbol{UL84})}{\boldsymbol{UL84}}, \left[\frac{\sin(UL84)}{UL84} \geq \frac{\sin(19)}{19}\right] OR \left[\text{C1orf105} \geq \frac{7}{treatment+2}\right] \end{matrix}} + \frac{7}{5} \right\} < 0$$

To facilitate validation the algorithm is also shown as a schematic in Fig 1.

A calculation for an individual patient, with a resulting value of less than zero indicated treatment non-response, and a value of zero or more indicated treatment response. The baseline values for the four gene expression markers, for each of the 59 patients, are present in S1 data`to enable verification of the algorithm.

Fig 1 Legend: Discovery Algorithm in Schematic Form

The variables were treatment assignment (placebo versus infliximab therapy) and four gene expression variables: SPTY2D1, C1orf105, KCTD4 and UL84 (Table 4).

Table 4 Gene Expression Variables

Expression Variable		Function
SPTY2D1	suppressor of Ty 5 homolog	tumor suppressor gene in thyroid cancer [15]
C1orf105	chromosome 1 open reading frame	vascular remodeling, coronary artery disease and atrial fibrillation [16,17]
KCTD4	potassium channel tetramerization domain containing 4	expressed in sepsis and esophageal cancer [18, 19]
UL84	gene for human CMV protein UL84	prior CMV infection associated with treatment response in RA [20,21]
PTPRC	Protein Tyrosine Phosphatase Receptor Type C or CD45	genetic variants associated with anti-TNF response in RA [22,23]
TPM3	Tropomyosin 3	tropomyosin isoform associated with neoplasia, myopathy and as an autoantigen in RA [24-26]
ARHGDIB	Rho GDP Dissociation Inhibitor Beta	negative regulator of Rho guanosine triphosphate (RhoGTP)ases, reduced in osteoarthritis synovial fluid and elevated in early RA synovial fluid [27,28]
SIAH1	Siah E3 Ubiquitin Protein Ligase 1	downregulated in both RA and tuberculosis, may influence amplitude of inflammatory gene response [29]

These variables were not those with the highest levels of expression. Agnostic evolutionary selection of treatment as a variable, either infliximab or placebo, into the predictive algorithm is evidence that the treatment has a mathematically significant impact on the response outcome for some patients. Although expected a priori, the treatment assignment variable was selected agnostically by the evolutionary process, and not pre-specified. Users of the software cannot influence or require any variable to be incorporated into an algorithm.

The performance metrics for the components of the discovery analyses are shown in Table 2, with 20 patients in the training set, 20 in the selection set and 19 in the test set. The overall accuracy was 100% with 100% sensitivity and specificity. Repeat analyses consistently yielded exactly the same variables, and the mathematical instructions encoded in all algorithms were, without exception, mathematically equivalent in terms of binary outcome: response or non-response. Omission of any of the mathematical instructions from the algorithm degraded predictivity. Therefore, the algorithm was optimized and devoid of superfluous calculations. Accuracy and reliability were consistent across training, validation and test sets analysed. When the initial four gene expression variables were excluded from subsequent analyses, four additional gene expression variables were identified as components of algorithms also with 100% accuracy: PTPRC, TPM3, ARHGDIB and SIAH1 (Table 4). The software selects the minimum number of variables for maximum accuracy. The fact that a second group of four variables provided algorithms with 100% accuracy implies that these variables are highly correlated with the original four, containing almost as much information. The x-y plots of the 8 individual markers versus DAS28 change are nonlinear: see S1 data.

Validation of gene expression variables by analysis of six independent datasets

We then tested the hypothesis that the eight variables from the discovery algorithms could be used to derive algorithms predicting individual treatment response outcomes from six additional published data sets, providing independent validation. 9-14). While the eights variables used were the same for all six dataset analyses, distinct algorithmic operations were necessary for each validation because of variation in quantitation instruments, measurement scale, data preparation and normalization methods across the individual datasets. Demographic details are shown in Table 1. In each case, the algorithms we produced had superior performance to the analyses presented in the original publications as shown in Table 5, even though only a subset of the eight discovery variables were available in some datasets. Table 5 shows the metrics for predictors presented in the original publications for each study, and also the metrics for the consistently superior algorithmic predictors we derived using the data from each study.

Table 5 – Discovery and validation studies of blood-based gene expression data of anti-TNF responsiveness in RA

First author	MacIsaac [7]	Lequerre [9]	Julia [10]	Bienkowska [11]	Toonen [12]	Nakamura [13]	Tanino [14]
Number of patients	59	30	44	46 – leaving out intermediate responses or 75 including intermediate responses	42	140 (infliximab treated)	68
Classifier Method	OLS regression	UHC	SVM, DDA, RF, & k-NN	RF	K-means cluster analysis	Logistic regression	LOO
Cross-validation analyses	None	LOO	LOO	LOO, NN, LDA and SVM	None	None	None
Original reported results	None reported	Sens 90% Spec 70%	Sens 94% Spec 86%	Sens 88% Spec 84%	Sens 61% Spec 71%	Sens 79% Spec 47%	Sens 68% Spec 86%

Current analysis	Sens 97% Spec 100%	Sens 100% Spec 100%	Sens 100% Spec 86%	Sens 100% Spec 86%	Sens 100% Spec 96%	Sens 95% Spec 91%	Sens 96% Spec 84%

UHC – unsupervised hierarchical clustering; OLS – orthogonal least squares; LOO – leave one out; LDA - linear discriminant analysis; SVM – support vector machine; RF – random forest; DLDA – Diagonal Linear Discriminant Analysis; DQDA – Diagonal Quadratic Discriminant Analysis.

The weighted-average sensitivity and specificity of the individual models from the analyses we conducted across all 7 datasets, with a total of 400 subjects treated with anti-TNF therapy, was 98.5% and 90.9%, respectively surpassing the accuracy of all the prior published individual analyses, as shown in Fig 2. The biomarker variables held up across 7 patient datasets, 3 different TNF inhibitor drugs, 3 continents, multiple ethnicities, despite differences in response criteria, RNA expression and processing platforms and measurement scale differences.

Fig 2 Discovery and validation studies of blood-based gene expression data of anti-TNF responsiveness in RA

Weighted-average accuracy of individual analyses of 7 datasets, on all 400 subjects receiving anti-TNF therapy [7-14]. 98.5% sensitivity and 90.9% specificity surpassed all individual previously reported analyses, as did our 84.5% positive predictive value, and 99.2% negative predictive value. Accuracy for the marketed TNF predictor from two reports, Jones et al and Mellors et al, is also presented for comparison [30,31].

Discussion

We have shown that an EC based analytic approach incorporating mathematical functions derives transparent interpretable biomarker algorithms highly predictive of individual patient responses from clinical trials. The algorithms and their high predictivity are notable, since none of the predictive biomarkers we identified were identified in the original clinical trial publication used for discovery [8]. In each of the six validation analyses, the derived algorithms had sensitivity and specificity superior to those in the original publications, which used a variety of mainstream analytic

316 approaches: see Fig 2. None of the publications associated with the six validation
317 datasets reported clinical utility of the eight predictive biomarkers identified in our
318 analyses. Context for comparatively assessing the approach we used is provided by
319 an existing marketed predictive test, derived using machine learning, for TNF
320 response. This predictive test identified non-responders with a specificity of 86.8 and
321 sensitivity of 50.0% in one report by Mellors et al and specificity of 77.3% and
322 sensitivity of 60.2% in another study by Jones et al [30,31]. The predictive test
323 contains 23 variables: 19 gene expression variables, plus anti-cyclic citrullinated
324 protein serostatus, sex, body mass index, and a patient global assessment [30].
325 None of the gene expression variables in this RA predictor overlap with the 8
326 variables we identified. In contrast, the weighted average of our RA treatment
327 response predictor has a sensitivity of 98.5% and specificity of 90.4%, although this
328 is not a direct comparison using the same datasets. The underlying data used to
329 derive the marketed predictive test and its algorithm were not available to analyze for
330 a direct comparison with our analyses of the seven datasets we used for discovery
331 and validation.

332 The use of this form of EC to derive an explanatory and predictive algorithm
333 summarizing a clinical trial outcome is novel. EC has been claimed to have
334 significant advantages leading to its description as a major transition in AI [32]. EC
335 uses core concepts from evolutionary biology - inheritance, random variation, and
336 selection - to derive algorithms from data which may make it particularly suitable for
337 analysis of biomedical data [32]. This form of EC leverages the strengths of EC in
338 general and does not depend on prior assumptions, provides highly interpretable
339 algorithms, and can provide solutions from modest sample sizes [32]. In addition, the

340 EC approach is entirely agnostic, and does not use any arbitrary thresholds to
341 eliminate variables from consideration. These various advantages are confirmed by
342 our findings, and contrast with inflexible analytic methods, that may create the
343 erroneous impression of greater noise and randomness than is actually present in
344 the observed data. We presented an example of an algorithm to enable conclusive
345 confirmation of its validity and provide easy access to the data used for its derivation
346 in the supplementary material.
347
348 The eight specific gene expression variables in our predictive algorithms may
349 provide insights into the molecular mechanisms underlying responsiveness to anti-
350 TNF therapy in RA. The eight variables were not individually highly correlated with
351 response, likely reflecting the importance of nonlinear interactions in biology,
352 modelled by the mathematical components of the algorithms. There are no prior
353 reports directly linking SPTY2D1, KCTD4, and c1orf105 to either RA, immune-
354 mediated diseases or to responses to therapy. Prior cytomegalovirus (CMV)
355 exposure has been associated with poor responses to therapy in early RA [20,21,31]
356 which may explain the UL84 gene variable as it encodes for a CMV protein,
357 essential for viral replication. Mutations in PTPRC, also known as CD45, have been
358 associated with RA patient response to anti-TNF therapy [22,23]. TPM3 has been
359 reported as an autoantigen in in RA, and also to involved in myopathy and renal
360 cancer [24,25,26]. ARHGDIB expression has been reported to be upregulated in RA,
361 whereas SIAH1 is downregulated [28,29]. TPM3, ARHGDIB and SIAH1 have not
362 been associated with therapeutic outcomes.
363

364 Medical information is accumulating at an accelerating pace, creating challenges for
365 clinicians to process, discuss with patients and implement objective data driven
366 therapeutic recommendations [33,34]. The notion that the future of medicine will
367 ultimately be based on computer derived algorithms is not recent or novel [33-35].
368 More than 50 years ago, William Schwartz predicted that computing would be
369 augmenting and, in some cases, even replacing the intellectual functions of
370 physicians [35]. The practice of medicine has been "algorithmic" even before the
371 introduction of computers to medicine and the digitization of medical information [36].
372 Diagnostic and therapeutic clinical guidelines, often described as algorithms, for
373 many prototypical clinical syndromes, and codified by experts, have been integral to
374 the practice of medicine for decades [36,37]). As Margolis wrote in 1983, clinical
375 "algorithms are needed primarily to define the practice of rational medicine" [36]. The
376 progression to a more prominent role of software-derived algorithms in medicine is
377 inevitable, given the current explosion of data [34,35,38,39).

378 A novel analytic approach that reflects the entirety of the trial data to inform
379 individual treatment decisions, in accordance with the aspirations of precision
380 medicine is desirable, especially now that clinical trials are frequently collecting
381 digital and multi-omic data sets. The relatively simple, objective, transparent,
382 interpretable algorithms we report, the ease of their independent validation that we
383 have illustrated, and their superior performance are compelling advantages. In
384 contrast, the cryptic, opaque, black box products that typical AI machine learning
385 yields are not desirable for informing medical decisions [2,40].

386 The novel approach we report and the noteworthy findings we present all warrant
387 rigorous evaluation because of their potential ramifications. The biomarker variables

were discovered in a clinical study conducted in Romania and Moldova and were validated in 6 other independently conducted studies with different TNF inhibitors, on 3 continents, in multiple ethnicities, despite differences in response criteria, RNA expression and processing platforms and measurement scale differences, thereby providing extremely high confidence as to their validity. The software analysis is objective because no assumptions are made in order to influence its identification of the optimal variables or mathematical operations. Any potential bias would be a consequence of selection criteria and study designs of the original clinical studies that yielded the gene expression data. Our consistent findings across all the independent datasets argue strongly against any bias compromising the validity of the predictor variables. While the eight biomarkers are appropriate for clinical use, the discovery-stage gene expression instruments used to derive the datasets have different quantitation levels and reliability than clinical-grade assay processes and technology. Therefore, a diagnostic test is undergoing development for clinical use, incorporating the eight identified variables using clinical grade assay processes and technologies. A novel algorithm will then be derived and rigorously validated for the developed test to ensure optimal performance and clinical utility.

The algorithmic biomarkers we report can be updated and validated as new multi-faceted data become available, continually refining their accuracy and generalizability, and amplifying confidence in their utility. Such algorithms could be incorporated into dynamic product labels to ensure that individual treatment decisions optimally reflect the available data.

Conclusion

The provision of transparent highly predictive algorithmic biomarkers, with user friendly tools for deployment avoiding onerous calculations, will reduce the cognitive burden of physicians and help improve the quality of their consultations. Emancipating clinicians to rationally practice the art of medicine, on a quantitative, scientific foundation, informed by validated interpretable algorithms, derived by EC based quantitative AI, could accelerate the implementation of precision medicine, and enhance the quality of health care.

Acknowledgements: Richard Gallagher, Mary Beth Yacyshyn, Tony Ho, David Madigan and Poul Strange provided useful comments on early versions of the manuscript.

Author Contributions

Conceptualization: Kevin Horgan, Michael F. McDermott, Douglas Harrington, Vahan Simonyan, Patrick Lilley.

Data curation, project administration and software: Patrick Lilley

Methodology: Kevin Horgan, Vahan Simonyan, Patrick Lilley.

Interpretation: Kevin Horgan, Michael F. McDermott, Douglas Harrington, Vahan Simonyan, Patrick Lilley.

Formal analysis: Patrick Lilley.

Visualization: Vahan Simonyan, Patrick Lilley.

Writing – original draft: Kevin Horgan, Michael F. McDermott, Patrick Lilley.

Writing – review and editing: Kevin Horgan, Michael F. McDermott, Douglas Harrington, Vahan Simonyan, Patrick Lilley.

453 **References**

454 1. Wilkinson J, Arnold KF, Murray EJ, van Smeden M, Carr K, Sippy R et al. Time
455 to reality check the promises of machine learning-powered precision medicine.
456 Lancet Digit Health. 2020;12:e677-e680. PMID: 33328030.
457 2. Price WN. Big data and black-box medical algorithms. Sci Transl Med.
458 2018;10:471. PMID: 30541791
459 3. Nurse P. Life, logic and information. Nature. 2008;454(7203):424-6. PMID:
460 18650911.
461 4. Berger SI, Iyengar R. Role of systems pharmacology in understanding drug
462 adverse events. Wiley Interdiscip Rev Syst Biol Med. 2011;3(2):129-35. PMID:
463 20803507
464 5. Savic S, Caseley EA, McDermott MF. Moving towards a systems-based
465 classification of innate immune-mediated diseases. Nat Rev Rheumatol.
466 2020;4:222-237.PMID: 32107482.
467 6. Coffey DS. Self-organization, complexity and chaos: The new biology for
468 medicine. Nat Med. 1998;8:882-5. PMID: 9701230.
469 7. MacIsaac K, Baumgartner R, Kang J, Loboda A, Peterfy C, DiCarlo J, et al.
470 Pre-treatment whole blood gene expression is associated with 14-week
471 response assessed by dynamic contrast enhanced magnetic resonance
472 imaging in infliximab-treated rheumatoid arthritis patients. PLoS ONE.
473 2014;9(12):e111937. PMID: 25504080
474 8. Beals C, Baumgartner R, Peterfy C, Balanescu A, Mirea G, Harabagiu A, et al.
475 Magnetic resonance imaging of the hand and wrist in a randomized, double-
476 blind, multicenter, placebo-controlled trial of infliximab for rheumatoid arthritis:
477 Comparison of dynamic contrast enhanced assessments with semi-quantitative
478 scoring. PLoS ONE. 2017;12(12):e0187397. PMID: 29236711
479 9. Lequerre T, Gauthier-Jauneau A, Bansard C, Derambure C, Hiron M, Vittecoq
480 O, et al. Gene profiling in white blood cells predicts infliximab responsiveness in
481 rheumatoid arthritis. Arth Res Ther. 2006; 8: R105. PMID: 16817978
482 10. Julia A, Erra A, Palacio C, Tomas C, Sans X, Barcel P, et al. An eight-gene
483 blood expression profile. PloS ONE. 2009;4:e7556. PMID: 19847310
484 11. Bienkowska J, Dagin G, Batliwalla F, Allaire N, Roubenoff R, Gregersen PK, et
485 al. Convergent random forest predictor: Methodology for predicting drug
486 response from genome-scale data applied to anti-TNF response. Genomics.
487 2009; 94:423-32. PMID: 19699293
488 12. Toonen EJ, Gilissen C, Franke B, Kievit W, Eijsbouts AM, den Broeder AA, et
489 al. Validation study of existing gene expression signatures for anti-TNF
490 treatment in patients with rheumatoid arthritis. PLoS ONE. 2012;7(3):e33199.
491 PMID: 22457743
492 13. Nakamura S, Suzuki K, Iijima H, Hata Y, Lim CR, Ishizawa Y, et al.
493 Identification of baseline gene expression signatures predicting therapeutic
494 responses to three biologic agents in rheumatoid arthritis: a retrospective
495 observational study. Arth Res Ther. 2016;18:159. PMID: 27435242
496 14. Tanino M, Matoba R, Nakamura S, Kameda H, Amano K, Okayama T, et al.
497 Prediction of efficacy of anti-TNF biologic agent, infliximab, for rheumatoid
498 arthritis patients using a comprehensive transcriptome analysis of white blood

499 cells. Biochem Biophys Res Commun. 2009 18;387(2):261-5. PMID:
500 19577537.
501 15. Ramírez-Moya J, Wert-Lamas L, Acuña-Ruíz A, Fletcher A, Wert-Carvajal C,
502 McCabe CJ, et al. Identification of an interactome network between lncRNAs
503 and miRNAs in thyroid cancer reveals SPTY2D1-AS1 as a new tumor
504 suppressor. Sci Rep. 2022; 12(1):7706. PMID: 35562181
505 16. Li K, Kong R, Ma L, Cao Y, Li W, Chen R, et al. Identification of potential M2
506 macrophage-associated diagnostic biomarkers in coronary artery disease.
507 Biosci Rep. 2022;42(12):BSR20221394. PMID: 36222281
508 17. Zhang J, Huang X, Wang X, Gao Y, Liu L, Li Z, et al. Identification of potential
509 crucial genes in atrial fibrillation: a bioinformatic analysis. BMC Med Genomics.
510 2020;**13**(1):104. PMID: 32682418
511 18. Kim S, Noh JH, Lee MJ, Park YJ, Kim BM, Kim YS et al. Effects of
512 Mitochondrial Transplantation on Transcriptomics in a Polymicrobial Sepsis
513 Model. Int J Mol Sci. 2023;24(20):15326. PMID: 37895006
514 19. Zheng C, Yu X, Xu T, Liu Z, Jiang Z, Xu J et al. KCTD4 interacts with CLIC1 to
515 disrupt calcium homeostasis and promote metastasis in esophageal cancer.
516 Acta Pharm Sin B. 2023 13(10):4217-4233. PMID: 37799381
517 20. Colletti KS, Smallenburg KE, Xu Y, Pari GS. Human cytomegalovirus UL84
518 interacts with an RNA stem-loop sequence found within the RNA/DNA hybrid
519 region of oriLyt. J Virol. 2007; 81:7077-85. PMID: 17459920
520 21. Davis JM, Knutson KL, Strausbauch MA, Green AB, Crowson CS, Therneau
521 TM et al. Immune response profiling in early rheumatoid arthritis: discovery of a
522 novel interaction of treatment response with viral immunity. Arthritis Res Ther.
523 2013;15(6):R199. PMID: 24267267
524 22. Cui J, Saevarsdottir S, Thomson B, Padyukov L, van der Helm-van Mil A,
525 NitithamJ, et al. Rheumatoid arthritis risk allele PTPRC is also associated with
526 response to anti-tumor necrosis factor alpha therapy. Arthritis Rheum.
527 2010;62(7):1849-61. PMID: 20309874
528 23. Plant D, Prajapati R, Hyrich KL, Morgan AW, Wilson AG, Isaacs JD, et al.
529 Replication of association of the PTPRC gene with response to anti-tumor
530 necrosis factor therapy in a large UK cohort. Arthritis Rheum. 2012;64(3):665-
531 70. PMID: 21952740
532 24. Poulsen TBG, Damgaard D, Jørgensen MM, Senolt L, Blackburn JM, Nielsen
533 C, et al. Identification of Novel Native Autoantigens in Rheumatoid Arthritis.
534 Biomedicines. 2020;8(6):141. PMID: 32486012
535 25. Lambert MR, Gussoni E. Tropomyosin 3 (TPM3) function in skeletal muscle
536 and in myopathy. Skelet Muscle. 2023;13(1):18. PMID: 37936227
537 26. Galea LA, Hildebrand MS, Witkowski T, Joy C, McEvoy CR, Hanegbi U, Aga A.
538 ALK-rearranged renal cell carcinoma with TPM3::ALK gene fusion and review
539 of the literature. Virchows Arch. 2023 482(3):625-633. PMID: 36370168
540 27. Ritter SY, Subbaiah R, Bebek G, Crish J, Scanzello CR, Krastins B, et al.
541 Proteomic analysis of synovial fluid from the osteoarthritic knee: comparison
542 with transcriptome analyses of joint tissues. Arthritis Rheum. 2013;65(4):981-
543 92. PMID: 23400684
544 28. Lequerré T, Bansard C, Vittecoq O, Derambure C, Hiron M, Daveau M, et
545 al. Early and long-standing rheumatoid arthritis: distinct molecular signatures

546 identified by gene-expression profiling in synovia. Arthritis Res Ther.
547 2009;**11**:R99. PMID: 19563633
548 29. Badr MT, Häcker G. Gene expression profiling meta-analysis reveals novel
549 gene signatures and pathways shared between tuberculosis and rheumatoid
550 arthritis. PLoS ONE. 2019;14(3):e0213470. PMID: 30845171
551 30. Mellors T, Withers JB, Ameli A, Jones A, Wang M, Zhang L, et al. Clinical
552 validation of a blood-based predictive test for stratification of response to tumor
553 necrosis factor inhibitor therapies in rheumatoid arthritis patients. Network and
554 Systems Medicine. 2020;3.1:91-104. doi: 10.1089/nsm.2020.0007
555 31. Jones A, Rapisardo S, Zhang L, Mellors T, Withers JB, Gatalica Z, et al.
556 Analytical and clinical validation of an RNA sequencing-based assay for
557 quantitative, accurate evaluation of a molecular signature response classifier in
558 rheumatoid arthritis. Expert Rev Mol Diagn. 2021;21(11):1235-43. PMID:
559 34727834.
560 32. Sipper M, Olson RS, Moore JH. Evolutionary computation: the next major
561 transition of artificial intelligence? BioData Min. 2017;10:26. PMID: 28775766
562 33. Herant M. Working paper: the coming age of algorithmic medicine. In: Recon
563 Strategy 2015 Oct 25 [cited January 21st 2024]. Available
564 from https://reconstrategy.com/2015/10/working-paper-the-coming-age-of-
565 algorithmic-medicine/
566 34. Obermeyer Z, Emanuel EJ. Predicting the Future - Big Data, Machine Learning,
567 and Clinical Medicine. N Engl J Med. 2016;375(13):1216. PMID: 27682033
568 35. Schwartz WB. Medicine and the computer - the promise and problems of
569 change. N Engl J Med. 1970;283:1257-64. PMID: 4920342
570 36. Margolis CZ. Uses of clinical algorithms. JAMA. 1983;249(5):627-32. PMID:
571 6336813
572 37. Keffer JH. Guidelines and algorithms: perceptions of why and when they are successful
573 and how to improve them. Clin Chem. 2001;47(8):1563-72. PMID: 11468268.
574 38. Warraich HJ, Califf RM, Krumholz HM. The digital transformation of medicine
575 can revitalize the patient-clinician relationship. NPJ Digit Med. 2018;1:49.
576 PMID: 31304328
577 39. Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. N Engl J Med.
578 2019;380(14):1347. PMID: 30943338
579 40. Naughton J. Machine-learning systems are problematic. That's why tech
580 bosses call them 'AI'. The Guardian. 2022 November 5 [Cited January 21st
581 2024]. Available from
582 https://www.theguardian.com/commentisfree/2022/nov/05/machine-learning-
583 systems-are-problematic-thats-why-tech-bosses-call-them-ai
584
585
586
587
588

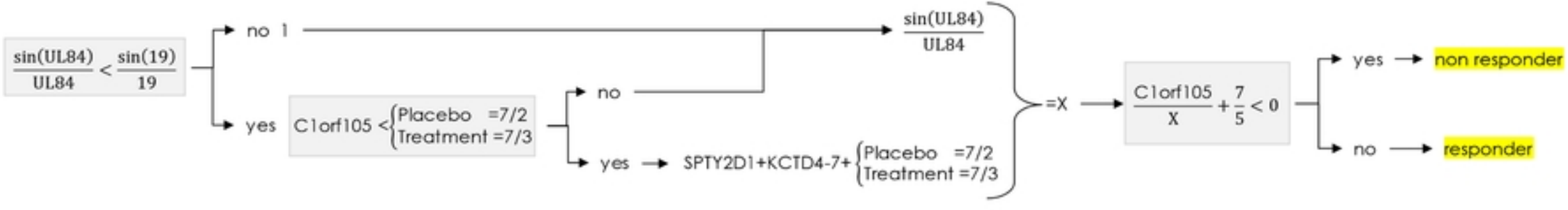
589 **Supporting information**

590 **S1 File**

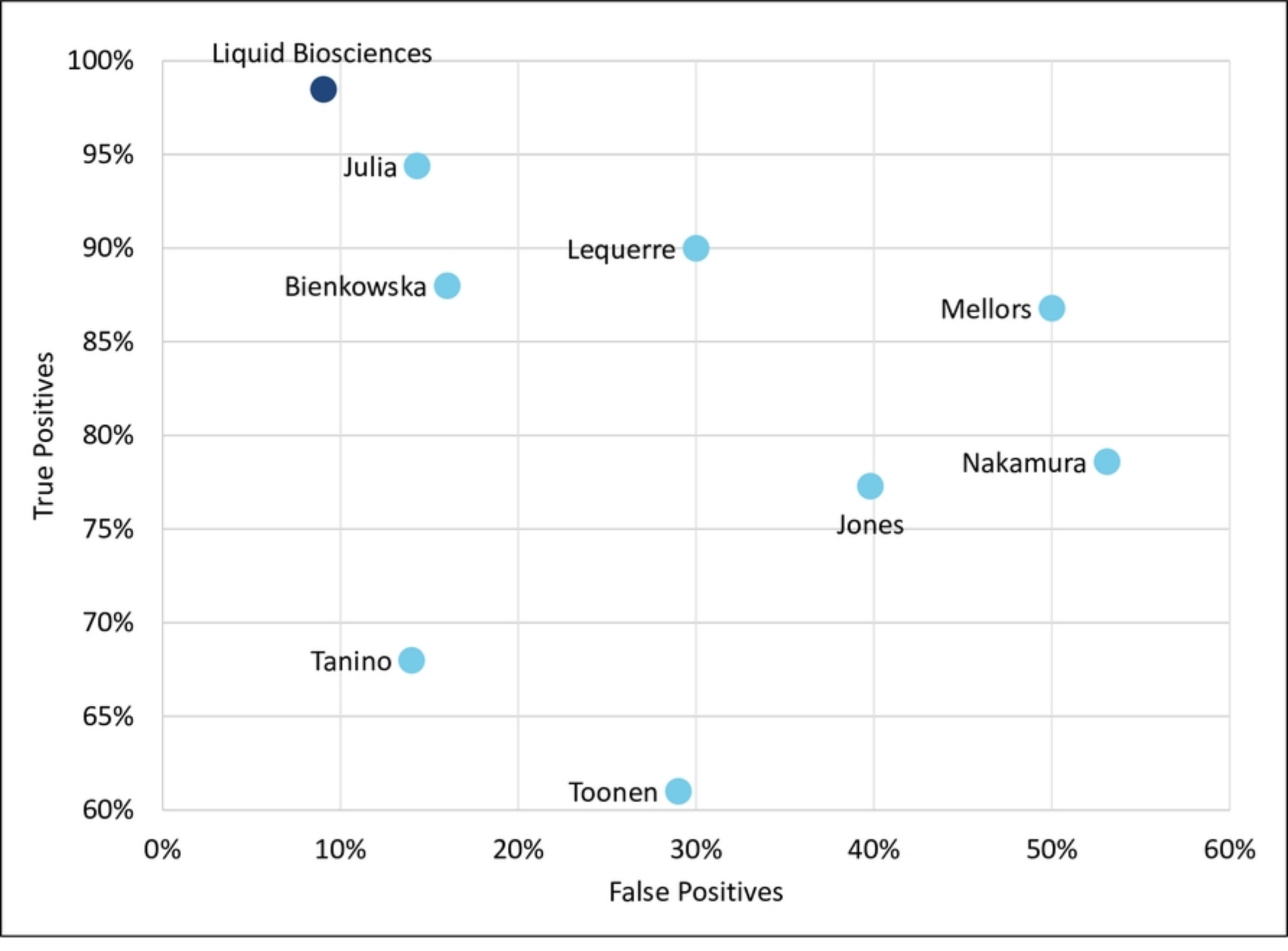
591 - Sheet 1: Data excerpted from GSE58795 on 4 gene expression variables in
592 discovery algorithm to enable validation
593 - Sheet 2: Summary gene expression for all eight variables by treatment group
594 - Sheets 3-8: x-y plots for DAS28 versus all eight variables by treatment group

595 **S2 File** – STARD Diagnostic Checklist

596


$\frac{\sin(UL84)}{UL84} < \frac{\sin(19)}{19}$
no 1
yes
C1orf105 < Placebo =7/2 Treatment =7/3
no
yes
SPTY2D1+KCTD4-7+ Placebo =7/2 Treatment =7/3
$\frac{\sin(UL84)}{UL84}$
=X
$\frac{C1orf105}{X} + \frac{7}{5} < 0$
yes
non responder
no
responder

Figure


Liquid Biosciences
Julia
Lequerre
Bienkowska
Mellors
Nakamura
Jones
Tanino
Toonen
100%
95%
90%
85%
80%
75%
70%
65%
60%
True Positives
0%
10%
20%
30%
40%
50%
60%
False Positives

Figure