Symbolic Regression for Discovery of Medical Equations: A Case Study on Glomerular Filtration Rate Estimation Equations

Kei Sen Fong*, Mehul Motani†
*†Department of Electrical and Computer Engineering, †Institute of Data Science, †N.1 Institute for Health,
†Institute for Digital Medicine (WisDM), National University of Singapore
Email: *fongkeisen@u.nus.edu, †motani@nus.edu.sg

***Abstract*—Symbolic Regression (SR) is a sub-field of machine learning that attempts to find a concise closed-form equation for regression tasks. The main strength of SR is in discovering new equations which are explainable and interpretable, in contrast to many black-box machine learning models. Thus, SR has become a first-class algorithm in various fields, including physics and engineering, and more recently healthcare. In this paper, we utilize and evaluate recent state-of-the-art (SOTA) SR methods and introduce our new proposed SR methods to the problem of estimating Glomerular Filtration Rate (GFR). GFR is a key indicator of kidney health and is an important determinant in certain diagnoses, such as Chronic Kidney Disease (CKD). As measuring GFR directly is expensive, the common practice for healthcare professionals is to use medical equations to estimate GFR from other biomarkers. However, most of these equations are built on specific populations and are potentially inaccurate for other populations. Additionally, there has been a growing need to tailor equations for cohorts with unique properties (e.g., young children, renal transplant patients). To address these issues, we propose to use SR to discover new equations to better estimate GFR for new populations or new cohorts. First, we utilize SOTA SR methods to discover equations that achieve better estimation performance than existing equations. Then, we introduce our novel method of taking existing GFR medical estimation equations as prior knowledge and evolving them via Constrained Genetic Perturbation (CGP). The prior knowledge used drastically reduces the search space and also discovers equations that closely resemble the functional forms familiar to medical professionals. We also introduce a variant that enables the equation to be modified with newly obtained features. Finally, we show that equations discovered by our methods demonstrate the best performance among all equations.**

***Index Terms*—Symbolic Regression, Glomerular Filtration Rate, Machine Learning, Chronic Kidney Disease**

I. Introduction

Symbolic Regression (SR) algorithms are machine learning algorithms that find concise closed-form equations for regression prediction tasks [1]. In contrast to black-box machine learning models, the equations produced by SR are explainable and interpretable. SR has thus become a first-class algorithm in various fields for equation discovery [2], including physics [3], material sciences [4] and engineering [5]. Recently, there is increased interest for SR in healthcare [6] because it produces explainable equations which are white-box, in contrast to black-box machine learning methods. In this paper, we utilize and evaluate recent state-of-the-art (SOTA) SR methods and introduce our new proposed SR methods to the problem of estimating the Glomerular Filtration Rate (GFR).

GFR is an important determinant in certain diagnoses, such as Chronic Kidney Disease (CKD) [7]. Measuring GFR directly is expensive and time-consuming, since it involves the plasma or urinary clearance of exogenous filtration markers (e.g., inulin and iohexol). Thus, the common practice is to use medical equations to compute an estimated GFR (eGFR) from other biomarkers. However, there are 3 major weaknesses of current eGFR equations: (i) Most of these eGFR equations are built on Caucasian and Afro-American populations and are potentially inaccurate for other populations [8]. (ii) Most eGFR equations have low performance on cohorts with unique properties (e.g., young children [9], renal transplant patients [10]), thus requiring development of new equations that are tailored to the specific cohort. (iii) The equation structure (i.e., equation excluding parameters) of current eGFR equations is often arbitrarily decided [9, 11, 12, 13, 14].

To address these weaknesses, we propose to use SR to discover new equations to better estimate GFR. Our SR methods (i) automate development of equations in new populations, (ii) allow for tailored equations to specific cohorts to improve prediction, and (iii) search a large variety of equation structures to pick the best structure based on data-driven evidence.

Our Contributions:

1) We discover equations using SOTA SR methods which select equation structures in a data-driven manner. These equations generate better estimates than current eGFR equations. We also propose a modification to an existing SR method to discover equations with conditions which perform even better.
2) We introduce our novel method of taking existing eGFR equations as prior knowledge and evolving them via Constrained Genetic Perturbation (CGP). The prior knowledge drastically reduces the search space and also discovers equations that closely resemble the functional forms familiar to medical professionals.
3) We develop further improvements on our method using a novel mechanism that enables existing eGFR equation to be modified with newly obtained features.
4) We show that eGFR equations discovered by our methods demonstrate robust outperformance over existing eGFR equations via a diverse range of prediction and complexity metrics.

979-8-3503-5409-6/24/$31.00 ©2024 IEEE
DOI 10.1109/CAI59869.2024.00245

TABLE I: **Existing Equations for Estimating GFR.** 10 commonly used eGFR equations from MDCalc. When ethnic factor correction is available, we present both versions: (i) with ethnic factor (WEF) (ii) no ethnic factor (NEF). The features used in the equations are $\{age$ in years, $gender$, serum creatinine in mg/dL (*SCR*), serum cystatin C in mg/L (*SCYS*), $height$ in cm$\}$.

Name of Equation	Condition	Equation
MDRD NEF (2006) [11]	female	$129.85 \times SCR^{-1.154} \times age^{-0.203}$
	male	$175 \times SCR^{-1.154} \times age^{-0.203}$
MDRD WEF (2006) [11]	female	$157.37 \times SCR^{-1.154} \times age^{-0.203}$
	male	$212.1 \times SCR^{-1.154} \times age^{-0.203}$
Schwartz Equation (2009) [9]	all	$0.413 \times height/SCR$
CKD-EPI Creatinine NEF (2009) [12]	female, $SCR \leq 0.7$	$128.06 \times SCR^{-0.329} \times 0.993^{age}$
	female, $SCR > 0.7$	$93.559 \times SCR^{-1.209} \times 0.993^{age}$
	male, $SCR \leq 0.9$	$135.02 \times SCR^{-0.411} \times 0.993^{age}$
	male, $SCR > 0.9$	$124.14 \times SCR^{-1.209} \times 0.993^{age}$
CKD-EPI Creatinine WEF (2009) [12]	female, $SCR \leq 0.7$	$148.42 \times SCR^{-0.329} \times 0.993^{age}$
	female, $SCR > 0.7$	$108.43 \times SCR^{-1.209} \times 0.993^{age}$
	male, $SCR \leq 0.9$	$156.49 \times SCR^{-0.411} \times 0.993^{age}$
	male, $SCR > 0.9$	$143.87 \times SCR^{-1.209} \times 0.993^{age}$
CKD-EPI Cystatin C (2012) [13]	female, $SCYS \leq 0.8$	$110.89 \times SCYS^{-0.499} \times 0.996^{age}$
	female, $SCYS > 0.8$	$92.166 \times SCYS^{-1.328} \times 0.996^{age}$
	male, $SCYS \leq 0.8$	$118.98 \times SCYS^{-0.499} \times 0.996^{age}$
	male, $SCYS > 0.8$	$98.89 \times SCYS^{-1.328} \times 0.996^{age}$
CKD-EPI Creatinine-Cystatin C NEF (2012) [13]	female, $SCYS \leq 0.8$, $SCR \leq 0.7$	$109.44 \times SCR^{-0.248} \times SCYS^{-0.375} \times 0.995^{age}$
	female, $SCYS \leq 0.8$, $SCR > 0.7$	$96.495 \times SCR^{-0.601} \times SCYS^{-0.375} \times 0.995^{age}$
	female, $SCYS > 0.8$, $SCR \leq 0.7$	$101.53 \times SCR^{-0.248} \times SCYS^{-0.711} \times 0.995^{age}$
	female, $SCYS > 0.8$, $SCR > 0.7$	$89.524 \times SCR^{-0.601} \times SCYS^{-0.711} \times 0.995^{age}$
	male, $SCYS \leq 0.8$, $SCR \leq 0.9$	$121.48 \times SCR^{-0.207} \times SCYS^{-0.375} \times 0.995^{age}$
	male, $SCYS \leq 0.8$, $SCR > 0.9$	$116.54 \times SCR^{-0.601} \times SCYS^{-0.375} \times 0.995^{age}$
	male, $SCYS > 0.8$, $SCR \leq 0.9$	$112.71 \times SCR^{-0.207} \times SCYS^{-0.711} \times 0.995^{age}$
	male, $SCYS > 0.8$, $SCR > 0.9$	$108.12 \times SCR^{-0.601} \times SCYS^{-0.711} \times 0.995^{age}$
CKD-EPI Creatinine-Cystatin C WEF (2012) [13]	female, $SCYS \leq 0.8$, $SCR \leq 0.7$	$118.19 \times SCR^{-0.248} \times SCYS^{-0.375} \times 0.995^{age}$
	female, $SCYS \leq 0.8$, $SCR > 0.7$	$104.21 \times SCR^{-0.601} \times SCYS^{-0.375} \times 0.995^{age}$
	female, $SCYS > 0.8$, $SCR \leq 0.7$	$109.66 \times SCR^{-0.248} \times SCYS^{-0.711} \times 0.995^{age}$
	female, $SCYS > 0.8$, $SCR > 0.7$	$96.686 \times SCR^{-0.601} \times SCYS^{-0.711} \times 0.995^{age}$
	male, $SCYS \leq 0.8$, $SCR \leq 0.9$	$131.2 \times SCR^{-0.207} \times SCYS^{-0.375} \times 0.995^{age}$
	male, $SCYS \leq 0.8$, $SCR > 0.9$	$125.86 \times SCR^{-0.601} \times SCYS^{-0.375} \times 0.995^{age}$
	male, $SCYS > 0.8$, $SCR \leq 0.9$	$121.72 \times SCR^{-0.207} \times SCYS^{-0.711} \times 0.995^{age}$
	male, $SCYS > 0.8$, $SCR > 0.9$	$116.77 \times SCR^{-0.601} \times SCYS^{-0.711} \times 0.995^{age}$
CKD-EPI Creatinine (2021) [14]	female, $SCR \leq 0.7$	$131.86 \times SCR^{-0.241} \times 0.9938^{age}$
	female, $SCR > 0.7$	$93.667 \times SCR^{-1.2} \times 0.9938^{age}$
	male, $SCR \leq 0.9$	$137.55 \times SCR^{-0.302} \times 0.9938^{age}$
	male, $SCR > 0.9$	$125.13 \times SCR^{-1.2} \times 0.9938^{age}$
CKD-EPI Creatinine-Cystatin C (2021) [14]	female, $SCYS \leq 0.8$, $SCR \leq 0.7$	$111.87 \times SCR^{-0.219} \times SCYS^{-0.323} \times 0.9961^{age}$
	female, $SCYS \leq 0.8$, $SCR > 0.7$	$99.63 \times SCR^{-0.544} \times SCYS^{-0.323} \times 0.9961^{age}$
	female, $SCYS > 0.8$, $SCR \leq 0.7$	$101.07 \times SCR^{-0.219} \times SCYS^{-0.778} \times 0.9961^{age}$
	female, $SCYS > 0.8$, $SCR > 0.7$	$90.011 \times SCR^{-0.544} \times SCYS^{-0.778} \times 0.9961^{age}$
	male, $SCYS \leq 0.8$, $SCR \leq 0.9$	$123.72 \times SCR^{-0.144} \times SCYS^{-0.323} \times 0.9961^{age}$
	male, $SCYS \leq 0.8$, $SCR > 0.9$	$118.61 \times SCR^{-0.544} \times SCYS^{-0.323} \times 0.9961^{age}$
	male, $SCYS > 0.8$, $SCR \leq 0.9$	$111.77 \times SCR^{-0.144} \times SCYS^{-0.778} \times 0.9961^{age}$
	male, $SCYS > 0.8$, $SCR > 0.9$	$107.16 \times SCR^{-0.544} \times SCYS^{-0.778} \times 0.9961^{age}$

II. Related Work

SOTA Symbolic Regression. In the field of SR, genetic programming (GP) has become the most common paradigm to tackle the large search space of equations [1, 2, 15]. Even SOTA SR algorithms incorporate GP at the core for regression [16, 17, 18] and classification [19]. GP-based SR operates by starting with a random population of equations, evaluating them, and modifying these equations (via predefined evolutionary operations such as crossover and mutation) based on their evaluation scores. In this work, we chose deep symbolic regression (DSR) [16], neural-guided genetic programming (NGGP) [17] and DistilSR [20] as our choice of SOTA SR methods based on recent SR benchmark performances in terms of prediction and complexity on SRBench [21].

Existing eGFR Equations. In this work, we selected all relevant eGFR equations from MDCalc, a well-known medical reference for medical equations [22, 23], used by millions of medical professionals globally (over 200 countries), inclusive of more than 65% of US physicians. The equations are made available in Table I, consisting of MDRD [11], Schwartz

equation [9], CKD-EPI Creatinine [12], CKD-EPI Cystatin and CKD-EPI Creatinine-Cystain C [13, 14]. When ethnic factor correction is available in the equation, we present both versions: (i) with ethnic factor (WEF) (ii) no ethnic factor (NEF). Note that the ethnic factor corrections only cater to a select few ethnic groups, which is one of the main weaknesses of current eGFR equations our SR methods addresses (via automated discovery of equations on new datasets).

III. Methodology

Dataset Details. In our work, we use a public dataset of measured GFR (mGFR) of Congolese adults [8]. This dataset obtained mGFR by measuring plasma clearance of iohexol. Other features include $\{age$ in years, $gender$, serum creatinine in mg/dL (*SCR*), serum cystatin C in mg/L (*SCYS*), $height$ in cm$\}$, which are present in equations in Table I.

Algorithm Details. For all SR algorithms, we set the functions in the primitive symbols set to $\{+, -, \times, /, \wedge\}$. The operands (features) in the primitive symbols set differ based on the algorithm, which we discuss below. The other hyperparameters are tuned based on mean squared error of the estimated GFR on a validation set against the mGFR. In this paper, we use a variety of SR algorithms, broadly categorized as: (i) existing SOTA SR methods, (ii) modification of an existing SR method and (iii) our proposed new SR methods.

(i) Existing SR: We use the 3 SOTA SR algorithms identified in related works, DSR, NGGP and DistilSR, on all features of the dataset. The top equation from each algorithm is registered as a new discovered eGFR equation.

(ii) Modified Existing SR: We note that in the existing eGFR equations (see Table I), most consist of conditions, which may be preferred by healthcare professionals. It is difficult for current SR algorithms to implicitly learn the structure of these conditions implicitly. Thus, we explicitly modify SR to use conditions, which we call conditioned-SR. In our work, we selected the conditions based on gender (i.e., the first condition is for females and the second condition is for males). For each condition, we run DistilSR on the dataset with all features and we pick the equations which perform well in MSE for each condition. However, to be consistent with medical eGFR equations, the final equations selected need to share the same equation structure across the multiple conditions, inspired by multi-level SR [18]. We picked DistilSR as the SR method to modify since in DistilSR, the search space of equations is deterministic, unlike DSR and NGGP.

(iii) Our Proposed New SR: In the methods above, the equations found are usually drastically different from existing eGFR equations and do not use any information from the existing eGFR equations. To this end, we introduce our novel method of taking existing GFR medical estimation equations as prior knowledge and evolving them via Constrained Genetic Perturbation (CGP), as outlined in Algorithm 1. In CGP, we utilize K-Expressions from Gene Expression Programming [24] in order to have a easily manipulated representation of an equation. As an example, the K-Expression "$* + -abcde$" represents $(a + b) * (c - d)$. K-Expressions have other properties which are desirable, such as variable length equations represented by fixed length strings and fulfilling the closure property without much overhead [24, 20]. In CGP, an existing eGFR equation is first converted to its K-Expression form. The K-Expression is then perturbed at 2 points, as shown in Step 4 to 10 of Algorithm 1. The K-Expression is then converted back into an equation, with numerical constants obtained by BFGS optimizer [25], and evaluated for the mean squared error of its estimation of GFR against the mGFR. In CGP, the prior knowledge used drastically reduces the search space since only equations 2 genetic perturbations away from the original equation are evaluated. Another benefit is that it also discovers equations that closely resemble the functional forms familiar to medical professionals if the input equation is already a widely accepted medical equation. Note that in CGP, we use a $primitive\_symbol\_set$ that only includes features that are in the original equations. Finally, we also introduce a variant, extraCGP, that enables the equation to be modified with all features by extending the $primitive\_symbol\_set$ to include all features. This is motivated by current research in eGFR equations, where researchers manually modify well-known equations by adding new features, such as the inclusion of *SCYS* which was previously omitted in CKD-EPI equations.

Key Differences Between Algorithms. In DSR, NGGP and DistilSR, it is hard to generate equations with conditions,

Algorithm 1: CGP Pseudo Code

Input: $original\_equation$, $primitive\_symbols\_set$, $\mathbf{X}$, $\mathbf{y}$, where $\mathbf{X}$ is the features and $\mathbf{y}$ is the mGFR
Output: $best\_modified\_equation$

1 $best\_score \leftarrow null$
2 $best\_modified\_equation \leftarrow null$
3 $original\_K\_exp \leftarrow$ ConvertToKExpression($original\_equation$)
4 **for** $i \in \{1, 2, \cdots, len(original\_K\_exp)\}$ **do**
5 **for** $j \in \{(i+1), (i+2), \cdots, len(original\_K\_exp)\}$ **do**
6 **for** $\alpha \in primitive\_symbols\_set$ **do**
7 **for** $\beta \in primitive\_symbols\_set$ **do**
8 $modified\_K\_exp \leftarrow original\_K\_exp$
9 $modified\_K\_exp[i] \leftarrow \alpha$
10 $modified\_K\_exp[j] \leftarrow \beta$
11 $modified\_K\_exp \leftarrow$ AppendPadding($modified\_K\_exp$) /* Padding ensures decoding to a valid equation */
12 $modified\_equation \leftarrow$ Decode($modified\_K\_exp$)
13 $modified\_equation \leftarrow$ BFGS($modified\_equation, \mathbf{X}, \mathbf{y}$) /* BFGS is a method for optimizing numerical constants */
14 $current\_score \leftarrow$ MSE($modified\_equation, \mathbf{X}, \mathbf{y}$)
15 **if** $best\_score > current\_score$ **then**
16 $best\_score \leftarrow current\_score$
17 $best\_modified\_equation \leftarrow modified\_equation$
18 **end**
19 **end**
20 **end**
21 **end**
22 **end**
23 **return** $best\_modified\_equation$

TABLE II: **New Equations (Our Contributions) for estimating GFR.**

SR Method Used	Name of Equation	Condition	Our Discovered Equation
DSR [16]	SR-1	all	$height/(SCR + SCYS^2/SCR)$
NGGP [17]	SR-2	all	$age \times height/(age \times SCYS^2 + age + 3 \times is\_female - SCYS)$
DistilSR [20]	SR-3	all	$height^{0.8677^{SCYS}}$
DistilSR [20] + Our modification	Conditioned-SR	female	$(SCYS + 72.333) \times SCYS^{-0.908}$
		male	$(SCYS + 89.42) \times SCYS^{-0.5543}$
Ours	CGP-1	female, $SCYS \leq 0.8$, $SCR \leq 0.7$	$243.41 \times 0.41121^{SCYS} \times 0.76392^{SCR}$
		female, $SCYS \leq 0.8$, $SCR > 0.7$	$251.56 \times 0.72774^{SCYS} \times 0.39669^{SCR}$
		female, $SCYS > 0.8$, $SCR \leq 0.7$	$102.43 \times 0.81339^{SCYS} \times 0.97084^{SCR}$
		female, $SCYS > 0.8$, $SCR > 0.7$	$149.82 \times 0.44582^{SCYS} \times 1.1073^{SCR}$
		male, $SCYS \leq 0.8$, $SCR \leq 0.9$	$17.063 \times 2.3259^{SCYS} \times 4.7351^{SCR}$
		male, $SCYS \leq 0.8$, $SCR > 0.9$	$3468.7 \times 0.64898^{SCYS} \times 0.034174^{SCR}$
		male, $SCYS > 0.8$, $SCR \leq 0.9$	$95.012 \times 0.73137^{SCYS} \times 1.4526^{SCR}$
		male, $SCYS > 0.8$, $SCR > 0.9$	$167.21 \times 0.66438^{SCYS} \times 0.81168^{SCR}$
Ours	CGP-2	female, $SCR \leq 0.7$	$4.4617 \times SCR^{4.193} \times (24.678 - age) + 110.34$
		female, $SCR > 0.7$	$0.49976 \times SCR^{-1.9447} \times (58.564 - age) + 72.42$
		male, $SCR \leq 0.9$	$2.2166 \times SCR^{7.555} \times (71.968 - age) + 84.107$
		male, $SCR > 0.9$	$0.64289 \times SCR^{-4.251} \times (51.547 - age) + 86.198$
Ours	CGP-3	female, $SCYS \leq 0.8$, $SCR \leq 0.7$	$SCR^{-0.23201} \times (92.956 \times SCYS^{-0.44557} - 0.30801 \times age)$
		female, $SCYS \leq 0.8$, $SCR > 0.7$	$SCR^{-0.40425} \times (79.963 \times SCYS^{-0.67882} - 0.32564 \times age)$
		female, $SCYS > 0.8$, $SCR \leq 0.7$	$SCR^{0.0115} \times (86.741 \times SCYS^{-0.097012} - 0.070911 \times age)$
		female, $SCYS > 0.8$, $SCR > 0.7$	$SCR^{0.18054} \times (112.28 \times SCYS^{-0.29257} - 0.68316 \times age)$
		male, $SCYS \leq 0.8$, $SCR \leq 0.9$	$SCR^{0.65914} \times (2.0637 \times SCYS^{1.3969} + 4.4278 \times age)$
		male, $SCYS \leq 0.8$, $SCR > 0.9$	$SCR^{-11.75} \times (0.14873 \times SCYS^{-13.426} + 1.3436 \times age)$
		male, $SCYS > 0.8$, $SCR \leq 0.9$	$SCR^{0.21501} \times (114.67 \times SCYS^{-0.31784} - 0.33236 \times age)$
		male, $SCYS > 0.8$, $SCR > 0.9$	$SCR^{-0.068742} \times (108.45 \times SCYS^{-0.12809} - 0.37943 \times age)$
Ours	extraCGP-1		Obtained Same Equations as CGP-1
Ours	extraCGP-2	female, $SCR \leq 0.7$	$298.76 \times age^{-0.11564} \times 0.40507^{SCYS}$
		female, $SCR > 0.7$	$300.3 \times age^{-0.21147} \times 0.5624^{SCYS}$
		male, $SCR \leq 0.9$	$273.6 \times age^{-0.17364} \times 0.6579^{SCYS}$
		male, $SCR > 0.9$	$242.21 \times age^{-0.17131} \times 0.71114^{SCYS}$
Ours	extraCGP-3		Obtained Same Equations as CGP-3

whereas in conditioned-SR the conditions are explicitly defined and in CGP, the conditions follow that of the original equation. In DSR and NGGP, the search space of equations is large and a stochastic search is done, whereas DistilSR, conditioned-SR and CGP does an exhaustive and deterministic search. In DSR, NGGP, DistilSR and conditioned-SR, the search space of equations are large and do not incorporate the current equations, whereas CGP drastically reduces the search space to K-Expressions that are 2 genetic perturbations away and also discovers equations that closely resemble the functional forms familiar to medical professionals.

Evaluation Details. We measure the performance of current eGFR equations and our discovered eGFR equation via a diverse range of prediction and complexity metrics. These experiments are done with a 70-30 training-test split. In-line with GFR literature, we chose the following 6 prediction metrics (where y is the mGFR, $\hat{y}$ is the eGFR, ρ is the pearson's correlation between mGFR and the eGFR, μ is the mean and σ is the standard deviation):

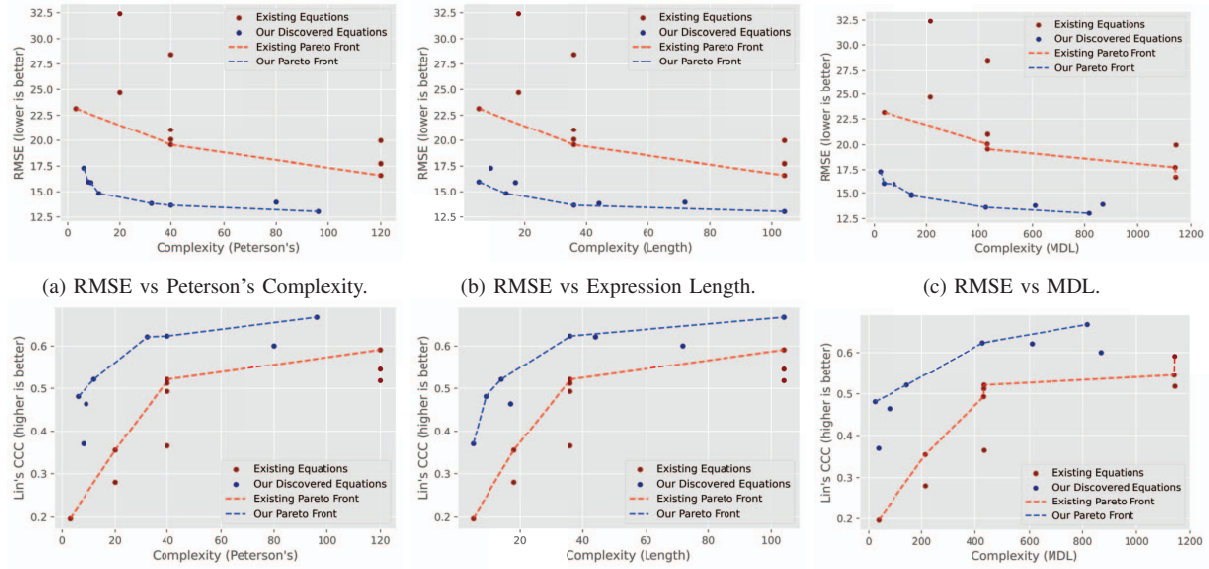
i) Root-mean-squared error (RMSE), $\sqrt{\left(\sum_{i=1}^{N}(y_i - \hat{y}_i)^2/N\right)}$.
ii) Mean absolute error (MAE), $\sum_{i=1}^{N}|y_i - \hat{y}_i|/N$.
iii) Lin's concordance correlation coefficient (CCC) [26], $2\rho\sigma_y\sigma_{\hat{y}}/((\mu_y - \mu_{\hat{y}})^2 + \sigma_y^2 + \sigma_{\hat{y}}^2)$.
iv) Proportion of estimates within $\pm 10\%$ of mGFR (P10) [8].
v) Proportion of estimates within $\pm 30\%$ of mGFR (P30) [8].
vi) Stage accuracy, the agreement of eGFR with mGFR in terms of categorizing individuals into the 5 guideline-recommended GFR stages (Stage 1: > 90, Stage 2: 60 to 89, Stage 3: 30 to 59, Stage 4: 15 to 29, Stage 5: < 15) [14].

In terms of complexity, we chose the following 3 metrics from SR literature:

i) Peterson's complexity, which is the sum of pre-defined scores assigned to equation tokens as detailed in [16].
ii) Equation length, which is the sum of occurrences of operations, constants and features in the equation [21].
iii) Minimum de*SCR*iption length (MDL), where an equation with features $\mathbf{x}$ and numerical constants, $\mathbf{p}$, given in the form of $f(\mathbf{x};\mathbf{p})$ has an MDL of $L_d(\mathbf{p}) + k\log_2 n$; k is the number of times the n basis functions appear [27]. For real numbers, $L_d(\mathbf{p})$ is given by $\frac{1}{2}\sum_i \log_2\left(1 + \left(\frac{p_i}{\varepsilon}\right)^2\right)$. In particular, we follow the implementation in [27], in which the precision floor, ϵ, is set to 2^{-30}.

TABLE III: Prediction performance analysis of all existing and new equations. Best performances are bolded.

Name of Equation	RMSE (lower is better)	MAE (lower is better)	Lin's CCC (higher is better)	P10 (higher is better)	P30 (higher is better)	Stage Accuracy (higher is better)
Existing Equations						
MDRD NEF (2006) [11]	24.704	16.04	0.355	0.387	0.836	0.581
MDRD WEF (2006) [11]	32.394	20.496	0.28	0.306	0.765	0.653
Schwartz Equation (2009) [9]	23.144	18.579	0.196	0.244	0.785	0.418
CKD-EPI Creatinine NEF (2009) [12]	19.564	13.735	0.512	0.5	0.806	0.663
CKD-EPI Creatinine WEF (2009) [12]	28.45	22.536	0.365	0.204	0.704	0.653
CKD-EPI Cystatin C (2012) [13]	21.042	15.832	0.521	0.428	0.836	0.581
CKD-EPI Creatinine-Cystatin C NEF (2012) [13]	16.58	11.825	0.591	0.5	0.908	0.663
CKD-EPI Creatinine-Cystatin C WEF (2012) [13]	19.975	15.409	0.518	0.367	0.836	0.653
CKD-EPI Creatinine (2021) [14]	20.083	14.416	0.493	0.397	0.806	0.673
CKD-EPI Creatinine-Cystatin C (2021) [14]	17.726	13.278	0.545	0.418	0.867	0.663
Our Discovered Equations (New Contributions)						
SR-1	17.294	13.217	0.481	0.438	0.908	0.591
SR-2	15.893	12.086	0.464	0.469	0.938	0.622
SR-3	15.95	12.235	0.37	0.438	0.938	0.612
Conditioned-SR	14.826	10.985	0.521	0.53	0.938	0.693
CGP-1/extraCGP-1	13.922	9.7528	0.6	0.571	0.938	0.734
CGP-2	13.808	10.222	0.621	0.561	**0.948**	0.683
CGP-3/extraCGP-3	**13.025**	**8.711**	**0.667**	**0.632**	**0.948**	0.755
extraCGP-2	13.63	9.4247	0.623	0.622	**0.948**	**0.775**



(a) RMSE vs Peterson's Complexity. (b) RMSE vs Expression Length. (c) RMSE vs MDL.

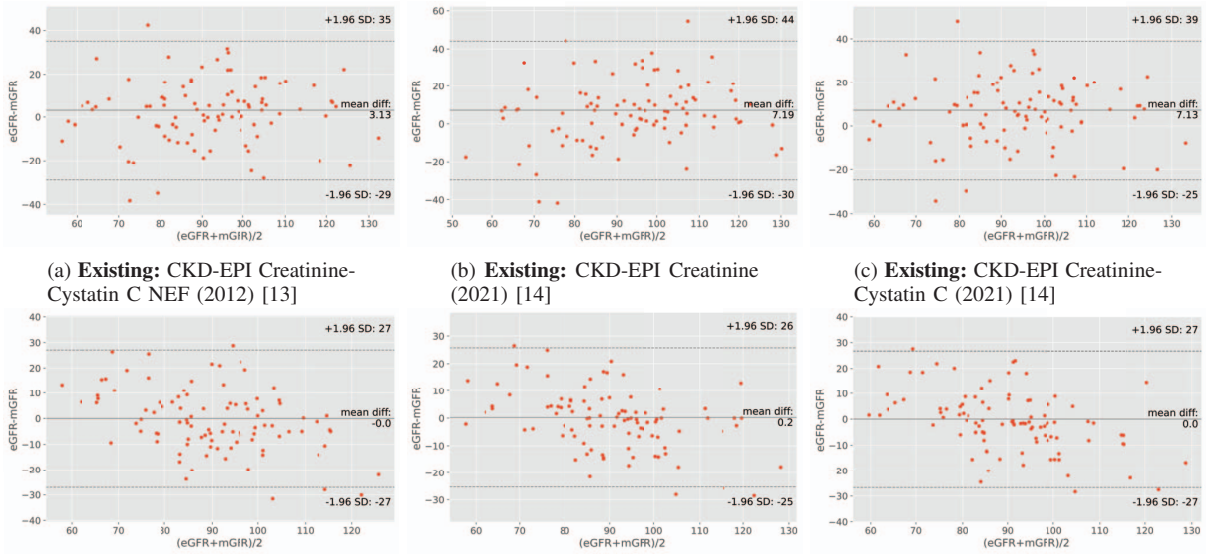
(d) Lin's CCC vs Peterson's Complexity. (e) Lin's CCC vs Expression Length. (f) Lin's CCC vs MDL.

Fig. 1: Plots of various prediction metrics against various complexity metrics.

IV. Results and Discussion

In Table II, we document the new eGFR equations we generated using SR. 8 new equations were discovered in total. SR-1, SR-2, SR-3 were found using existing SOTA SR methods. Conditioned-SR was found using DistilSR with our modifications. CGP-1, CGP-2, CGP-3, extraCGP-1, extraCGP-2 and extraCGP-3 were found using our new CGP algorithm (see Algorithm 1) which takes in an equation and modifies it. In particular, CGP-1 and extraCGP-1 were modified from 'CKD-EPI Creatinine WEF (2012)'. CGP-2 and extraCGP-2 were modified from 'CKD-EPI Creatinine (2021)'. CGP-3 and extraCGP-3 were modified from 'CKD-EPI Creatinine-Cystatin C (2021)'. The difference between CGP and extraCGP is that CGP uses only features provided by the original equation, whereas extraCGP uses all features in the dataset. Note that CGP-1 and extraCGP-1 are the same, suggesting that the equation with genetic perturbations including the extra features did not result in better performance. The same applies to CGP-3 and extraCGP-3.

Do our new eGFR equations from SR provide better estimates than current equations? We record the prediction performances of the equations in Table III. In all 6 prediction performance metrics, the best equations were our discovered equations using SR. Notably, CGP-3/extraCGP-3 performed the best in 5 out of 6 metrics, coming second in the remaining



(a) **Existing:** CKD-EPI Creatinine-Cystatin C NEF (2012) [13]

(b) **Existing:** CKD-EPI Creatinine (2021) [14]

(c) **Existing:** CKD-EPI Creatinine-Cystatin C (2021) [14]

(d) **New (Ours):** CGP-2

(e) **New (Ours):** CGP-3/extraCGP-3

(f) **New (Ours):** extraCGP-2

Fig. 2: Bland-Altman analysis of 3 top existing eGFR equations and 3 top eGFR equations we discover.

metric. It should also be noted that CGP-1/extraCGP-1, CGP-2, CGP-3/extraCGP-3 and extraCGP-2 are superior to all existing equations on all 6 performance metrics.

How does the prediction-complexity tradeoff of the new eGFR equations compare to existing equations? Although our methods perform best in terms of pure prediction performance as discussed above, it is important to verify that this is not at the expense of higher complexity (which translates to less explainability and interpretability). We can observe qualitatively that our discovered eGFR equations in Table II are simple and similar in complexity to existing equations. Quantitative-wise, we use 3 widely accepted complexity metrics. In Fig. 1, we plot various prediction metrics against various complexity metrics. In these 6 plots, our discovered eGFR equations pareto-dominates existing eGFR equations. In other words, for each of our top eGFR equations, there is no existing eGFR equation which has both better prediction and lower complexity.

Do the new eGFR equations diagnose better compared to existing equations? On the macro level, as seen in Table III, the stage accuracy of our discovered equations show the best performance. To analyze the micro level, we perform a Bland-Altman analysis for 3 top existing eGFR equations (see Fig. 2a,2b,2c) and 3 top new eGFR equations we discovered (see Fig. 2d,2e,2f). The Bland-Altman analysis is commonly used in healthcare to evaluate the agreement between a estimate and a measured value [8]. The Bland-Altman plots in Fig. 2 demonstrate that our eGFR equations perform better than existing equations. In the existing eGFR equations, we see that the mean difference is positive, suggesting that existing eGFR equations tend to overestimate the mGFR, which may result in missed diagnosis of a weak kidney (low GFR). On the other hand, our discovered equations have a near 0 mean difference. Additionally, the standard deviation of the difference (eGFR-mGFR) for existing equations are larger than our new equations. Furthermore, despite the smaller standard deviation of our equations' eGFR, more of the differences lies within ±1.96 standard deviation (95% limits of agreement).

Limitations: Our work is based upon retrospective data and hence future work using on a prospective cohort is required to provide further validation for the new equations.

V. Conclusion

In this paper, we utilize and demonstrate a variety of ways for SR to discover medical eGFR equations, inclusive of methods using SOTA SR and our own proposed SR innovations. First, using SOTA SR methods, we discover eGFR equations that achieve better estimation performance than existing equations. Then, we introduce our novel method, Constrained Genetic Perturbation, which takes existing eGFR equations and modifies them to achieve better performance. Prior knowledge, in the form of the existing equation, drastically reduces the search space and also discovers equations that closely resemble the functional forms familiar to medical professionals. We then introduce a variant that also enables the equation to be modified with newly obtained features, which produced the equation extraCGP-2, the best equation in terms of diagnosing the kidney stage. Finally, we show that equations discovered by our methods demonstrate the best performance among all equations over a large range of prediction and complexity metrics. Future work could explore expanding the primitive set. We hope that our work will motivate healthcare professionals to use SR following our methodology to discover new medical equations in an automated and data-driven way.

ACKNOWLEDGMENT

This research/project is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG3-PhD-2023-08-052T), and A*STAR, CISCO Systems (USA) Pte. Ltd and National University of Singapore under its Cisco-NUS Accelerated Digital Economy Corporate Laboratory (Award I21001E0002).

REFERENCES

[1] J. R. Koza, "Genetic Programming. On the Programming of Computers by Means of Natural Selection," *Complex Adaptive Systems*, 1992.

[2] K. S. Fong, S. Wongso, and M. Motani, "Rethinking Symbolic Regression: Morphology and Adaptability in the Context of Evolutionary Algorithms," in *The Eleventh International Conference on Learning Representations*, 2022.

[3] S.-M. Udrescu and M. Tegmark, "AI Feynman: A Physics-Inspired Method for Symbolic Regression," *Science Advances*, vol. 6, no. 16, p. eaay2631, 2020.

[4] Y. Wang, N. Wagner, and J. M. Rondinelli, "Symbolic Regression in Materials Science," *MRS Communications*, vol. 9, no. 3, pp. 793–805, 2019.

[5] J. Martinez-Gil and J. M. Chaves-Gonzalez, "A Novel Method Based on Symbolic Regression for Interpretable Semantic Similarity Measurement," *Expert Systems with Applications*, vol. 160, p. 113663, 2020.

[6] N. J. Christensen, S. Demharter, M. Machado *et al.*, "Identifying Interactions in Omics Data for Clinical Biomarker Discovery using Symbolic Regression," *Bioinformatics*, vol. 38, no. 15, pp. 3749–3758, 2022.

[7] A. S. Levey and L. A. Inker, "GFR as the "Gold Standard": Estimated, Measured, and True," *American Journal of Kidney Diseases*, vol. 67, no. 1, pp. 9–12, 2016.

[8] J. B. Bukabau, E. K. Sumaili, E. Cavalier *et al.*, "Performance of Glomerular Filtration Rate Estimation Equations in Congolese Healthy Adults: the Inopportunity of the Ethnic Correction," *PloS one*, vol. 13, no. 3, p. e0193384, 2018.

[9] G. J. Schwartz, A. Mun, M. F. Schneider *et al.*, "New Equations to Estimate GFR in Children with CKD," *Journal of the American Society of Nephrology*, vol. 20, no. 3, pp. 629–637, 2009.

[10] M. A. Hossain, H. Elmoselhi, A. A. Elshorbagy, and A. Shoker, "The Sask Formula to Estimate Glomerular Filtration Rate in Renal Transplant Patients," *Nephron Clinical Practice*, vol. 117, no. 2, pp. c135–c150, 2011.

[11] A. S. Levey, J. Coresh, T. Greene *et al.*, "Using Standardized Serum Creatinine Values in the Modification of Diet in Renal Disease Study Equation for Estimating Glomerular Filtration Rate," *Annals of Internal Medicine*, vol. 145, no. 4, pp. 247–254, 2006.

[12] A. S. Levey, L. A. Stevens, C. H. Schmid *et al.*, "A New Equation to Estimate Glomerular Filtration Rate," *Annals of Internal Medicine*, vol. 150, no. 9, pp. 604–612, 2009.

[13] L. A. Inker, C. H. Schmid, H. Tighiouart *et al.*, "Estimating Glomerular Filtration Rate from Serum Creatinine and Cystatin c," *New England Journal of Medicine*, vol. 367, no. 1, pp. 20–29, 2012.

[14] L. A. Inker, N. D. Eneanya, J. Coresh *et al.*, "New Creatinine and Cystatin C–Based Equations to Estimate GFR without Race," *New England Journal of Medicine*, vol. 385, no. 19, pp. 1737–1749, 2021.

[15] M. Schmidt and H. Lipson, "Distilling Free-Form Natural Laws from Experimental Data," *Science*, vol. 324, no. 5923, pp. 81–85, 2009.

[16] B. K. Petersen, M. L. Larma, T. N. Mundhenk *et al.*, "Deep Symbolic Regression: Recovering Mathematical Expressions from Data via Risk-Seeking Policy Gradients," *The International Conference on Learning Representations*, 2019.

[17] T. N. Mundhenk, M. Landajuela, R. Glatt *et al.*, "Symbolic Regression via Neural-Guided Genetic Programming Population Seeding," in *Advances in Neural Information Processing Systems*, 2021.

[18] K. S. Fong and M. Motani, "Multi-Level Symbolic Regression: Function Structure Learning for Multi-Level Data," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2024, pp. 2890–2898.

[19] K. S. Fong and M. Motani, "Symbolic Regression Enhanced Decision Trees for Classification Tasks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 11, 2024, pp. 12 033–12 042.

[20] K. S. Fong and M. Motani, "DistilSR: A Distilled Version of Gene Expression Programming Symbolic Regression," in *Proceedings of the Companion Conference on Genetic and Evolutionary Computation*, 2023, pp. 567–570.

[21] W. La Cava, P. Orzechowski, B. Burlacu *et al.*, "Contemporary Symbolic Regression Methods and their Relative Performance," *Neurips Track on Datasets and Benchmarks.*, 2021.

[22] A. Elovic and A. Pourmand, "MDCalc Medical Calculator App Review," *Journal of Digital Imaging*, vol. 32, pp. 682–684, 2019.

[23] N. Soleimanpour and M. Bann, "Clinical Risk Calculators Informing the Decision to Admit: A Methodologic Evaluation and Assessment of Applicability," *Plos one*, vol. 17, no. 12, p. e0279294, 2022.

[24] C. Ferreira, "Gene Expression Programming in Problem Solving," *Soft Computing and Industry: Recent Applications*, pp. 635–653, 2000.

[25] C. G. Broyden, "The Convergence of a Class of Double-Rank Minimization Algorithms," *IMA Journal of Applied Mathematics*, vol. 6, no. 1, pp. 76–90, 1970.

[26] I. Lawrence and K. Lin, "A Concordance Correlation Coefficient to Evaluate Reproducibility," *Biometrics*, pp. 255–268, 1989.

[27] S.-M. Udrescu, A. Tan, J. Feng *et al.*, "AI Feynman 2.0: Pareto-Optimal Symbolic Regression Exploiting Graph Modularity," *Advances in Neural Information Processing Systems*, vol. 33, pp. 4860–4871, 2020.