Cappadocia 7th International Congress on Scientific Research / August 2-4, 2025 by IKSAD Institute - <https://www.cappadociacongress.org/books> - <https://en.cappadociacongress.org/_files/ugd/d0a9b7_e2b4441287f14d57991624907b0633bb.pdf> (pages in file :13/67/302-303//in book:209-210)

Automated derivation of the analytical value of subjective health assessment data assets I.

Zsolt FEJES (0000-0001-9065-5358)\*

László PITLIK (0000-0001-5819-0319)\*

László, PITLIK (Jr) (0000-0002-8058-9577)\*

Mátyás PITLIK (0000-0002-1991-3008)\*

Diána SZŰCS (0009-0004-7680-6948)\*

János RIKK (0000-0002-3846-6661)\*

Melissza MÉSZÁROS (0009-0006-1391-6220)\*

Róbert GÁNCS (009-0003-6570-3129)\*

Bence LOJEK (0009-0007-1643-868X)\*

\*(MY-X research team)

Keywords: consistency, validation, quality assurance, artificial intelligence, context free characteristics, questionnaire

# Abstract

When collecting data through questionnaires, there is a risk that the data quality may be critically low—i.e., not significantly different from random chance—due to problems affecting the respondent’s intuition. However, the more the questionnaire questions are a substitute for measurement, the greater the chance of generating authentic data. In order to derive meaning from responses to non-measurable questionnaire items, those responses should be modelled in relation to measurement-based items, such as frequency of sport, sleep disturbances, or smoking status (X-array). In addition, all steps of the modelling must also be performed based on a random X-array. In an AI-oriented and automatable modelling-based data asset evaluation, it is expected that the quality indicators of random data assets will perform worse overall than those of real (questionnaire-based) data assets. Further work is needed to define the concept of ideal data assets and to investigate the issues stemming from the dynamic nature of correlation—both of which pose additional risks to traditional data asset evaluation. Data asset evaluation should be primarily content-independent (context free), but the presentation of content-dependent aspects is also an integral part of the specific case study. In the context of the examined data assets, a non-measurement-type X-array can also be constructed based on the subjective assessment of various factors affecting sleep quality. The impact of the two X-array types on the model's goodness of fit will be discussed in a dedicated paper. The result shows that the focused layer of the examined data set has higher quality than a data set generated by random numbers. (<https://miau.my-x.hu/miau/325/abs_operationalized_data_quality_evaluation_process.docx> - accepted)

# Introduction

In this article, based on health data derived from self-reports processed within the framework of Attila Csizmár’s thesis (2025a) (cf. *The study of the prevalence and comorbidity of obstructive sleep apnea (OSA) among the personnel of the Hungarian Defence Forces between 2011–2022 – Semmelweis University*) and the IKEB report (Csizmár:2025b), the general hypothesis of data asset quality assurance is examined: Can data assets obtained from personnel affected in different years be equally reliable in different ways?

Csizmár’s thesis itself (cf. p. 24) explicitly states: “Self-rated health is a subjective indicator of global, general health status, which, according to research findings, has good predictive value in relation to objective indicators of health status” (cf. Hornyák, B. and Ördögh, I. (2015). *Obstructive Sleep Apnea Syndrome (OSAS) among the personnel of the Hungarian Defence Forces II. The prevalence and comorbidity of OSAS among the personnel of the Hungarian Defence Forces.* Hadtudományi Szemle, 8(3), 101–110). When characterizing self-rated health status, participants were asked to respond to the question “How would you rate your general health status?” on a 5-point scale. The possible answers were: 1 = Very poor, 2 = Poor, 3 = Satisfactory, 4 = Good, and 5 = Excellent. Naturally, considering the total number of respondents each year as 100%, the sum of the percentage distribution across the five response categories had to equal 100% every year.

Similar statements appear in Hornyák’s doctoral dissertation (Hornyák, 2022) and thesis (Hornyák, 2011): cf. “Self-rated health is a subjective indicator of global, general health status, which, according to research findings, has good predictive value in relation to objective indicators of health status (Ilder, Benyamani, 1997; Burstroem, Fredlund, 2001). Therefore, in examining the relationship between health status and the Mental Resilience Test (MÁQ), we also examined the development of subjective indicators alongside objective ones (e.g., disease occurrence). According to our research findings, individuals with high MÁQ scores rated their health status significantly better (Hornyák, 2011).” (MÁQ = Mental Resilience Test)

The indispensability of questionnaires is also reinforced by the following work (with 21 mentions of the keyword “questionnaire”): Szilágyi, Zsuzsanna; Csukonyi, Csilla; Sótér, Andrea; Hornyák, Beatrix: *Steps in the introduction of mental resilience assessments in the Hungarian Defence Forces (2006–2009)*, Hadtudományi Szemle 7:1, pp. 158–178, 21 pages (2014) - <https://tudasportal.uni-nke.hu/xmlui/static/pdfjs/web/viewer.html?file=https://tudasportal.uni-nke.hu/xmlui/bitstream/handle/20.500.12944/10608/2014_1_tt_szilagyi.pdf?sequence=2&isAllowed=y>

The practice and culture of analysing time-series questionnaire data is also evident in the authors’ professional networks: cf. Novák, Attila; Hornyák, Beatrix; Rázsó, Zsófia; Szalánczi, Szabolcs; Sótér, Andrea; Juhász, Zsolt; Nyakas, Csaba (2019) – *The introduction of health behaviour profiles in the Hungarian Defence Forces: a cluster analysis of lifestyle factors according to the health screening tests performed in 2011–2015*, INTERNATIONAL JOURNAL OF OCCUPATIONAL MEDICINE AND ENVIRONMENTAL HEALTH 32:1, pp. 99–114, 16 pages – <https://pubmed.ncbi.nlm.nih.gov/30855101/>

Thus, the hypothesis test in this article essentially demonstrates the task of a robot-researcher who, before analysing the data of a given data asset, seeks to objectively (methodically) verify whether the data asset to be analysed conforms to the experiences of Csizmár–Hornyák–Ördögh–Ilder–Benyamani–Burstroem–Fredlund, which, although seemingly universally valid, may not hold true for any specific data asset—i.e., the responses of the given year’s participants may be of a quality similar to that of randomly generated numbers.

The estimation of the value of data assets is carried out not only through the AI-based procedure that is the focus of this article but also with the involvement of Copilot. Although absolute value estimation is not customary, within a benchmarking process, the concept of value can be operationalized. A minimal element of the benchmarking process must be the creation of a random data asset. (A separate article will address the other end of the value scale, i.e., the methodology for constructing the ideal data asset.)

The evaluation methodologies presented in this article must necessarily be context-free, meaning that value judgments must be made based on a sequence of steps fundamentally independent of the phenomena being examined.

# Literature overview

Although the data asset under examination concerns health and lifestyle data of individuals, due to the expectation of content-independence, it is necessary and appropriate to refer exclusively to methodological precedents:

* **Exploration of relationships between questionnaire items**
(cf. László Pitlik: *How can modelling/ChatGPT reveal which two questions in a questionnaire most contradict each other?* – 2025: <https://miau.my-x.hu/miau/319/kerdoiv.xlsx>)
* **Scanning of data asset value within the COCO STEP-IX framework (Component-based Object Comparison for Objectivity)**
(cf. János Angyal: *AI-supported, price/performance-optimized server procurement* – 2024: <https://miau.my-x.hu/miau/315/aj_plus.pdf>)
* **Exploration of quasi-perfect modelability between physical phenomena**
(cf. Marcell Pitlik: *Optical analysis of volumetric combustion and AI-based evaluation* – 2021: <https://miau.my-x.hu/miau/277/szakdoga_PM_final.pdf>, and *Noise-based analysis of burns using artificial intelligence* – 2021: <https://miau.my-x.hu/miau/279/tdk_zaj/TDK_PM_zaj_final.pdf>)
* **Exploration of relationships between complex phenomena through modelling**
(cf. István Németh, Ferenc Szani, László Pitlik: *Comparative analysis of educational materials based on eye movement/emotion/energy/EEG and mouse/keyboard log data* – 2022: <https://miau.my-x.hu/miau/288/2d_3d_complex_study.docx>)
* **Exploration of the potential for deception in questionnaire-based data assets**
(cf. László Pitlik, Marcell Pitlik: *Lie detection in questionnaires* – 2019: <https://miau.my-x.hu/miau/256/torrent/liar-detection-in-questionnaires.docx>)
* **Detection of inconsistency patterns using questionnaires**
(cf. László Pitlik: *Possibilities for inconsistency analysis based on questionnaire responses* – 2017: <http://miau.my-x.hu/miau/232/elegedettsegi_hermeneutikak_v2.docx>)
* **Suspicion generation based on questionnaires**
(cf. László Pitlik: *Nominally scaled interpreted surveys in suspicion generation* – 2012: + <http://miau.my-x.hu/miau/162/93_3.docx>) + <https://miau.my-x.hu/miau/159/93_2.docx>)
* **Lie detection based on questionnaires**
(cf. László Pitlik, Csaba Horváth, László Szabó: *Financial awareness from another perspective, or lie if you can during the survey* – 2011: <http://miau.my-x.hu/miau/158/la158.docx>)

The references cited above, along with their own citations, form an integral part of the general-purpose modelling logic developed around questionnaires as object-attribute matrices. In this logic, each questionnaire response must be estimated based on the other data, and the interpretation of all fact-estimate deviations—either per question or per respondent—in an anti-discriminatory manner leads to the identification of the most critical questions and/or respondents.

In addition to context-free analysis, the issue of content-dependent question relationships also exists and can be addressed using similar logic.

# Experiences with Copilot

A special supplementary layer of the literature background today is the attempt to solve the examined problem using LLM services (e.g., Copilot).

Prompt:

“We are looking for the ideal X data matrix, which consists of 11 rows (objects: O1…O11) and 11 columns (X1…X11). An X matrix is considered more ideal the more accurately it can derive every position in all 5 layers of the Y matrix. Four scenarios have been created (#1, #2, #3, #4), in which the Y matrix is identical across all layers, but the X matrix differs. Question: What is the ideality ranking of the scenarios? Justify the ranking you establish in as much detail as possible.”



**Figure 1: Copilot-based data asset evaluation** (Source: own illustration)

Copilot response (with potential anomalies):

1. place: Scenario#1 = average R²: 0.9936

R² of the 5 layers: [1.0, 0.993, 0.986, 0.997, 0.993] – fact check = 0.9938

2. place: Scenario#4 = average R²: 0.9797

R² of the 5 layers: [1.0, 0.956, 0.977, 0.968, 0.998] – fact check = 0.9798

3. place: Scenario#3 = average R²: 0.9477

R² of the 5 layers: [1.0, 0.871, 0.956, 0.912, 1.0] – fact check = 0.9478

4. place: Scenario#2 = average R²: 0.8281 0.8282

R² of the 5 layers: [1.0, 0.919, 0.744, 0.717, 0.761] – fact check = 0.8282

Figure 1 (without yet detailing the origin of the scenarios) provides a clear guide on how, according to Copilot, the ideality of the relationship between the X-array and the Y-array can be algorithmically assessed. (The numbers/characters marked in red point to a strange anomaly, namely that real calculations are taking place behind Copilot’s opinion, yet they are not consistent - meaning that even the average calculation may be questionable beyond rounding.)



**Figure 2: Copilot response to identical prompt as before** (Source: own illustration)

Comparing Figures 1 and 2 reveals that Copilot is not even consistent with itself: Scenarios 1 and 4 lead to the same R² value, Scenario 2’s R² matches the value calculated from the partial results, while in Scenario 3, the three partial results differ significantly (0.9472 vs. \*77 vs. \*78).

These numerically existing (but quantitatively minor) errors do not change the essence: between Scenarios 1 and 4, Scenario 1 (the questionnaire-based reality) is better, while Scenario 2 (the random number-based benchmark) is worse. Scenarios 3 and 4 are partial results of the ideal X-array to be presented in a separate article, and according to Copilot, they are not better than the questionnaire data realized in reality.

In addition to R², the RMSE values also clearly confirm the same ranking as R².

However, based on correlations (cf. Figure 3):



**Figure 3: Correlations of Y-layers with X-array variables** (Source: own illustration)
the winning (ideal?) scenario is Scenario 4, and Scenario 1 (real) is better than Scenario 2 (random), since the maximum of the average sign-independent correlation values is 0.83, and the minimum is 0.36.

# Data asset

The processed data asset is time-series in nature: the questions cover the years between 2011 and 2022 (11 years in total). Due to technical reasons, the data from 2014 could not yet be processed. The record count for the year 2018 is only a fraction (1:10) of the record counts for the other years — the reason for this discrepancy still requires clarification. However, this quantitative difference necessitated the use of percentage-based data interpretation, meaning that for each of the 11 examined attributes, the total number of respondents per year had to be considered as 100% (cf. Figure 1).

The 11 attributes are linked to 4 questions, each with 2, 3, or 4 response options.

| Label | Response Code | Response Label  |
| --- | --- | --- |
| Physical activity | 1 | Yes |
| Physical activity | 2 | No |
| Frequency of physical activity | 1 | 4–5 times per week |
| Frequency of physical activity | 2 | 2–3 times per week |
| Frequency of physical activity | 3 | Once a week |
| Frequency of physical activity | 4 | Less frequently than the above |
| Smoking | 1 | No, and has never smoked |
| Smoking | 2 | No, quit smoking |
| Smoking | 3 | Yes |
| Do you suffer from sleep disorders? | 1 | No |
| Do you suffer from sleep disorders? | 2 | Yes |

**Figure 4: Structure of the X-matrix** (source: own illustration)

| Variable label | Response Code | Response Label  |
| --- | --- | --- |
| Self-rated health status | 1 | Very poor |
| Self-rated health status | 2 | Poor |
| Self-rated health status | 3 | Fair |
| Self-rated health status | 4 | Good |
| Self-rated health status | 5 | Excellent |

**Figure 5: Structure of the Y-matrix** (source: own illustration)

As shown in Figure 4, physical activity is a binary input, similar to the phenomenon of sleep disorders. Smoking has 3 options, and sports activity has 4 options. Each X-phenomenon is quasi-measurable and largely free from the subjective challenges of classical self-reporting (cf. “How much do you like the “Piréz” people? 1 < 5”). In contrast to the quasi-measurable data generated through self-reported questionnaires, the Y-matrix (Figure 5) typically represents the “explosion” of subjectivity. This is why it is rational to examine the Csizmár–Hornyák–Ördögh–Ilder–Benyamani–Burstroem–Fredlund hypothesis in relation to this specific data asset.

The full data asset covered over 20,000 individuals across the 11 years. While representativeness cannot be claimed for the current study, there is a significant inertia effect — that is, the stability of annual attribute ratios.

The correlation values (cf. Figure 3) once again confirm that there are strong relationships between the X-matrix and the Y-matrix. However, correlation-based analysis loses its time-series nature, meaning that it cannot answer the question: can each year be equally consistent in a different way? In other words: does the honesty or intuitive potential of the ever-changing personnel vary over the 11–12 years under examination?

Prompt: Please provide all details of the linear regression calculations:

How many models were created? What are their coefficients? What are the predicted values?

During the linear regression analysis, a separate model was fitted for each of the 5 Y-variables (y1–y5) in each scenario (4 in total), meaning:

Total number of models:

4 scenarios × 5 Y-variables = 20 linear regression models

**Figure 6: Details of Copilot regression calculations** (source: own illustration)

Figure 6 reveals the structure of the regression calculations: each Y-layer is generated in 11 (annual) repetitions based on the 11 X(i) variables, meaning that even in regression-based calculations, the comparability of year-to-year similarity is lost!

# Methodology

Now that the relevant details of the questionnaire-based data asset are known, and the methodological alternatives available in the literature (without artificial intelligence) are also understood, it becomes clear that classical interpretations are unsuitable for comparing data across years. When evaluating questionnaire data assets, it is not sufficient to interpret only the entire dataset if serious risks may arise along its time-series layers. The hypothesis is therefore simple: \*\*can each year be modelled differently but equally well? This hypothesis primarily requires an AI-based, anti-discriminatory optimization approach (cf. COCO-Y0: <https://miau.my-x.hu/miau/196/My-X%20Team_A5%20fuzet_HU_jav.pdf>). However, as we will see, when modelling based on random numbers leads to surprisingly good partial results—and consequently, modelling errors in the real data asset are hard to find—it may be unnecessary to "use a cannon to shoot a sparrow."

In the follow-up articles, where additional scenarios will be developed to explore the nature of the ideal data asset, anti-discriminatory modelling will become indispensable—just like addressing the risks associated with interpreting correlation derived from regression coefficients or understanding non-measurement-type X-matrices.

According to Copilot (cf. Figure 6), 20 regression models were created: 4 scenarios × 5 Y-layers = 20 data constellations. In these OAMs (object-attribute matrices), the 11 years are not interpreted as a time series. Only the 20 OAMs have indicators resulting from the regression steps. The situation is similar for correlation values (cf. Figure 3), where even the Y-layers do not need to be separated. In regression calculations, the average of the fact vs. prediction correlations (R² values) per Y-layer is used to evaluate the scenarios. This, incidentally, raises the question: why were regression calculations needed at all, especially with the assumption of linearity? In other words: instead of the arbitrariness of linearity, would it have made sense to choose another functional form? And more broadly: does a module exist in human knowledge that can determine the best regression form? (Answering this meta-level question would again require a separate article…)

Regarding the hypothesis "can each year be modelled differently but equally well," a modelling framework is needed that is more flexible than the logic of regression calculations. Step functions are such structures, as instead of the “a” parameter in the regression form “Y = a\*X + b,” a separate parameter is assigned to each rank level of the input signals. That is, the actual magnitude of the input data is not processed—only their rank order is used as the control signal. (Whether this leads to actual information loss is again a topic for a separate article…)

Why is a more flexible modelling framework necessary?

First, because in regression, the “a” values for each input variable are signed numbers—meaning the regression model itself decides what kind of causal mechanism to assume between Xi and Y. In contrast, with step functions, this decision is made by the modeler. And it is precisely this decision that leads to one of the promised follow-up articles, where the pitfalls of using Y vs. Xi correlations (i.e., direct or inverse proportionality) as control signals will be explored.

Second, the more parameters a modelling framework uses (cf. artificial neural networks), the greater the chance that predictions and actual values will match. This is necessary because, in the case of raw data composed of random numbers, a rigid modelling framework (cf. regression, especially in its linear form) will likely produce incorrect predictions for almost every observation (year)—even for a later-to-be-discovered ideal data asset. In such cases, it is inconceivable that all real and random-number-based OAM scenarios would lead to perfect fact vs. prediction matches. With the flexibility of step functions, however, the chance of this happening—as will soon be shown in detail—is high.

This article therefore presents the results of a parallel analysis in which the first and second OAM scenarios from Figure 1 are modelled using five step functions each, corresponding to the five Y options. The expert-level ranking parameters (direction vectors) used for the raw data enforce the following direct or inverse proportionalities instead of the free sign parameters of regression calculations:

| Category | Question | Option Code | Option | Expert’s Direction | Correlation |
| --- | --- | --- | --- | --- | --- |
| Question | Physical activity  | 1 | Yes | 0 | 0.93 |
| Question | Physical activity  | 2 | No | 1 | -0.93 |
| Question | Frequency of physical activity  | 1 | 4–5 times per week | 0 | 0.88 |
| Question | Frequency of physical activity  | 2 | 2–3 times per week | 0! | -0.94! |
| Question | Frequency of physical activity  | 3 | Once a week | 0 | 0.46 |
| Question | Frequency of physical activity  | 4 | Less frequently than the above | 0! | -0.08! |
| Question | Smoking | 1 | No, and has never smoked | 0 | 0.43 |
| Question | Smoking | 2 | No, quit smoking | 1 | -0.37 |
| Question | Smoking | 3 | Yes | 1 | -0.06 |
| Question | Do you suffer from sleep disorders?  | 1 | No | 0 | 0.85 |
| Question | Do you suffer from sleep disorders?  | 2 | Yes | 1 | -0.85 |

Figure 7: Expert-defined direction vector (Source: own illustration)

Figure 7 contains two critical points for the attentive reader: - x4 – exercising 2–3 times per week, where the expert direction code is 0, meaning that the higher the proportion (%) of respondents in this category, the higher the proportion should be in any Y-layer. The correlation values refer to Y5, the layer representing the proportion of “excellent” health (cf. Figure 3, top layer, Y5 columns for variables x3–x6). In the case of this critical variable, the correlation contradicts the expert expectation of direct proportionality (direction code = 0), as it is strongly negative (–0.94). The other critical variable is x6, which represents one end of the scale.

Why was the correlation for x4 ignored? The answer is simple and partially context-dependent: If an element in a four-option scale that is not at an extreme position behaves differently from its neighbouring elements, it must be considered illogical (cf. polynomial vs. linear regression interpretability). If x4 is not at the edge of the option list (1–2–3–4), then there is no reason for it to behave differently from the other three elements.

In the case of physical activity being represented as a binary variable and/or sleep disorder having a similarly binary structure, one option is the opposite of the other. This is also reinforced by the correlation values in both cases—necessarily so, since knowing one value in a binary setup forces all further mathematical or logical operations onto a constrained path due to the rule that 100 minus one equals the other.

In the case of binary variables like physical activity or sleep disorder, one option is the opposite of the other. This is reinforced by the correlation values in both cases—necessarily so, since knowing one value in a binary setup automatically determines the other (100 – one = the other), which constrains all further mathematical/logical operations.

Why doesn’t this constraint apply to x6 (sport frequency)? The answer is not simple. One might assume that options 1–2–3 reflect direct proportionality, and based on the rule 100 – first – second – third = fourth, option 4 (x6) could reflect inverse proportionality. BUT: None of the four levels of sport frequency clearly represent a denial of the phenomenon (unlike binary cases). Therefore, if x4’s correlation must be overridden, then x6 must inherit the homogeneous direction code of the other three options.

Why focus on correlations with Y5? Even though Figure 3’s top layer provides correlation values for x3–x6 across Y4–Y1 as well, the answer is relatively simple: If Y5 is considered an extremely positive (excellent) value, then all others are its opposite. Thus, only the aggregate of the remaining layers matters, which—based on the 100 – p1 = p2 rule—is constrained. So, it is sufficient to use Y5 as the reference.

Is there a contradiction in the smoking variable with three options? The answer is simple: no. Why? Because among the three smoking options, “never smoked” is the opposite of the other two, which are gradations of smoking. Therefore, if the direction code for “never smoked” is 0 (supported by correlation), then the other two (its opposites) must necessarily have a direction code of 1, which is also confirmed by the correlations.

The above reasoning justifies the claim of partial context-independence: No logical operation used keywords like “sport” or “smoking”—only the number of options and their oppositional nature were considered. This logic could apply to countless other phenomena, even though some domain-specific interpretation still occurred.

The parameters for the step functions were derived using the \*\*MY-X FREE COCO online service: <https://miau.my-x.hu/myx-free/> + <https://miau.my-x.hu/myx-free/coco/index.html>

# Results

Based on the real data, the flexible step functions allowed for zero prediction errors across all five Y-layers for every year—fully meeting expectations derived from high correlations and the quality of the data asset (cf. <https://miau.my-x.hu/miau/326/s1c1/s1c1_overview.xlsx>).

In contrast, when modelling with random data assets, one would expect to detect modelling errors, especially when expert-defined direction codes (cf. methodological section) are strictly followed. However, even when the X-matrix was generated using random numbers (while respecting the structural rules of 2–3–4-option sets), it turned out that all five model layers could still be predicted flawlessly.

This means that, contrary to the validity and consistency ranking suggested by Copilot’s regression-based (R²-oriented and/or error-oriented) and correlation-based approaches, the more flexible modelling frameworks highlight that the questionnaire-based data asset—when constrained by expert opinion—does not inherently possess a trivial value advantage over the randomly generated data asset.

The suspiciously (naively) clear-cut conclusions found in the literature were challenged by the correlation-based perspective, where one of the supposedly ideal scenarios (#3 or #4) outperformed both the real questionnaire-based scenario (#1) and especially the random-number-based scenario (#2) (cf. Figure 3). Meanwhile, in the regression-based evaluation, scenarios #3 and #4 were positioned between the real and random scenarios.

However:

While the issue of sign direction in regression coefficients still warrants a dedicated article, at this point the process of validating the value of the real questionnaire-based data asset must be extended with further steps (cf. Appendix: Copilot’s self-guided error correction logic as experienced in this example).

The next logical step is to proceed with modelling using dynamic direction codes, meaning that the direction code for each variable in the fixed X-matrix is determined by the sign of its correlation with the current Y-layer. All other methodological steps remain unchanged.

Resulting fact vs. prediction outcomes:

In the case of the real (questionnaire-based) model series, the modelling remains flawless for Y5, Y4, Y3, and Y2 layers. However, for Y1, the first antagonism appears in the X-matrix—this refers to a situation where, for a pair of years (objects), one year’s data is better or at least not worse across all variables compared to another year (cf. 2011 vs. 2021), yet the comparison of Y-values does not reflect this systematic advantage in the X-matrix.

This type of antagonism typically arises in object-attribute matrices (OAMs) when comparing row indices. Atypical patterns include comparisons based on the actual values of row indices. In this case, for variable x5, for example, 2011 ranks 9th and 2021 ranks 7th—so in all other respects, 2011’s Y1 should be higher than 2021’s, but in reality, Y1 for 2011 is lower than for 2021.

In the case of the random-number-based X-matrix, antagonisms appear in three Y-layers (Y1, Y2, Y3), while Y4 and Y5 still yield flawless fact vs. prediction pairs for every year. In Y2 and Y3, one typical and one atypical antagonism can be identified, while in Y1, the correlation between prediction and actual values drops sharply to 0.54 (cf. Y2–Y3: correlation = 0.99, Y4–Y5: correlation = 1.00).

# Discussion

The regression- and correlation-based calculations found in the literature (and followed by Copilot) are naive and lack objectivity. While they may appear to yield trivial and reassuring results in comparing the quality of real (questionnaire-based) and random raw data scenarios, they are fundamentally limited. These methods do not allow for meaningful year-to-year comparisons, nor can they determine the ideal X-matrix. This is evident from the evaluation of test scenarios using MS Excel Solver (#3: all x1–x11 correlations equal to 1.00 for y5 only; #4: maximize the average absolute correlation across all Y-layers), where the results differed significantly depending on whether regression- or correlation-based logic was applied. In other words, there is no universally accepted definition of an ideal X-matrix in the literature.

Using direction codes as dynamic control signals is not foreign to regression modelling—in fact, the concept of regression itself underpins the correlation-dependence of direction codes. Therefore, replacing fixed expert-defined direction codes with dynamic ones can be seen as a sensitivity analysis of the hypothesis: can each year be modelled differently but equally well?

Just like the challenge of defining the ideal X-matrix, there is a challenge at the other end of the data asset quality scale: what does the most random randomness mean? After all, even the real (questionnaire-based) X-matrix—or a future ideal version—can be generated randomly. Regarding randomness, it is worth noting that, for example, the entropy of the universe is not at its most extreme when matter/energy is evenly distributed—because homogeneity itself is a kind of pattern.

Whether dealing with an ideal or a random X-matrix, one could theoretically generate alternatives using brute-force (combinatorial) methods. At that point, virtually infinite alternative X-matrices would need to be evaluated based on modelling outcomes. Anti-discriminatory modelling is capable of converting the ideality of both ideal and random patterns—or all perceived X-matrix variants—onto a single scale, using function symmetry and self-validation. On this scale, the preferred model outcomes represent one extreme (ideal), and their mirror image represents the other (random). All other patterns represent normative behaviour. For real (questionnaire-based) data assets, it is expected that they fall above the norm—i.e., they should exhibit ideal characteristics.

Within this self-validation framework, depending on the size of the OAM (object-attribute matrix), a fourth interpretation category may also emerge: non-validatable X-matrix patterns, which can occur anywhere between the extremely ideal and the seemingly random. These indicate that, even with sufficiently flexible modelling frameworks, it may be risky to make definitive claims about their value given a specific quantity and content of data.

The modelling framework used here (COCO-STD, for production function generation) can be further enhanced with the COCO MCM (exploratory) method, which becomes necessary when direction codes cannot be logically determined. As demonstrated by the Y1 model driven by a dynamic correlation-based control signal (which resulted in a correlation of 0.54), larger OAMs and/or more complex X–X relationships will produce low-correlation prediction error patterns across all Y-layers. Their relative ideality or advantage can once again only be transformed into a scale using COCO Y0 (anti-discriminatory) modelling.

# Conclusions

The naivety of the solutions found in the literature—namely, the inability to formulate an anti-discriminatory counter-hypothesis (can each year be modelled differently but equally well?)—in itself reveals the methodological arbitrariness suggested by the literature. This arbitrariness (naivety) is further exposed by the lack of a clear definition of the ideal and the contradictory evaluations of potential ideal X-matrices. So, to the question of whether we can be satisfied with the foundations provided by the literature (cf. Copilot’s responses), the answer at a higher level of complexity is: NO.

The fact that it is possible to perfectly predict non-measurement-oriented questionnaire responses (y1–y5) based on random numbers should be sobering for all analysts and readers familiar with classical statistical interpretations (and especially for those who are not). In other words, the logic of sensitivity analysis, which is common in economics, should also be extended to the world of modelling and hypothesis interpretation: in this case, without knowing the results produced by the same methodology applied to identically sized but random data assets, one should not make claims about the value of real data assets.

A specific slice of this expectation is the conversion of the concept of the ideal to a Knuth-level formalism (cf. [https://miau.my-x.hu/miau2009/index\_tki.php3?\_filterText0=\*knuth](https://miau.my-x.hu/miau2009/index_tki.php3?_filterText0=*knuth)).

The potential value of real data assets—which may appear equivalent to randomness—is vividly illustrated within the framework of correlation-driven modelling by the emergence of antagonisms. Random data assets, when viewed from the right angle, exhibit classic/typical antagonisms—something that a real or valuable data asset should not display. Of course, atypical antagonisms are also risk factors, but the explanation of their occurrence belongs in the article that aims to define the ideal. For now, it is enough to declare that in the case of a Y-matrix with internal structural constraints (i.e., the sum of layers y1–y5 is always 100%), the existence of a constant ideal X-matrix—or rather, the existence of many alternative idealities instead of a single ideal state—must be examined as a hypothesis, so that we do not fall back into the world of naive/arbitrary interpretations.

The production-function-based, automatable procedure presented here can and should be extended with the validation, anti-discriminatory, and exploratory layers/modules mentioned in the discussion, in order to increase universality—because the concept of GOODNESS can and should be made infinitely complex.

# References

Csizmár, A (2025a). Az obstruktív alvási apnoe (OSA) gyakoriságának és komorbiditásának vizsgálata a Magyar Honvédség személyi állománya körében 2011-2022 között, Diplomamunka, Semmelweis Egyetem

Csizmár, A (2025b). Beszámoló kutatási eredményekről - Az obstruktív alvási apnoe (OSA) gyakoriságának és komorbiditásának vizsgálata a Magyar Honvédség személyi állománya körében 2011-2022 között, MHEK-IKEB-jelentés

Pitlik László (2025): Hogyan leplezhető le modellezéssel/ChatGPT-vel, hogy egy kérdőív mely két kérdése mond leginkább ellent egymásnak? – <https://miau.my-x.hu/miau/319/kerdoiv.xlsx>

Angyal János (2024): Szerverek MI-támogatással optimalizált ár/teljesítmény-alapú beszerzése - <https://miau.my-x.hu/miau/315/aj_plus.pdf>

Németh István, Szani Ferenc, Pitlik László (2022): Tananyagok összehasonlító elemzése szemmozgás/érzelem/energia/EEG és egér/billentyűzet-log-adatok alapján - <https://miau.my-x.hu/miau/288/2d_3d_complex_study.docx>

Hornyák, B. (2022): A mentális állóképesség mérésének és fejlesztésének lehetősége a Magyar Honvédség hivatásos és szerződéses személyi állományának körében, PhD-értekezés, Nemzeti Közszolgálati Egyetem, Hadtudományi Doktori Iskola, <https://hdi.uni-nke.hu/document/hdi-uni-nke-hu/Hornyak_Beatrix_doktori%20%C3%A9rtekez%C3%A9s.pdf>

Pitlik Marcell (2021): Térfogati égés optikai vizsgálata és mesterséges intelligencia alapú elemzése - <https://miau.my-x.hu/miau/277/szakdoga_PM_final.pdf>

Pitlik Marcell (2021): Égések zaj alapú elemzése mesterséges intelligenciával - <https://miau.my-x.hu/miau/279/tdk_zaj/TDK_PM_zaj_final.pdf>

Pitlik László, Pitlik Marcell (2019): Hazugságvizsgálat kérdőívezés keretében - <https://miau.my-x.hu/miau/256/torrent/liar-detection-in-questionnaires.docx>

Novák, Attila; Hornyák, Beatrix; Rázsó, Zsófia; Szalánczi, Szabolcs; Sótér, Andrea; Juhász, Zsolt; Nyakas, Csaba (2019) - The introduction of health behavior profiles in the Hungarian Defense Forces: a cluster analysis of lifestyle factors according to the health screening tests performed in 2011–2015 - INTERNATIONAL JOURNAL OF OCCUPATIONAL MEDICINE AND ENVIRONMENTAL HEALTH 32: 1 pp. 99-114., 16 p. - <https://pubmed.ncbi.nlm.nih.gov/30855101/>

Pitlik László (2017): Inkonzisztencia-vizsgálatok lehetősége kérdőíves válaszok alapján – <https://miau.my-x.hu/miau/232/elegedettsegi_hermeneutikak_v2.docx>

Hornyák, B. és Ördögh, I. (2015): Obstruktív alvási apnoe szindróma (OSAS) a Magyar Honvédség személyi állománya körében II. Az OSAS gyakorisága és komorbiditása a Magyar Honvédség személyi állományának körében. Hadtudományi Szemle, 8(3), 101-110., <https://epa.oszk.hu/02400/02463/00027/pdf/EPA02463_hadtudomanyi_szemle_2015_02_099-111.pdf>

Szilágyi, Zsuzsanna; Csukonyi, Csilla; Sótér, Andrea; Hornyák, Beatrix (2014): A mentális állóképesség-vizsgálatok bevezetésének lépései a Magyar Honvédség állományában (2006-2009) HADTUDOMÁNYI SZEMLE 7 : 1 pp. 158-178. , 21 p. - <https://tudasportal.uni-nke.hu/xmlui/static/pdfjs/web/viewer.html?file=https://tudasportal.uni-nke.hu/xmlui/bitstream/handle/20.500.12944/10608/2014_1_tt_szilagyi.pdf?sequence=2&isAllowed=y>

Pitlik László (2012): Nominalskaliert interpretierte Befragungen in der Verdachtsgenerierung I/II – <http://miau.my-x.hu/miau/162/93_3.docx> + <https://miau.my-x.hu/miau/159/93_2.docx>

Hornyák, B. (2011): Mentális állóképesség összefüggése az egészségmagatartással, Diplomamunka: DEOEC Népegészségügyi Kar Egészségpszichológiai Szak

Pitlik László Horváth Csaba, Szabó László (2011): Pénzügyi tudatosság más szemmel, avagy hazudj, ha tudsz a kérdőívezés során – <http://miau.my-x.hu/miau/158/la158.docx>

# Annex

Details: <https://miau.my-x.hu/miau/326/s1c1> + Copilot - self-fixing experiments:

Technical Issue: It appears that during raw data processing, an error occurred: the number of columns in the data did not match the number specified in the header. This likely happened because the data was tab-delimited, and the first column (e.g., o1, o2, etc.) was not handled properly.

To resolve this, I will restructure the data processing to include object identifiers and ensure that each row aligns exactly with the header. I’ll fix this and rerun the file generation.

However, the error persisted: the number of columns in the data still didn’t match the header. This is likely because the first column (object ID) wasn’t included in the header.

Next step:

✅ I will include the first column in the header as well, so each row will contain exactly 17 columns (1 ID + 11 X + 5 Y).

I’ll fix this and rerun the file generation. The file creation still failed because the first column (e.g., o1, o2, etc.) was not in the header, so the rows had 16 values while the header had only 15 columns.

📌 Solution: The header must be extended with the first column, e.g., obj\_id, so the full header will be: obj\_id, x1, x2, ..., x11, y1, y2, y3, y4, y5

I’ll now make this correction and try generating the Excel file again.

Eventually, the file creation failed because the system tried to open an external file (scenarios\_data.txt) that doesn’t exist — this was a technical error during processing.

📌 Solution: I will now embed the data directly from the conversation and not reference an external file. This will ensure the Excel file is successfully created.

I’ll restart the process with the correctly embedded data.

Unfortunately, the file creation still failed because the header had 16 columns, while the data rows had 17 — the first column (e.g., o1, o2, etc.) was treated as data but wasn’t included in the header.

📌 Final fix: I will explicitly insert the obj\_id column into the header.

Each data row will then match the 17-column header exactly. I’ll now apply this fix and generate the Excel file. You were redirected to the Microsoft 365 Copilot app, making it easy to access Copilot and agents. Update your bookmark to the new URL: m365.cloud.microsoft/chat.

Communication was interrupted…

Why is it problematic to return the details if the results of those details were already communicated earlier?

Would you like me to now generate the Excel file with all 20 regression models and their details?