Università degli studi di Modena e Reggio Emilia

Dipartimento di Scienze Fisiche, Informatiche e Matematiche

Corso di Laurea Magistrale in Matematica

From Grades to Recovery: A New Way to Quantify Intrinsic Capacity

A Theoretically Grounded Approach to Clinical Index Construction Using Social Ranking and Symbolic Machine Learning

Relatore:
Prof.ssa Federica Mandreoli

Candidato:
Chiara Garofano

Correlatore:
Ph.D. Veronica Guidetti

Anno Accademico 2024/2025

Abstract

The concept of Intrinsic Capacity (IC), introduced by the World Health Organization (WHO), reflects the composite of an individual's physical and mental capacities and plays a crucial role in assessing functional health in older adults. However, due to its multidimensional nature, there is currently no standardized method to quantitatively assess IC using real-world clinical data.

This thesis proposes a novel, interpretable, and data-driven framework to estimate, rank, and model Intrinsic Capacity from a set of 14 validated clinical questionnaries. While these scores are universal and applicable to any population, our goal is to develop a tool, a direct formula, that enables the construction of an Intrinsic Capacity index tailored to a specific population. This tool should be suitable for use in monitoring contexts and based on the clinical characteristics relevant to that particular group; in our case, individuals affected by Long COVID. As a first step, we aim to construct a single rank derived from the IC features. To achieve this, we need to aggregate 14 distinct variables into a unified score, which will also enable a ranking.

The Majority Judgment (MJ) method is then applied to generate a majority grade for each individual, resulting in a population-wide ranking of IC. As a social ranking method, MJ aggregates the judgments of multiple evaluators to produce a global ranking of individuals. Specifically, it takes as input, for each individual (candidate), the 14 values assigned by the IC features, which are interpreted as scores or votes given by 14 judges, each representing a distinct IC dimension. For this process to be meaningful and interpretable, these scores must be both comparable and ordinal. This necessitates transforming the original continuous variables into five ordinal categories, ranging from very low to very high, using an optimized binning strategy designed to preserve variance and inter-patient diversity. The consistency of the ranking is then

validated by its correlation with the clinical outcome healed.

To develop a more tailored index for the patients under consideration and thereby enhance the clinical applicability of the model, symbolic regression, a symbolic machine learning technique aimed at discovering mathematical expressions that best reproduces the IC ranking, is performed not on the original 14 IC features, but on a new set of more specific-purpose covariates.

This approach yields an interpretable and portable formula based on covariates whose define a precise population: Long COVID's patients. The expression output closely reflects the IC grade obtained in the first phase, providing a clinically meaningful tool to characterize the IC ranking of this defined population. In addition, it reliably predicts recovery outcomes.

Contents

Introduction

The progressive aging of the global population presents an urgent need to redefine how health is measured, monitored and supported in adults. Traditional clinical assessments often focus on disease-centric metrics, which, although essential, fail to capture the broader functional capabilities of individuals. In response to this gap, the World Health Organization (WHO) introduced the concept of Intrinsic Capacity (IC) (John R. Beard et al. [2016]), defined as the composite of all the physical and mental capacities a person can draw upon. This concept promotes a shift from a disease-oriented model to a function-oriented approach to healthy aging.

Despite its importance, Intrinsic Capacity remains a largely theoretical construct due to the lack of standardized tools for its measurement in clinical practice. IC is inherently multidimensional, involving physiological, cognitive and psychological components, making its quantification a complex task. To address this challenge, this thesis proposes a data-driven framework that aims to estimate IC from clinical data, provide a meaningful ranking across individuals and, ultimately, construct a model that can be used in broader healthcare settings.

The study relies on a medical dataset comprising 14 features (validated clinical questionnaires) that reflect various domains related to IC: they are crucial for understanding the structure and variability of IC. Thus, these features are used as a foundation to analyze IC in depth and define an initial scoring mechanism. We started from a set of questionnaires that describe the IC 5 health domains including locomotion, sensory, vitality, cognitive and psychosocial, namely the Depression Anxiety Stress Scales (DASS), EuroQol 5-Dimension 5-Level (EQ-5D-5L), Insomnia Severity Index (ISI), CD-RISC (Connor-Davidson Resilience Scale), and Short Form Health Survey with 36 items (SF-36). Together, the scores from these tools provide a multi-

dimensional representation of an individual's IC, aligning with the WHO framework. The Majority Judgment (MJ) method is applied to aggregate the values of these 14 features into a single majority grade, representing each individual's overall level of Intrinsic Capacity. Originally proposed for voting systems, this method offers an innovative way to summarize multidimensional inputs into a robust and ordinal measure. In particular, we use the questionnaires as judges and their values as votes for constructing a vote matrix where the candidates are the patients. In our case votes are not ordinal but continuos, to fix this problem, make Majority Judgment applicable and Intrinsic Capacity a more readable and unambiguous scale, we need to render the votes categorical which means transforming the 14 continuous clinical variables into five ordered categories: very low, low, medium, high, and very high, using an optimized binning process. This transformation aims to preserve the discriminative power of the original features by minimizing intra-bin variance and promoting diversity across bins. This categorization enhances interpretability, enabling the application of decision-theoretic methods such as Majority Judgment, and reduces the rumor arising from the features, aligning with the questionnaires' goal: stratify the population in different levels of well being.

The resulting majority grades are eventually used to rank patients according to their IC, offering a transparent and interpretable comparison across the population. Acquired this relative ranking, which is obtained by universal variables such as validated questionnaires, we now want to construct an absolute index, specifically designed for a target population, that can measure Intrinsic Capacity. To achieve this, we collect variables typical of the population of interest (in our case, COVID-19), including demographics, anthropometric characteristics, lifestyle factors, and COVID-related symptoms. We then investigate whether these variables can explain the ranking derived from the IC constructed through Majority Judgment. For this task, a symbolic regression model is introduced. Unlike traditional machine learning models, symbolic regression produces human-readable mathematical expressions that best approximate the target of interest, in our case the IC rank. The model was trained with the set of covariates COVID-related and generates a formula that closely approximate the IC score. Lastly, to validate this newly derived index, we use it as input to a logistic regression model to predict the healing from Long COVID.

Chapter 1

Methods

1.1 Majority Judgment

The Majority Judgment framework offers a principled and robust approach for aggregating ordinal evaluations into a collective ranking. Unlike traditional scoring systems or pairwise comparisons, MJ allows each subject to be independently evaluated across a set of clinical features related, in this case, to IC, using a shared and totally ordered scale. These evaluations are then summarized through a central statistic known as the *majority grade*, which reflects the consensus among the IC-derived features. This grade serves as the foundation for ranking patients, while ensuring key properties from social choice theory such as anonymity, neutrality, and monotonicity are respected.

1.1.1 Majority Ranking

Social choice theory provides formal tools to combine individual evaluations into a coherent collective decision. In our setting, we consider a set of candidates (e.g., patients)

$$C = \{C_1, C_2, \ldots, C_m\},$$

evaluated by a set of judges: clinical features representing aspects of intrinsic capacity

$$J = \{J_1, J_2, \ldots, J_n\},$$

using a common grading scale

$$\Lambda = \{\alpha, \beta, \gamma, \ldots\},$$

which is assumed to be totally ordered (e.g., from "very poor" to "excellent" or, numerically speaking, from 1 to 5).

Each feature J_j assigns a grade α_{ij} to each subject C_i, forming an evaluation matrix Φ, where

$$\Phi(i, j) = \alpha_{ij}.$$

The goal is to derive a ranking over subjects that satisfies the following desirable properties:

- *Anonymity*: The final ranking should be unaffected by the identity or ordering of the features.
- *Neutrality*: Renaming or reordering subjects should not change their ranking.
- *Transitivity*: If subject A is ranked above B, and B above C, then A must be ranked above C.
- *Independence of Irrelevant Alternatives*: The relative ranking between two subjects should not be influenced by the inclusion or exclusion of others.

In this work, we adopt the Majority Judgment framework introduced by Balinski and Laraki [2010]. Here, each subject is evaluated based on IC-derived features using a shared ordinal scale, resulting in a judgment matrix Φ. The aggregation function $F : \Lambda^{m \times n} \to \Lambda^m$ processes this matrix and assigns a final grade F_i to each subject C_i.

The core of MJ is the *majority grade*, defined as the central grade in the ordered list of evaluations for a given subject:

$$F_i^{MJ} = \begin{cases} r_{(n+1)/2} & \text{if } n \text{ is odd,} \\ r_{(n+2)/2} & \text{if } n \text{ is even,} \end{cases}$$

where $r = (r_1, \ldots, r_n)$ is the vector of grades for C_i, sorted in non-increasing order. This majority grade is the grade that a majority of features assign at least as high a score to, and no majority assigns a lower score. As such, it captures the central consensus in the most faithful way possible, honoring stronger agreement among features with a more favorable final grade.

While the majority grade provides a stable summary, situations may arise where multiple subjects share the same majority grade. In such cases, MJ introduces a tie-breaking mechanism based on the *majority value*, a sequence that reflects the full structure of agreement beyond the central grade.

To construct a subject's majority value, we begin with the ordered list of grades assigned by the IC features. At each step, we identify the central grade(s) of the remaining list and select the lower of the two if the list has even length, or the central grade if it has odd length. This selected grade is then removed, and the process is repeated on the reduced list. This procedure continues until all grades have been extracted. The resulting sequence fans outward from the center and reflects how the evaluations deviate from consensus, in a way that prioritizes central agreement while incorporating the full judgment distribution.

Once majority values are computed, subjects can be ranked using the *majority ranking* relation, $\succ_{MJ}$. Given two subjects, C_1 and C_2, with majority values $r_v^{C_1}$ and $r_v^{C_2}$, we say $C_1 \succ_{MJ} C_2$ if and only if $r_v^{C_1} \succ_{lexi} r_v^{C_2}$, where $\succ_{lexi}$ denotes the lexicographic order on sequences. In this way, MJ refines the collective ranking by fairly resolving ties using the entire distribution of evaluations.

A full pipeline on a toy esample is illustrated in Figure 1.1: from raw IC-derived feature evaluations, to sorting and identification of majority grades, to tie-breaking via majority values, and ultimately the final ranking of subjects.

	J_1	J_2	J_3
C_1	1	3	2
C_2	4	2	3
C_3	2	1	1
C_4	3	3	4

→

	r_1	r_2	r_3
C_1	3	2	1
C_2	4	3	2
C_3	2	1	1
C_4	4	3	3

→

	r_v
C_1	[2,1,3]
C_2	[3,2,4]
C_3	[1,1,2]
C_4	[3,3,4]

→

	r_v	**rank**
C_4	[3,3,4]	1
C_2	[3,2,4]	2
C_1	[2,1,3]	3
C_3	[1,1,2]	4

Figure 1.1: Transformation steps leading to majority ranking definition. The original grade profile where three judges J_i grade four candidates C_j (leftmost) is transformed into a sorted equivalent profile (center-left) through anonymity. The majority grade is identified by the central column (light gray). Subsequently, the majority value sequences are computed (center-right). The lexicographic order of the candidate majority values determines the majority ranking (rightmost).

Majority Judgment is well suited in clinical evaluation contexts as it enables interpretable and nuanced aggregation of multi-feature profiles without relying on numerical assumptions or artificial scoring. It is particularly robust to heterogeneity among features and produces consistent, fair, and transparent stratifications.

1.2 Data Binning

In many data processing and analysis tasks, particularly when dealing with continuous variables, it becomes useful, or even necessary, to discretize numerical data. This transformation, known as *binning*, reduces the resolution of the data by grouping similar values into intervals, or *bins*. The result is a compact, often interpretable representation that can simplify downstream modeling, improve robustness to noise, and reduce computational cost.

Binning is a fundamental step in histogram construction, where the data domain is partitioned into a finite set of contiguous intervals. Each bin typically stores a summary statistic, such as the frequency or mean of the values it contains. While binning introduces approximation error, it enables fast querying, efficient visualization, and noise-tolerant learning, especially in large-scale or streaming data settings.

1.2.1 Common binning strategies

Several binning techniques have been proposed in the literature, ranging from naive heuristics to sophisticated optimization-based methods. Among the most widely used are the **Equal-Width** (uniform) and **Equal-Frequency** (quantile) binning approaches. These methods are simple, fast, and often sufficient in applications where speed is prioritized over accuracy.

Equal-Width (Uniform) Binning In equal-width binning, the domain of the variable X is divided into B intervals of the same length. Let $X_{\min}$ and $X_{\max}$ denote the minimum and maximum values of X, then the bin width is given by:

$$w = \frac{X_{\max} - X_{\min}}{B}.$$

Each bin b_i covers the interval:

$$[X_{\min} + (i-1)w,\, X_{\min} + iw), \quad \text{for } i = 1, \ldots, B-1,$$

with the final bin including the upper bound $X_{\max}$. The bin assignment for a value x is computed in constant time by:

$$b(x) = \left\lfloor \frac{x - X_{\min}}{w} \right\rfloor + 1.$$

This method is extremely efficient and easy to implement. However, it performs poorly on skewed or non-uniform distributions, as it may produce bins with highly unbalanced frequencies, leading to inaccurate approximations.

Equal-Frequency (Quantile) Binning In equal-frequency binning, also known as quantile binning, the data is sorted and divided into B bins such that each bin contains (approximately) the same number of data points. This method ensures that all bins are equally populated, which can help stabilize learning algorithms that are sensitive to frequency imbalance.

Formally, let N be the number of data points. The data is first sorted in ascending order: $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(N)}$. Then, each bin is constructed by selecting breakpoints at ranks:

$$\left\{ x_{\left(\lceil \frac{N \cdot i}{B} \rceil\right)} \right\}_{i=1}^{B-1}.$$

This process results in bin intervals that vary in width depending on the local density of the data, with narrower bins in dense regions and wider bins in sparse ones. Quantile binning is more adaptive than equal-width binning and handles skewed distributions better, but requires sorting the dataset.

While equal-width and quantile binning offer fast and simple mechanisms for histogram construction, they are limited in terms of accuracy. Neither method accounts for the actual distribution of data frequencies within bins, nor do they minimize any global error metric. As a result, they may produce suboptimal representations, particularly when the number of bins is small or the data distribution is highly irregular.

1.2.2 Optimal Binning via Sum Squared Error Minimization

The problem of constructing an optimal histogram centers on how to choose bucket boundaries to best approximate the original frequency distribution. This task becomes critical when histograms are used not only for visualization, but also for downstream quantitative tasks such as query optimization, selectivity estimation, or algorithmic decision-making. One of the most widely adopted and theoretically grounded approaches to this task involves minimizing the total Sum of Squared Errors (SSE) between the true frequencies and their bucket-level approximations.

The method described in this section, developed by Güvenir and Uysal [2004], focuses on optimal binning under a space constraint: given a limited number of buckets, the goal is to produce a partitioning of the data domain that minimizes SSE.

Let us define the problem formally. We consider a set of N frequency values $F = \{f_1, f_2, \ldots, f_N\}$, where each $f_i \in \mathbb{R}^{\geq 0}$ represents the frequency (or count) associated with a data value $v_i \in \mathbb{Z}$, sorted in increasing order. The task is to partition F into B contiguous intervals (buckets) $\{I_1, I_2, \ldots, I_B\}$, with each interval $I_b = [i_b, j_b]$ summarized by a single representative value h_b.

The most common choice for h_b is the average frequency within the bucket:

$$h_b = \frac{1}{j_b - i_b + 1} \sum_{k=i_b}^{j_b} f_k.$$

Given this, the error for bucket I_b is computed as:

$$\mathrm{SSE}(I_b) = \sum_{k=i_b}^{j_b} (f_k - h_b)^2.$$

The objective is to choose bucket boundaries such that the total SSE across all buckets is minimized:

$$\mathrm{SSE}_{\text{total}} = \sum_{b=1}^{B} \mathrm{SSE}(I_b).$$

We refer to this task as the *space-bounded histogram problem* under the SSE metric.

To solve this problem optimally, we employ a dynamic programming approach. Let $\mathrm{SSE}^*(i, k)$ denote the minimal SSE that can be achieved by partitioning the prefix $F[1 : i]$ into k buckets. This leads to the following recurrence:

$$\mathrm{SSE}^*(i, k) = \min_{j<i} \left\{ \mathrm{SSE}^*(j, k-1) + \mathrm{SSE}([j+1, i]) \right\},$$

with the base case defined as $\text{SSE}^*(i, 1) = \text{SSE}([1, i])$.

To compute $\text{SSE}([i, j])$ efficiently, we precompute two prefix arrays:

$$P[i] = \sum_{k=1}^{i} f_k \qquad \text{and} \qquad PP[i] = \sum_{k=1}^{i} f_k^2.$$

These allow the SSE of any interval $[i, j]$ to be computed in constant time using the identity:

$$\text{SSE}([i, j]) = PP[j] - PP[i-1] - \frac{(P[j] - P[i-1])^2}{j - i + 1}.$$

This results in an overall time complexity of $\mathcal{O}(N^2 B)$ for computing the optimal histogram.

Although the basic dynamic programming approach is exact, its quadratic dependence on N may be impractical for large datasets. Fortunately, SSE satisfies useful monotonicity properties that enable pruning of redundant computations. Specifically, for any $i < k < j$, the following inequality holds:

$$\text{SSE}([i, j]) \geq \text{SSE}([i, k]) + \text{SSE}([k + 1, j]).$$

This implies that once a candidate boundary results in an SSE exceeding the best known so far, further splits beyond that point can be skipped. Using this insight, one can apply binary search to accelerate the inner loop of the recurrence, and early stopping to reduce unnecessary evaluations. These optimizations significantly reduce the practical computation time while preserving optimality.

These features make SSE-based optimal histograms a powerful tool for data summarization, query optimization, and machine learning model evaluation, particularly when statistical accuracy and interpretability are required.

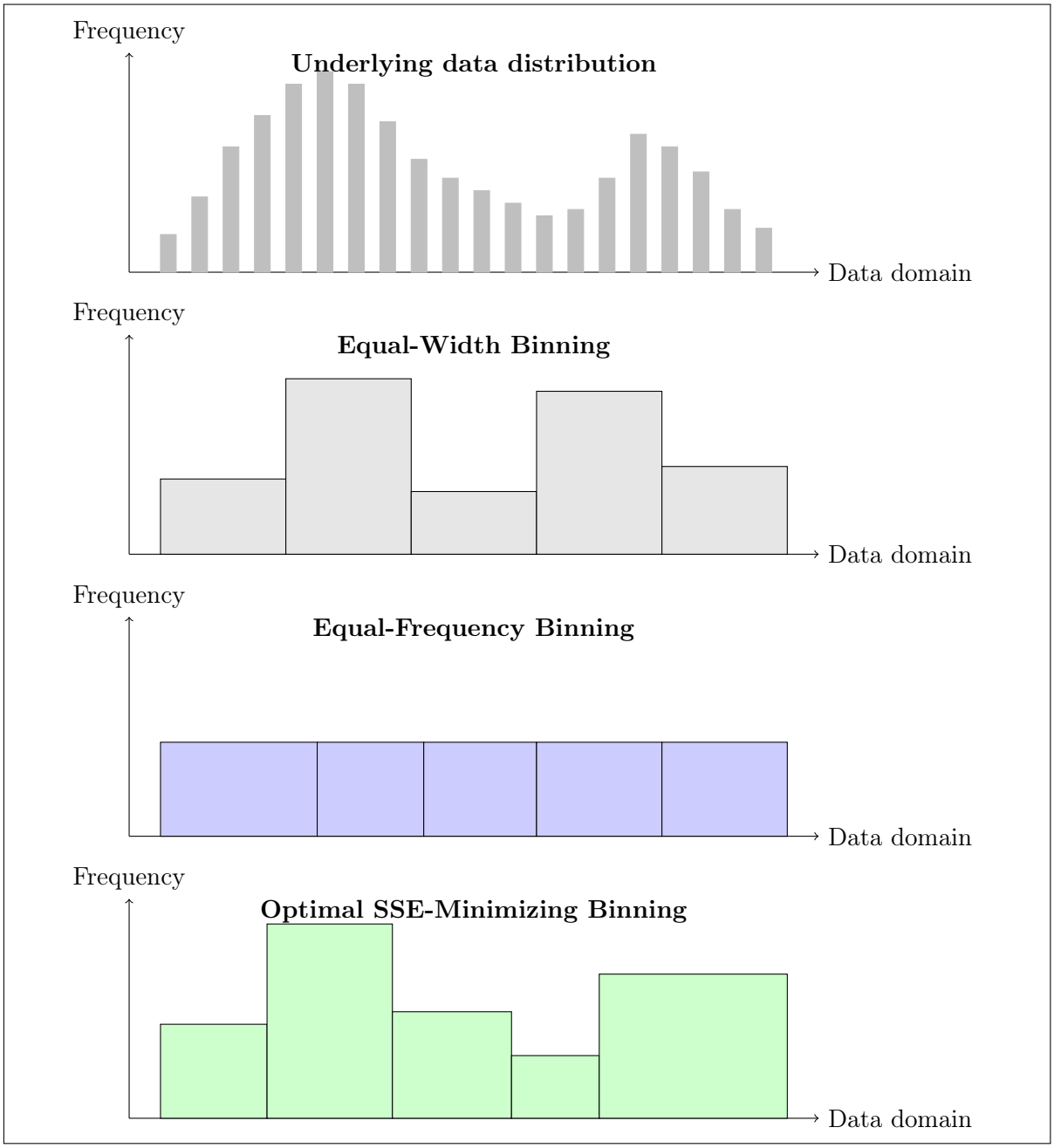

Figure 1.2: Visual comparison of binning strategies. The top plot shows a stylized data distribution. Equal-width binning divides the domain into fixed-size intervals, while equal-frequency ensures each bin contains an equal number of data points. Optimal binning dynamically adjusts bin boundaries to minimize the total squared reconstruction error (SSE).

1.3 Symbolic Regression with Genetic Programming

Symbolic regression (SR) is a powerful machine learning technique used to discover analytical expressions that best describe a dataset, without requiring the specification of a fixed functional form. Unlike traditional regression methods, symbolic regression searches both the space of model structures and their parameters simultaneously. This flexibility allows SR to uncover underlying laws in the form of interpretable mathematical expressions, making it particularly useful in scientific and engineering contexts.

The most prominent approach to symbolic regression is based on genetic programming (GP): an evolutionary algorithm inspired by biological evolution, explored in [Poli et al., 2008]. GP evolves a population of candidate mathematical expressions using operators such as crossover, mutation, and selection. Each expression is typically represented as a tree structure, where internal nodes denote mathematical operators and leaf nodes represent variables or constants.

1.3.1 Symbolic expression representation

Symbolic expressions in SR are typically represented as syntax trees. Each node of the tree corresponds to an element of a predefined function set (e.g., $+, -, \times, \div, \sin, \exp$), while the leaves correspond to input variables or constants. For instance, the expression:

$$f(x, y) = x \cdot \sin(y + 1)$$

can be represented as the tree in Figure 1.3.

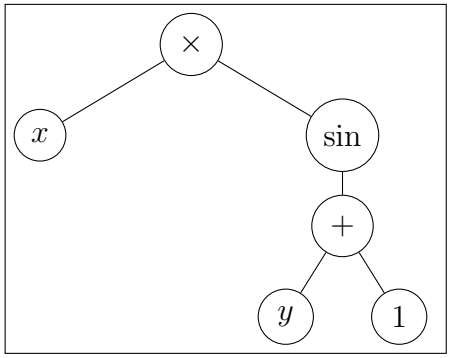

Figure 1.3: Tree representation of the expression $x \cdot \sin(y + 1)$.

This representation supports the application of genetic operations such as subtree crossover and mutation, which alter the structure of candidate solutions during the evolutionary process.

1.3.2 Genetic programming algorithm

The symbolic regression process begins by generating an initial population of random expression trees. The main steps of the genetic programming algorithm are as follows:

1. *Initialization*: Randomly generate a population of expressions, subject to a maximum depth or complexity constraint.
2. *Evaluation*: Compute a fitness score for each expression. A common metric is the sum of squared errors (SSE) between predicted and actual outputs.
3. *Selection*: Select a subset of high-performing individuals (expressions) for reproduction, typically via tournament or roulette-wheel selection.
4. *Crossover*: Swap subtrees between pairs of parent expressions to create offspring.
5. *Mutation*: Randomly modify parts of expressions by replacing a node or subtree.
6. *Replacement*: Form a new population from selected and newly generated expressions.
7. *Termination*: Repeat the above steps until a stopping criterion is met (e.g., maximum number of generations or convergence of fitness).

This process is highly flexible, allowing customization of the function set, tree constraints, and evolutionary parameters to suit the problem domain.

1.3.3 Symbolic Regression for Prediction

In medical decision-making contexts, predictive models must often satisfy two complementary requirements: accurate estimation of clinical outcomes and interpretability of the underlying reasoning process. SR provides a principled way to address both needs by discovering closed-form mathematical expressions that relate input features (e.g.,

patient measurements, clinical scores or, in our case, Intrinsic Capacity) to a continuous target variable. While SR is commonly applied to regression tasks, it can also be adapted to classification problems, particularly when the goal is not just to predict a binary label, but to estimate a meaningful and interpretable risk score that correlates with patient status.

In such applications, SR is used to model a continuous surrogate signal that is associated with the target of interest. The symbolic model learns a function $f(\mathbf{x}) \in \mathbb{R}$ from a predefined set of mathematical operators, where $\mathbf{x} \in \mathbb{R}^d$ denotes the vector of input features. This function is optimized to approximate a reference value that encodes clinical judgments or aggregate evaluations informative with respect to the target. Once trained, the symbolic expression serves as a transparent scoring function that can be evaluated on new patient data. The output $f(\mathbf{x})$ can then be:

- interpreted directly as a continuous indicator of patient state;
- used in regression or logistic models to predict a second target output.

This approach preserves the interpretability of the model, since the symbolic formula is explicitly readable, and allows for downstream classification without requiring opaque black-box mechanisms. Moreover, the structure of the learned formula can offer insights into the nonlinear interactions between features and their influence on the outcome, which is especially valuable in clinical and biomedical research settings.

In summary, symbolic regression offers a flexible and interpretable approach for deriving meaningful scores correlated with a given set of features. Beyond this, it can serve as an initial step in predicting a secondary target known to be related to the first, such as, in our case, predicting the healing outcome based on the IC grade. These models effectively bridge the gap between human expert judgment and algorithmic prediction, making them particularly valuable in fields where transparency and domain validation are paramount.

Chapter 2

Dataset

This study is based on a clinical dataset comprising records from **983 patients** undergoing follow-up evaluation for long COVID. The dataset integrates standardized indicators of health, detailed clinical covariates, and a recovery label. It is used for both ordinal ranking and interpretable prediction.

2.1 Overview and General Statistics

Each row in the dataset corresponds to a unique patient visit and includes three types of variables:

- **WHO-based functional indicators (IC features)**, which serve as judges for deriving consensus-based patient rankings;
- **Clinical covariates**, used for interpretable symbolic modeling;
- **Target variable** `healed`, indicating whether the patient was considered clinically recovered.

Among the patients, **62.6%** were labeled as `healed = 1`, indicating clinical recovery, and **37.4%** as `healed = 0`.

2.2 Feature Categories and Measurement

IC Indicators (Judges)

A total of 14 IC indicators were used to evaluate the multidimensional functional state of each patient. These indicators were derived from questionnaires covering the five IC health domains: **locomotion, sensory, vitality, cognitive, and psychosocial**.

The questionnaires included:

- **Depression Anxiety Stress Scales (DASS):**
  Evaluates levels of depression, anxiety, and stress, providing insight into psychosocial well-being.

- **EuroQol 5-Dimension 5-Level (EQ-5D-5L):**
  Assesses health-related quality of life across mobility, self-care, pain, and emotional states, reflecting locomotion, vitality, and psychosocial capacity.

- **Insomnia Severity Index (ISI):**
  Measures the severity of sleep disturbances, impacting vitality and cognitive performance.

- **Connor-Davidson Resilience Scale (CD-RISC):**
  Assesses psychological resilience, indicating emotional regulation and psychosocial functioning.

- **Short Form Health Survey with 36 items (SF-36):**
  Covers physical functioning, energy levels, emotional well-being, and social engagement, relating to locomotion, vitality, cognitive health, and psychosocial status.

Together, these tools provide a multidimensional representation of an individual's intrinsic capacity, consistent with the WHO framework.

Clinical Covariates for Symbolic Modeling

The symbolic regression model relies on 25 covariates describing the patient's objective health condition. These features span the following domains:

- **Quick IC domain assessment via short non-validated questionnaires** (`Q_cognitive`, `Q_psychosocial`, `Q_locomotion`, `Q_sensory`, `Q_vitality`);
- **Cognitive and psychosocial function:** cognitive impairment via Cogstate Brief Battery (`cognitive_impaired`);
- **Physical and locomotor capacity:** including SPPB, gait speed (adjusted by sex and height), and hand grip strength;
- **Respiratory function:** evaluated via the Modified Medical Research Council dyspnea scale (`mMRC_scale`);
- **Nutritional status:** BMI, significant weight loss post-hospitalization, and protein intake balance;
- **Symptom burden:** number of mild, moderate, and severe symptoms reported;
- **Comorbidities and lifestyle:** presence of fibrosis, steatosis, smoking status, physical activity, polypharmacy.
- **Demographics:** age and sex.

Continuous covariates are standardized (zero mean, unit variance). Binary indicators are encoded as 0/1.

2.3 Exploratory Data Analysis

Distribution of IC Features

The IC indicators show a generally right-skewed distribution, with most patients reporting moderate to high levels of functioning. This is particularly evident in quality-of-life, energy, and general functioning scores. In contrast, variables such as stress, insomnia, and emotional well-being are more dispersed, suggesting greater variability in psychological domains.

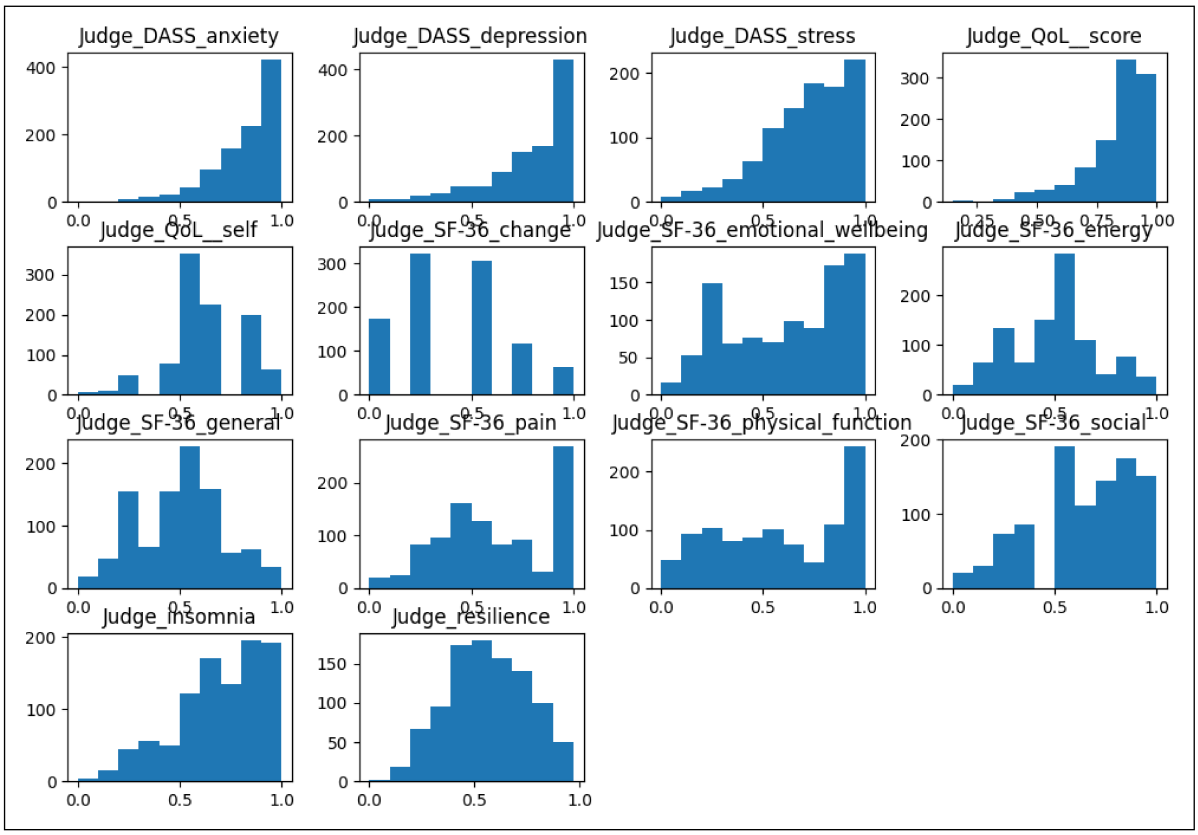

Figure 2.1: IC-features' distributions

Missing Data and Cleaning Procedure

To avoid imputation biases in symbolic regression and ordinal ranking, all records with missing values in any IC indicator or covariate were excluded when needed. This filtering reduced the initial cohort from 983 to **737 patients** with complete information. This subset was used for all subsequent modeling tasks.

Feature Statistics and Correlation with Recovery

We analyzed the statistical properties of covariates and their association with the target variable `healed`. Depending on variable type, the appropriate correlation metric was used: Spearman correlation for ordinal and continuous variables and Point-biserial correlation for binary features. The following covariates showed the strongest correlations with recovery status:

- **mMRC scale** ($\rho = -0.31$): breathlessness is inversely correlated with healing, as expected in post-COVID respiratory dysfunction.
- **Cognitive function (Q_cognitive)** ($\rho = +0.25$): higher cognitive scores correlate with a greater chance of recovery.
- **Number of moderate symptoms** ($\rho = -0.24$): more persisting symptoms reduce the likelihood of healing.
- **Locomotor ability (Q_locomotion)** ($\rho = +0.24$): better movement capability improves recovery outcomes.
- **SPPB $\geq$ 10** ($\rho = +0.21$): reaching a high threshold on physical performance test is a strong positive signal.
- **Steatosis** ($\rho = -0.17$): liver dysfunction appears negatively associated with healing.
- **Psychosocial function (Q_psychosocial)** ($\rho = +0.17$): stronger psychosocial resilience supports recovery.
- **Hypoproteic status** ($\rho = -0.15$): protein deficiency undermines healing capacity.
- **Hyperproteic status** ($\rho = +0.15$): high protein levels correlate with better recovery.
- **Weak hand grip** ($\rho = -0.15$): muscular frailty is linked to worse outcomes.

These results are consistent with clinical observations and confirm the relevance of selected covariates for symbolic modeling and ordinal health stratification.

Chapter 3

Experimentation

This chapter details the experimental methodology developed and implemented entirely in Python throughout the research. The primary aim is to construct a robust and interpretable computational pipeline capable of producing a single, continuous scoring system that accurately approximates the Intrinsic Capacity as defined by the World Health Organization. By integrating standardized functional indicators with a comprehensive set of covariate features, this score seeks to capture the multifaceted aspects of patient health and functional ability in a data-driven yet clinically meaningful manner. Subsequently, we evaluate the validity and utility of this derived score by testing its capacity to serve as a predictive basis for long COVID outcomes, an endpoint already known to be correlated with Intrinsic Capacity.

The first stage of experimentation involved analyzing different strategies to discretize the 14 IC-features, which represent multidimensional indicators of patient function. Several binning techniques were tested and compared in order to obtain a suitable voting matrix for ranking procedures.

Building on this, various ranking algorithms were evaluated to determine how to aggregate the discretized IC scores into a consistent and ordinal ranking of patients.

Once a reliable ranking framework was defined, the focus shifted to symbolic regression. This phase aimed to discover interpretable mathematical expressions, constructed from clinical covariates, that could approximate the consensus-based functional ranking. Special attention was given to the design of an objective function that aligned well with the ordinal nature of the ranking task.

The final symbolic models were then assessed not only for their agreement with the derived rankings, but also for their ability to reflect the true recovery status, as indicated by the `healed` variable.

3.1 Discretization of IC Features

The 14 Intrinsic Capacity features considered in this study represent diverse and complementary dimensions of patient functionality. These include cognitive capacity, emotional resilience, physical performance, perceived well-being, and social participation. Each variable originates from clinically validated scales, such as DASS-21, SF-36, and domain-specific sleep or resilience scores, and was pre-processed by normalization to the $[0, 1]$ interval. This standardization facilitates direct comparison between heterogeneous indicators and simplifies subsequent analysis.

However, in order to construct a rank-based consensus of patient health status using tools from social choice theory, these continuous IC features need to be converted into discrete, ordinal categories. This transformation enables the use of aggregation mechanisms that operate on symbolic or rank-level inputs, such as majority judgment and lexicographic comparison. In this context, each IC domain is interpreted as a "voter" evaluating every patient and assigning them a grade from a fixed set $\Lambda = \{1, 2, 3, 4, 5\}$, where higher values indicate better functional status.

From a clinical and modeling standpoint, this discretization provides multiple benefits:

- It transforms numerical scores into interpretable qualitative judgments (e.g., poor, fair, good, very good, excellent);
- It smooths out minor variations and noise, reducing the effect of outliers in judgment;
- It provides a common language across domains, enabling fair aggregation and ranking.

Moreover, many of the IC features already have natural or clinically motivated interpretations that align with 5 ordinal levels, for this motive we decided to bin the features into 5 buckets (grades).

3.1.1 From Continuous Scores to Discrete Grades

The transformation process involved applying binning procedures to each of the 14 IC features separately. The goal was to assign to each patient, for each IC feature, a grade $g \in \Lambda$ indicating their level of functioning in that domain. This grade corresponds to the bin into which the patient's normalized score falls. The output of this process is a new symbolic matrix $\Phi \in \Lambda^{m \times 14}$, where m is the number of patients and each entry $\Phi(i, j)$ is the grade assigned to patient i by IC feature j.

To rigorously assess which binning strategy yields the most meaningful discretization, we applied and compared three methods:

1. Equal-width binning (uniform);
2. Quantile binning;
3. Optimal binning based on within-bin variance minimization.

Each method was independently applied to all 14 features, resulting in three complete versions of the dataset, each with binned values from 1 to 5. These datasets were then used in the ranking and modeling phases to compare performance and interpretability.

Equal-width binning. This method partitions the interval $[0, 1]$ into five equal-sized subintervals. It is straightforward to implement and provides a fixed reference scale. However, it does not adapt to the actual data distribution. If most patient scores are concentrated in a narrow range, as often happens in clinical data, this method can produce unbalanced bins, some sparsely populated or even empty.

Quantile binning. Here, bin thresholds are determined by computing the empirical quantiles of each IC feature so that each bin contains (approximately) the same number of patients. This approach ensures balanced bin frequencies and avoids sparsity issues. However, it can produce irregular bin widths, and if many patients share identical or similar values, some grades may collapse into indistinct groupings, which may reduce interpretability.

Optimal binning. To achieve the best compromise between interpretability and data representation, we also implemented a supervised binning method that minimizes the total intra-bin variance (sum of squared errors, SSE). This dynamic programming algorithm computes the $k = 5$ bin boundaries that minimize the SSE across all patients for each IC feature. The resulting discretization adapts to the feature's distribution, reflecting clusters or significant value transitions.

3.1.2 Comparison and Practical Application

All three binning strategies were implemented and tested systematically on the 14 IC features. For each method, we generated a full version of the discretized dataset, with grades ranging from 1 to 5 for every patient-feature pair. These symbolic matrices were then used as inputs to downstream ranking and modeling procedures.

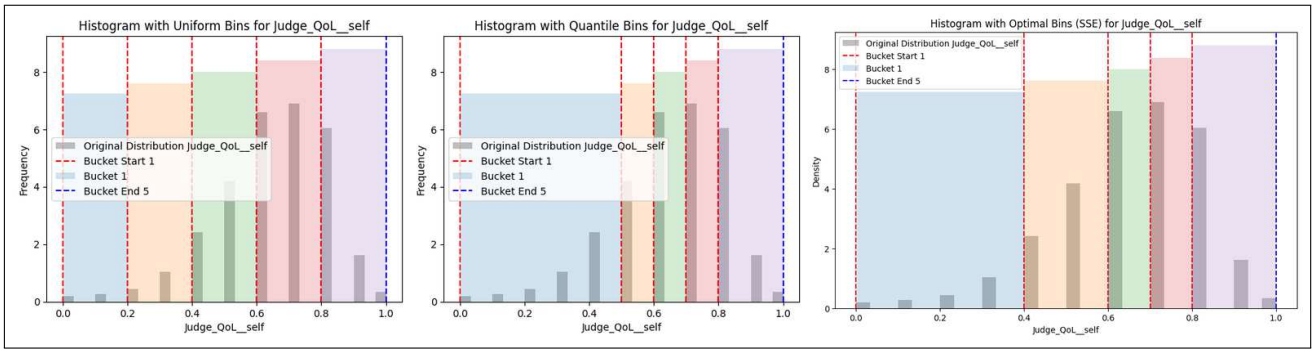

Figure 3.1: Comparison of the three binning methods on the `QoL_self` feature. Uniform binning divides the range evenly, quantile binning balances the number of patients per bin, and optimal binning captures distributional shifts.

Uniform binning, while simple, proved sensitive to data skewness and led to frequent imbalances across bins. Quantile binning guaranteed bin balance but was less effective in distinguishing clinically relevant subgroups. Optimal binning, although computationally more intensive, provided the best compromise, adapting flexibly to each feature's distribution and yielding interpretable grade distinctions.

3.2 Patient Ranking from Discretized IC-Values

3.2.1 Building the Ranking Systems

Once each IC questionnaire has been discretized into an ordinal scale from 1 to 5, we obtain for each patient a vector of 14 symbolic votes. These vectors, one per patient, represent the evaluations expressed by the 14 "judges" (i.e., IC features) on the patient's functional condition. The next step is to transform this multidimensional symbolic information into a single ordered list: a patient ranking that reflects their overall functional status based solely on intrinsic capacity.

This requires defining a *social ranking function*: a method to aggregate these symbolic evaluations into a transitive, interpretable, and ideally robust ordering. In this work, we explored three aggregation strategies:

1. The standard **lexicographic order**;
2. A modified q-**lexicographic order**;
3. The **Majority Judgment** method.

LEXICOGRAPHIC Lexicographic (dictionary-style) ranking compares patients by sorting their grade vectors in decreasing order and then comparing them element by element, starting from the first. Given two sorted vectors, the first position at which the grades differ determines the ranking. While simple, this method is highly sensitive to the initial values, which may exaggerate the influence of one or two features.

Example.
Consider three candidates (patients) and their corresponding raw votes (feature values). First, we sort each patient's votes in descending order:

Patient A

Raw: [5, 2, 3, 2, 4, 4, 5, 4, 4, 3, 3, 5, 1, 4]

Sorted: [5, 5, 5, 4, 4, 4, 4, 4, 3, 3, 3, 2, 2, 1]

Patient B

Raw: [4, 2, 3, 2, 4, 4, 4, 4, 4, 3, 3, 4, 4, 4]

Sorted: [4, 4, 4, 4, 4, 4, 4, 4, 4, 3, 3, 3, 2, 2]

Patient C

Raw: [3, 3, 4, 4, 4, 4, 5, 3, 3, 4, 5, 5, 3, 3]

Sorted: [5, 5, 5, 4, 4, 4, 4, 4, 3, 3, 3, 3, 3, 3]

We then compare the sorted vote sequences element-wise:

- Patient A vs. Patient B: the first element is 5 vs. 4 $\Rightarrow$ **Patient A $>$ Patient B**
- Patient A vs. Patient C:
  - 1st to 11th elements are equal
  - 12th element: 2 (A) vs. 3 (C) $\Rightarrow$ **Patient C $>$ Patient A**

Therefore, the lexicographic ranking order is:

Patient C $>$ Patient A $>$ Patient B

q-LEXICOGRAPHIC To reduce the influence of the top-most grades, we use a q-lexicographic variant where the comparison begins at a more central point in the distribution. Specifically, we start from the 11^{th} grade (the 75th percentile in a 14-element vector) then we delete it from the sequence and place it into another string (which will be the grade), then we can restart the process itiratevely from the 75th percentile of the new sequence until it has only one vote left. This places greater emphasis on the body of the distribution rather than the extremes, resulting in rankings that are more robust to local variations.

Example.

Let us compare the same patients as in the example from subsection 3.2.1. At each iteration, we extract the element at position $\tilde{m} = \lfloor 0.75 \times \text{length} \rfloor$ from the sorted list of votes, remove it from the sequence, and proceed to the next step.

Step	Current Sequence	Extracted Value
1	[5,5,5,4,4,4,4,4,3,3,3,2,2,1]	3 (11th)
2	[5,5,5,4,4,4,4,4,3,3,2,2,1]	3 (10th)
3	[5,5,5,4,4,4,4,4,3,2,2,1]	3 (9th)
4	[5,5,5,4,4,4,4,4,2,2,1]	2 (9th)
5	[5,5,5,4,4,4,4,4,2,1]	4 (8th)
6	[5,5,5,4,4,4,4,2,1]	4 (7th)
7	[5,5,5,4,4,4,2,1]	4 (6th)
8	[5,5,5,4,4,2,1]	4 (5th)
9	[5,5,5,4,2,1]	5 (4th)
10	[5,5,5,2,1]	5 (3rd)
11	[5,5,2,1]	5 (3rd)
12	[5,2,1]	2 (2nd)
13	[5,1]	5 (1st)
14	[1]	1 (1st)

Applying this method to all three the patients we obtain that the Final q-Lex Grade are:

Patient A

3 3 3 3 4 4 4 4 5 5 5 2 5 1

Patient B

3 3 3 4 4 4 4 4 4 4 2 2 4 2

Patient C

3 3 3 3 4 4 4 4 5 5 5 3 5 3

The q-lexicographic ranking, derived from the lexicographic order of these final grades, is:

Patient B > Patient C > Patient A

since Patient B's grade at the 4th position is 4, which is greater than Patient C's 3, and Patient C's grade at the 12th position is 3, which is greater than Patient A's 2.

MAJORITY JUDGMENT The Majority Judgment (MJ) method assigns a consensus grade by identifying the lowest grade that at least 50% of the judges have given or exceeded. This corresponds to finding the *median* grade, with one important detail: when the number of grades is even, the median is taken as the lower of the two middle values.

To enable complete comparability and ranking between patients, we represent their full majority judgment as a 14-digit string using the following procedure:

1. Sort the 14 grades in descending order.

2. Iteratively extract the median grade from the current list according to:

$$\tilde{m} = \begin{cases} \frac{n+1}{2}, & \text{if } n \text{ is odd} \\ \frac{n+2}{2}, & \text{if } n \text{ is even} \end{cases}$$

   where n is the length of the current list. That is, if n is odd, take the middle element; if n is even, take the lower of the two central elements. Remove this element from the list and append it to the output string.

3. Repeat this process until all grades have been extracted and the list is empty.

4. The resulting sequence of 14 extracted grades forms a string that can be converted into an integer for direct comparison and ranking.

Example. Using the same three patients from the previous example, we apply the Majority Judgment (MJ) method to derive their full consensus grades.

Patient A (sorted grades)

$$[5, 5, 5, 4, 4, 4, 4, 4, 3, 3, 3, 2, 2, 1]$$

Step	Current List	Extracted Median (position)
1	[5,5,5,4,4,4,4,4,3,3,3,2,2,1]	4 (8th)
2	[5,5,5,4,4,4,4,3,3,3,2,2,1]	4 (7th)
3	[5,5,5,4,4,4,3,3,3,2,2,1]	3 (7th)
4	[5,5,5,4,4,4,3,3,2,2,1]	4 (6th)
5	[5,5,5,4,4,3,3,2,2,1]	3 (6th)
6	[5,5,5,4,4,3,2,2,1]	4 (5th)
7	[5,5,5,4,3,2,2,1]	3 (5th)
8	[5,5,5,4,2,2,1]	4 (4th)
9	[5,5,5,2,2,1]	2 (4th)
10	[5,5,5,2,1]	5 (3rd)
11	[5,5,2,1]	2 (3rd)
12	[5,5,1]	5 (2nd)
13	[5,1]	1 (2nd)
14	[5]	5 (1st)

Applying the same procedure to Patients B and C, we obtain the following MJ's grades:

Patient A

4 4 3 4 3 4 3 4 2 5 2 5 1 5

Patient B

4 4 4 4 3 4 3 4 3 4 2 4 2 4

Patient C

4 4 3 4 3 4 3 4 3 5 3 5 3 5

Finally, the lexicographic comparison of these MJ grades leads to the ranking:

Patient B > Patient C > Patient A

This ordering follows because Patient B's grade at the 4th position (4) is higher than Patient C's (3), and Patient C's grade at the 9th position (3) is higher than Patient A's (2), reflecting a stronger overall consensus in favor of Patient B, followed by Patient C, then Patient A.

3.2.2 Comparison of Ranking Strategies and Final Method Selection

In this section, we compared three aggregation methods for ranking patients based on their discretized intrinsic capacity (IC) features: the standard lexicographic ranking, the q-lexicographic variant, and the Majority Judgment (MJ) approach. The lexicographic method is straightforward and intuitive. It compares patients element-by-element, starting from the highest grade in the sorted vector. However, it tends to overweight the most prominent values, making the final rank sensitive to a few extreme features. For instance, in our comparative example, Patient C's early high grades pushed them above others despite worse values in the remaining dimensions, demonstrating the potential distortive effect of outliers.

To mitigate this issue, the q-lexicographic variant modifies the comparison strategy: rather than beginning from the top-most grade, it starts at the 75^{th} percentile and proceeds iteratively. This method emphasizes the middle-to-upper part of the distribution, aiming to reduce bias from extreme values. However, its iterative logic introduces complexity and it often still selects values close to the edges of the distribution, limiting its robustness.

In contrast, the Majority Judgment method addresses these limitations through a consensus-driven mechanism. At each iteration, it extracts the median grade from the remaining values (selecting the lower median when the number is even), forming a symbolic sequence that reflects the most central and agreed-upon assessments. This sequence is then converted into a unique numeric grade for each patient, which serves both ranking and modeling purposes. MJ focuses on collective agreement rather than individual extremities, producing rankings that are more balanced and interpretable.

To empirically assess each method, we evaluated the statistical association between the resulting grades or rankings and the binary clinical outcome `healed`. The analysis was conducted across the three binning strategies used to discretize the "Judges" (IC-features) in section 3.1: uniform, quantile, and optimal (SSE-based), using two complementary statistical indicators:

- Chi-squared test and Cramér's V assess the association between grades (expressed as 14-digit strings) and the binary healing outcome. Cramér's V ranges from 0 (no

association) to 1 (perfect association), with values above 0.6 generally considered strong.

- Point-biserial correlation (r_{pb}) measures the linear relationship between a patient's numeric rank (as determined by the ranking method) and the binary healing status. Values close to –1 or 1 denote strong associations, while values near 0 indicate weak or no relationship.

Table 3.1: Statistical association of patient grades and rankings with the binary outcome `healed`, across binning strategies. Bold highlights the best result in each line.

Method / Metric	**Uniform**	**Quantile**	**Optimal**
Majority Grade *(Chi², Cramér's V)*			
Chi² (p-value)	458.6 (0.027)	458.6 (0.027)	**485.7 (0.006)**
Cramér's V	0.796	0.796	**0.820**
Majority Ranking *(Point-biserial)*			
r_{pb}	**–0.318 (<0.001)**	**–0.318 (<0.001)**	–0.311 (<0.001)
Lex Grade *(Chi², Cramér's V)*			
Chi² (p-value)	456.5 (0.027)	501.4 (0.084)	**483.5 (0.005)**
Cramér's V	0.795	0.833	**0.818**
Lex Ranking *(Point-biserial)*			
r_{pb}	**–0.332 (<0.001)**	–0.316 (<0.001)	–0.294 (<0.001)
q-Lex Grade *(Chi², Cramér's V)*			
Chi² (p-value)	458.6 (0.027)	515.8 (0.101)	**485.7 (0.006)**
Cramér's V	0.796	0.845	**0.820**
q-Lex Ranking *(Point-biserial)*			
r_{pb}	–0.290 (<0.001)	–0.291 (<0.001)	**–0.294 (<0.001)**

As shown, the combination of Majority Judgment with optimal binning consistently outperformed the other methods across both statistical metrics. It achieved the strongest association between patient grade and healing status, with a Cramér's V of 0.820 ($p < 0.01$), and maintained a solid point-biserial correlation between rank and outcome ($r_{\text{pb}} = -0.311$, $p < 0.001$).

Although lexicographic ranking reached a slightly higher correlation coefficient in one instance ($r_{\text{pb}} = -0.332$ under uniform binning), its performance was more sensitive to the binning strategy, and its categorical grades showed less stable associations with the outcome.

In contrast, the q-lexicographic method exhibited inconsistent behavior: despite achieving the highest Cramér's V (0.845) under quantile binning, the associated p-value was not statistically significant ($p = 0.101$), suggesting the presence of ties and flat distributions that undermine its reliability. This instability further supports the conclusion that Majority Judgment, particularly when paired with optimal binning, offers the best balance of robustness, interpretability, and clinical relevance.

3.3 Interpretable Scoring of Intrinsic Capacity: A Symbolic Approach

Having obtained a consensus-based ranking and a symbolic grade for each patient through the Majority Judgment (MJ) method (applied to optimally binned values), the next logical step is to reproduce this ranking using only the available covariate features, those not involved in the original IC-based evaluation. This decouples the ranking process from intrinsic capacity scores and enables generalization to new patients, based solely on physical, clinical, and demographic predictors.

The objective is therefore to learn a symbolic function, interpretable and compact, that maps a patient's covariates to an ordinal value that mimics the MJ-based functional grade. Ideally, this output should preserve the same semantic structure as the original MJ scale and correlate with the clinical outcome `healed`, as the MJ ranking itself has already been shown to do.

To achieve this, we employ Symbolic Regression (SR), a method that evolves mathematical expressions to fit data while maintaining interpretability. However, standard SR optimization (e.g., minimizing squared error with respect to a real-valued target) is not directly suitable for our goal, which involves reproducing a symbolic scale and ordering.

Thus, we implemented and tested three customized objective functions to guide the

symbolic search:

1. **Spearman correlation with the MJ rank:** the symbolic output is optimized to have the highest monotonic correlation with the MJ-derived patient ranking. This encourages the function to learn the correct patient ordering, even if the predicted values are not numerically identical.

2. **Wasserstein distance from the grade distribution:** here, the symbolic output is expected to reproduce the empirical distribution of MJ grades. By minimizing the Wasserstein distance, we ensure that the learned function generates an output scale statistically aligned with the original grades.

3. **Residual difference between binned symbolic output and MJ grades:** this third approach is designed to capture ordinal structure directly. Since the MJ grades belong to a 5-level symbolic scale, we first discretize the symbolic output, produced by the regression model, into 5 bins using either quantile or optimal binning. The same binning procedure is applied to the original MJ grades. The loss is then defined as the average absolute difference between the binned values of the symbolic formula and the corresponding MJ grades.

   This method is particularly well-suited for our setting, as it treats the symbolic output as a source of ordinal information rather than a numeric estimate. It emphasizes alignment in symbolic categorization, not just in distribution or correlation. In essence, we are encouraging the model to reproduce the same semantic grade levels observed in the MJ evaluation, thereby capturing a meaningful stratification of patients. By focusing on minimizing discrepancies in binned levels, this objective supports the creation of interpretable, discretized scores that reflect a functional ordering consistent with clinical judgment.

3.3.1 Fitness function selection

The goal is to discover an analytical expression that best models a given dataset by evolving symbolic representations of candidate solutions. To evaluate how well each candidate performs, it is necessary to define one or more fitness functions. A **fitness function** assigns a numerical score to each expression, indicating its quality

with respect to certain desired properties. In this work, we compare and integrate three different fitness functions specifically designed to address the ranking quality and distributional properties of the model output. The first is based on the Spearman Correlation (ρ) between the model output and the Majority Ranking (based on Intrinsic Capacity), the second fitness function uses the Wasserstein Distance to compare the distribution of model outputs to the empirical distribution of the IC majority grade and the third function, Binning Diversity, assesses the similarity between the model output and the IC grade after both are discretized using a quantile or a optimal binning scheme (with five bins).

Alongside these accuracy, and distribution-oriented objectives, we also incorporate a structural complexity term to favor more interpretable expressions. When only a single criterion is optimized, the regression problem can be treated as a single-objective optimization task, where the goal is, in this case, to minimize a single fitness value: $1 - \rho$ for the first strategy, *Wasserstein Distance* for the second or, lastly, *Binning Discrepancy*. However, the inclusion of multiple conflicting objectives requires a multi-objective optimization framework.

In multi-objective symbolic regression, each candidate is evaluated with respect to multiple fitness functions simultaneously. Since these objectives often conflict with one another, it is usually not possible to find a single expression that optimizes all objectives at once. Instead, the evolutionary algorithm maintains and evolves a set of non-dominated solutions known as the **Pareto front**. A solution is considered non-dominated if there is no other solution in the population that performs equally well or better across all objectives and strictly better in at least one. The Pareto front, therefore, represents the set of optimal trade-offs found during the evolutionary process. Rather than selecting a single best model, this approach provides a spectrum of alternatives, each representing a different balance between the competing objectives. This allows domain experts to select the most appropriate model based on the specific needs of the task.

To assess the effectiveness of the proposed fitness functions and determine which configurations yield the most meaningful symbolic expressions, we performed a validation step prior to applying the final target transformation. Specifically, we evaluated each symbolic model based on how well its output aligns with the IC Majority ranking. This

step did not involve the "healed" target, ensuring that the analysis focused purely on the ability of the symbolic expression to reproduce the derived ordinal structure of the data.

The comparison was carried out by computing the rank loss between the output of each symbolic expression and the IC rank. Rank loss quantifies the number of pairwise ranking inversions between the predicted and true ranks, with lower values indicating better agreement. For each fitness function combination tested, we extracted several expressions from the first Pareto front and evaluated their rank loss on the training data.

This validation approach allowed us to compare the relative strengths of the fitness criteria in producing outputs that are consistent with the MJ-derived ranking. It also provided an empirical basis for selecting the most appropriate fitness function set to use in subsequent experiments, including those involving the "healed" target. By deferring the use of the final target variable, we ensured that this step remained purely diagnostic and focused on assessing alignment with the input data's intrinsic ranking structure.

First, the dataset was split into a training set and a test set using a 75%-25% proportion. This resulted in 544 records for training and 179 records for testing. To evaluate the relative performance of the different fitness functions, we conducted a validation test, on the training set, where each fitness function was used individually. The symbolic regression algorithm was allowed to evolve expressions for 15 minutes and for each run we selected the last expression from the corresponding Pareto front.

The performance of each resulting expression was assessed by computing the rank loss between its output and the IC rank. The results are summarized below:

- **Spearman correlation**: the selected formula was

$$
\begin{aligned}
&-3.157 + 2.820 \cdot (Q_{\text{vitality}} \cdot (|Q_{\text{psychosocial}}|^{3.443} \cdot \\
&\quad \cdot (|\text{age}|^{Q_{\text{locomotion}}} \cdot |Q_{\text{psychosocial}} + 0.391 \cdot \text{mMRCScale}|^{-1.191})))
\end{aligned}
$$

  The rank loss with respect to the IC rank was **0.25**, indicating a positive correlation.

- **Wasserstein distance**: the selected formula was

$$-14.968 \cdot Q_{\text{locomotion}}$$

The rank loss with respect to the IC rank was **0.60**, suggesting a poor alignment with the IC-based ranking.

- **Quantile Binning diversity** : the selected formula was

$$Q_{\text{psychosocial}} \cdot |-8.893 \cdot Q_{\text{vitality}} + 1.544 \cdot \text{mMRCScale}|^{6.263}$$

The rank loss with respect to the IC rank was **0.26**, also indicating a positive but slightly weaker correlation than the Spearman-based fitness.

- **Optimal Binning diversity**: the selected formula was

$$Q_{\text{psychosocial}}$$

The rank loss with respect to the IC rank was **0.33**.

These results indicate that the fitness functions based on Spearman rank correlation and quantile binning diversity tend to produce expressions more consistent with the majority ranking structure, as reflected by their lower rank loss values. Motivated by this observation, we also explored the combination of these two fitness functions.

Building on these findings, we ran the symbolic regression model for 30 minutes using both the Spearman correlation and binning diversity as joint fitness functions. The resulting expression was:

$$Q_{\text{psychosocial}} + (Q_{\text{vitality}} \cdot (-0.527 \cdot \text{mMRCScale} + \quad Q_{\text{psychosocial}} \cdot (\text{age} + |Q_{\text{cognitive}}|^{9.585})))$$

The rank loss with respect to the IC rank was **0.24**, representing the best performance observed across the configurations; we also tested the combination of Spearman correlation with optimal binning diversity as fitness functions, however, this setup yielded worse results, with an average rank loss above **0.35**. For this reason, this last configuration was excluded from the final analysis.

To visually compare the distribution of predicted bins, derived from the final symbolic expression reported above, and true IC grades, we plotted their histograms overlapped with partial transparency. This allows us to assess how well the model's output captures the overall distributional characteristics of the target grades.

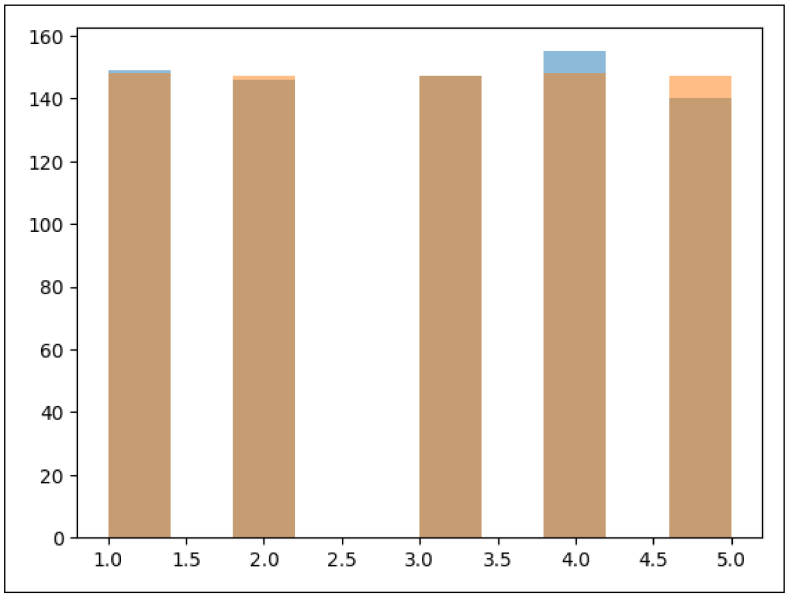

Figure 3.2: Quantile-binned distributions, in orange: IC-grades and in blue: formula output.

The strong overlap between the two distributions indicates that the symbolic regression model successfully captures the overall distributional pattern of the target variable. Minor discrepancies appear in some bins, but the general alignment supports the model's validity in reproducing the rank structure and distribution of the IC grades.

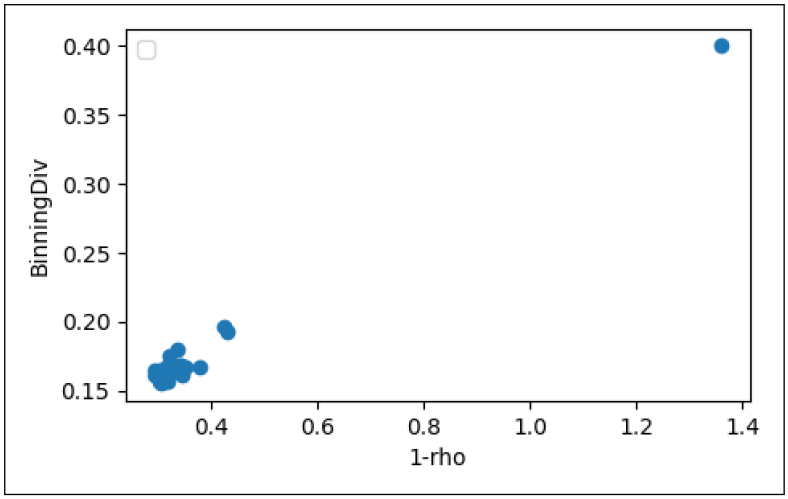

Figure 3.3: Scatter plot of the first Pareto front solutions, showing the trade-off between $1-\rho$ (Spearman correlation loss) and Binning Diversity, without considering formulas' complexity. Each point corresponds to a symbolic regression candidate evaluated on both objectives.

This scatter plot demonstrates the inherent tension between optimizing for rank correlation and binning similarity. Models with superior rank preservation ($1 - \rho$ close to zero) sometimes show reduced binning diversity, and vice versa. Such a distribution underscores the necessity of a multi-objective framework to balance conflicting goals and select models that best fit the problem context (the dots near the origin).

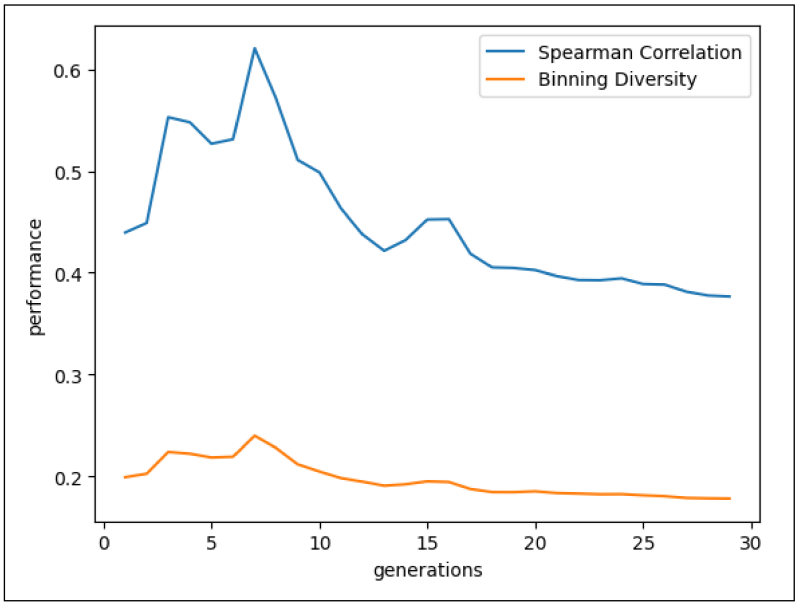

Figure 3.4: Evolution of average fitness values over generations for Spearman Correlation and Binning Diversity. Both metrics improve over time, indicating effective convergence of the symbolic regression algorithm.

The plot of fitness progression across generations highlights consistent improvement in both objectives. The Spearman correlation fitness shows a smooth upward trend, reflecting steady gains in rank alignment, while the Binning Diversity metric exhibits more variability, capturing intermittent improvements in ordinal bin similarity. Overall, these results confirm that the evolutionary algorithm effectively balances and enhances multiple aspects of model quality during the search.

3.3.2 Prediction Phase

For the final prediction task, which aim to validate the rank obtained from the COVID covariates, we selected the best-performing symbolic formula obtained during the validation phase, namely:

$$Q_{\text{psychosocial}} + \left(Q_{\text{vitality}} \cdot \left(-0.527 \cdot \text{mMRCScale} + Q_{\text{psychosocial}} \cdot \left(\text{age} + |Q_{\text{cognitive}}|^{9.585}\right)\right)\right)$$

This formula was applied to the entire training dataset to generate the IC index which was then used together with the healed target variable for training a logistic regression classifier. The model was subsequently evaluated on the test set, where the symbolic formula's output was used as input to predict the healed status.

The predictive performance on the test set is summarized below.

Table 3.2: Confusion Matrix

	Predicted 0	**Predicted 1**
Actual 0	23	51
Actual 1	9	96

Table 3.3: Classification Report

Class	**Precision**	**Recall**	**F1-score**	**Support**
0.0	0.72	0.31	0.43	74
1.0	0.65	0.91	0.76	105
Accuracy		0.665		
Macro Avg	0.69	0.61	0.60	179
Weighted Avg	0.68	0.66	0.63	179

It is important to note that the dataset is imbalanced towards class 1, which affects the classification metrics. The model demonstrates a high recall for the majority class 1, while the recall for the minority class 0 is comparatively lower. Overall, achieving an accuracy of approximately 66.5% is encouraging, especially given the relatively small dataset size and the entirely data-driven approach that does not rely on handcrafted features or domain-specific knowledge.

Conclusions

This thesis set out to reframe the problem of quantifying Intrinsic Capacity through a lens that values interpretability, robustness and practical relevance. Recognizing the complexity and the lack of a unifying theoretical formulation of IC, the work proposed a coherent and transparent pipeline grounded in both clinical reasoning and computational rigor.

The process began with 14 IC-related features, which were discretized into ordinal grades through an optimized binning algorithm tailored to balance information preservation and interpretability. These symbolic evaluations were then aggregated using the Majority Judgment method, a technique from social choice theory chosen for its ability to capture consensus in multidimensional assessments. MJ allowed for the derivation of a robust and relative (but derived from universal validated questionnaires) IC grade, which proved to correlate meaningfully with recovery outcomes in a Long COVID cohort.

To enable the construction of a practical and absolute index, particularly in populations with specific conditions, a symbolic regression model was trained to reconstruct the MJ-based IC ranking using a set of targeted clinical covariates related to COVID-19. The resulting symbolic expression closely mimics the original MJ-derived ranking and was clinically validated by demonstrating that the new ranking effectively predicts patient recovery.

Beyond the clinical domain, the methodology proposed in this thesis offers a generalizable strategy for transforming complex, multidimensional data into transparent and defensible summaries. By combining discretization, consensus-based aggregation, and symbolic modeling, the framework can be applied wherever multiple heterogeneous signals must be synthesized into a meaningful index.

Future developments could involve validating the approach on larger and more diverse clinical datasets. Finally, exploring the use of this pipeline in non-clinical domains, such as education systems, environmental vulnerability indices, or socio-economic assessments, may offer valuable insights into how multi-feature evaluation can be both rigorous and accessible.

Bibliography

Michel Balinski and Rida Laraki. *Majority Judgment: Measuring, Ranking, and Electing*. MIT Press, 2010.

John R. Beard et al. The world report on ageing and health: a policy framework for healthy ageing. *The Lancet*, 387(10033):2145–2154, 2016.

H. A. Güvenir and I. Uysal. An algorithm for optimal discretization of continuous attributes. *Information Sciences*, 168:1–29, 2004.

Riccardo Poli, William B. Langdon, and Nicholas F. McPhee. A field guide to genetic programming. *Lulu Press*, 2008.