Theses of Doctoral (PhD) Dissertation

DYNAMIC MODELLING OF THE HUMAN MILK COMPOSITION AT MOLECULAR LEVEL AS A FUNCTION OF ENVIRONMENTAL FACTORS

Tünde Vámosiné Pacza

PhD candidate

Dissertation supervisor:

József Baranyi, PhD

UNIVERSITY OF DEBRECEN

DOCTORAL SCHOOL OF NUTRITION AND FOOD SCIENCES

Debrecen, 2025

1. BACKGROUND AND OBJECTIVES OF THE DOCTORAL THESIS

Although the impact of diet on health was recognized for over thousands of years, nutritional databases still do not list the significant proportion molecules that can be found in food. *(Hooton és mtsai., 2020)*.

Human milk, our first and most important food, is essential for the growth and development of the newborn immediately after birth and an irreplaceable source of nutrition for the survival of the infant. *(Rossum és mtsai., 2005; Agostoni és mtsai., 2009)*, is a complex biological system of nutritional and bioactive components which are in constant interaction with each other. *(Christian és mtsai., 2021; Samuel és mtsai., 2020)*. In recent decades, there have been extensive research efforts to understand the components of human milk and the factors influencing their composition, but data generation and collection, statistical and modelling methods, and interpretation of results rarely focus on the dynamic nature of the 'mother-milk-infant' system. *(Ballard & Morrow, 2013; Perrella és mtsai., 2021; Sánchez és mtsai., 2021; Carr és mtsai., 2021; De Weerth és mtsai., 2022; Samuel és mtsai., 2020)*.

To the best of our knowledge, no quantitative statistical analysis has yet been performed to evaluate and prioritise the causes of heterogeneity in human milk composition. One explanation for this is the lack of longitudinal data on human milk composition *(Christian és mtsai., 2021; Shenhav & Azad, 2022)*, especially for individual mothers. One aim of our research was to highlight how this knowledge gap can be filled by creating a database *(Pacza és mtsai., 2022)* to serve as a recommended template; a database with ontology and data for dynamic modelling.

This research focused on mapping the composition of human milk at molecular level and mathematical modelling its changes over time. The primary objective was to create a database of published data from scientific research and then to use this database to develop a model of the dynamics of human milk components over time.

The first part of this thesis describes the MilkyBase database, which contains data on the biochemical composition of human milk. The data were extracted and digitised from scientific publications, partly by machine learning and partly manually. Our primary goal was to define an ontology to support the understanding of how the composition of human milk depends on different factors. A secondary goal is to provide an ontology that

can be used for research on other types of food, where users can store and publish their own data in a similar format.

The second part of the thesis uses data from the MilkyBase database to demonstrates how predictive modelling methods, already widely used in food microbiology, can be applied to human milk research.

The aim was to use the model to gain a more accurate picture of the temporal and individual variation in human milk components and the factors that most influence them, to helping design further research on human milk components. In addition to presenting the method, the final part of the paper will also highlight the potential limitations and potential applications of 'predictive breast milk research'.

2. MATERIAL AND METHODS

The methods used in the research can be divided into 3 main stages. Firstly, an ontology for capturing human milk components was defined and the database was populated with information based on publications, then the core database was extended using advanced search, and finally the obtained data was analysed and modelled.

Since our goal was to create an ontology that could be a useful tool for nutrition professionals and researchers, as well as for industry and regulation, we had to make a number of compromises during the ontology definition process to find a balance between Big Data's four main pillars - volume, velocity, variety and veracity.

Based on these principles, Microsoft Excel, a powerful spreadsheet program, was chosen to capture the data into a database. It is the most widely used and best-known application, capable of linking spreadsheets and offering data visualisation and analysis tools. In addition, its functions can be extended using the object-oriented programming language Visual Basic for Applications (VBA). The VBA utilities we have created allow for input validation during record capture, and assist in data analysis (for example, comparing your own and others' results of similar published data), encouraging data curators to record relevant data in our database. This is the realisation of the so-called wiki philosophy, i.e. adding knowledge to the commons, which can potentially lead to a much larger database.

2.1. MILKYBASE ONTOLOGY

In our data collection process, we first defined the ontology. (Figure 1.).

For a start, we searched for relevant publications (**Literature search**) on human milk composition. Using FoodMine, a natural language processing algorithm, we got a list of publications, then the result was refined and extended with manual methods. (**Analyse selected source**) The collected papers were read and scored by us, human reviewers. The main selection criteria were that the publication should contain quantitative data on the human milk components; what is more, so-called dynamic data, showing the variation of the components over time. These ideally, were organized in tables which were easier to integrate into the Excel spreadsheets we chose to store the data in.

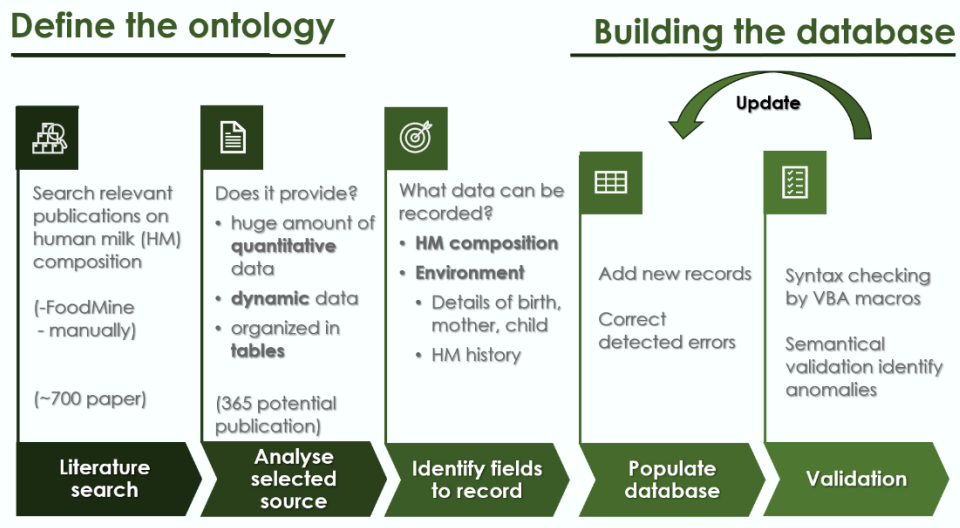

Figure.1. The process of creating MilkyBase database

Finally, we manually evaluated the selected publications to identify appropriate datasets for recording. (**Identify fields to record**) As our aim was to collect as much information as possible, we gathered data not only on the human milk composition but also on the conditions under which the data were produced.

While populating the database continuous data curation and **validation** were performed. On the one hand, macros for syntax checks were developed; on the other hand, semantic validation, was used to identify anomalies in the publications. The latter quality check is much more difficult than the syntax check, as semantics requires human intelligence.

MS Excel VBA macros are used to check the syntax of the database. The MBmacros.xlsm file containing the macros is available to any user in the Figshare repository. By periodically running a "Syntax check" validation macro during data capture, it is possible to ensure that the recorded data is entered in the required format of the fields, thus facilitating data analysis.

During the semantic check we also found many ambiguities or inconsistencies in the terminology used by the authors. In these cases, we have defined and quantified the published data to the best of our knowledge.

2.2. EXTENDING THE CORE MILKYBASE DATABASE

In the second phase of the research, additional data were added to the core MilkyBase database for more accurate data evaluation. When expanding the database, the literature search focused on the regional situation of breast milk research and added data from two priority regions, Asia and South America.

The literature search was conducted using SciELO - Scientific Electronic Library Online (SciELO) to search Scopus, PubMed, Web of Science (WOS) online publication databases.

Out of 190 publications found in the literature search for database expansion, 23 were finally found to be suitable for expansion of our database. As we placed a strong emphasis on searching for dynamic data, our database expansion has multiplied the number of dynamic data by almost one and a half times. (Figure 2.)

MilkyBase version	Publikáció	Rekord (Master lap)			Összetevő	DynVal
		Összesen	Statikus	Dinamikus		
Core (1.0.)	**140**	**840**	205	635	**752**	**7666**
Adatbázis bővítés	**23**	**54**	11	43	**71**	**2965**
2.0.	**163**	**894**	216	678	**823**	**10631**

Figure 1. Records of the MilkyBase database after the data extension

2.3. ANALYSIS AND MODELLING THE TEMPORAL TRAJECTORIES OF HUMAN MILK COMPONENTS

In addition to creating an ontology, the aim was to use the data collected in the database to build a model that describes the trajectories of human milk components during the critical first four months of human life, when infants' nutritional needs ideally come primarily from human milk.

Our primary model was developed as follows:

The focus interval was the first four months. During this time, the $y(t)$ concentration of a human milk components was assumed to follow **a two-phase model**.

1. In the first, initial phase (colostrum), the HMC concentration changed with time in a linear in time interval [0, λ],
2. followed by an exponentially convergent saturation phase:

(1.)
$$y(t)=\begin{cases} y_0 + a \cdot t & (0 \le t < \lambda) \\ y_\lambda \cdot e^{-r\cdot(t-\lambda)} + y_{End}(1 - e^{-r\cdot(t-\lambda)}) & (\lambda \le t) \end{cases}$$

where $y_\lambda = y_0 + a\cdot\lambda$, $0 \le r$, $0 \le \lambda$.

Where:

- ***y(t)*** : concentration of a human milk component as a function of time from birth
- ***t*** : post-partum time (*t=0* is the time of birth)
- y_0 : concentration of a human milk component at the time of delivery (initial concentration)
- ***a*** : the rate of change in the concentration of the human milk component in the initial phase
- $y_\lambda = y(\lambda)$: the rate of change in the concentration of the human milk component at the end of the initial phase
- ***r*** : saturation rate
- λ : length of initial (colostrum) phase
- y_{End} : the final concentration level

The initial parameters, y_0 and ***a***, as well as those of the second phase, ***r***,, and the final concentration level y_{End} depend on various factors, primarily on some characteristics of the mother.

The second, convergent phase is called the **saturation model**.

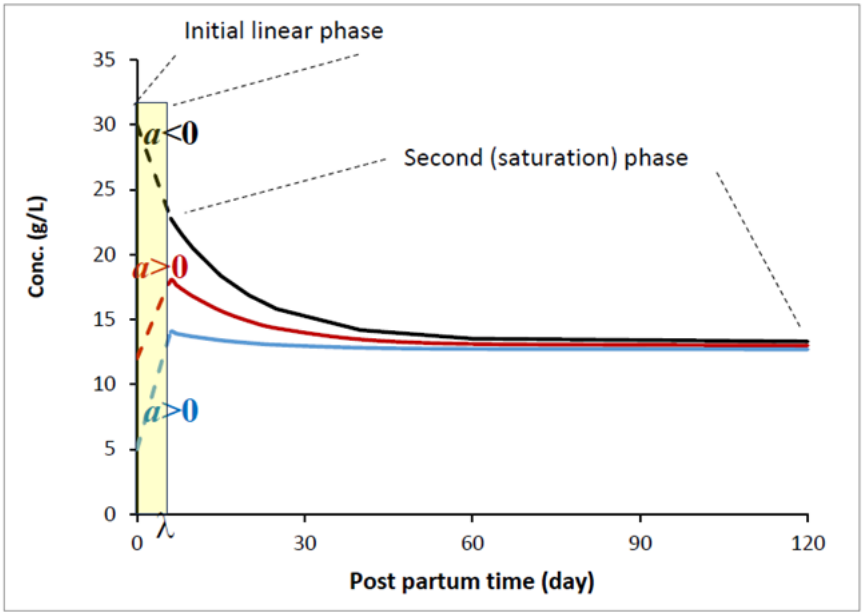

Figure 2. *Two-phase saturation model.* Broken line: initial phase (colostrum). Continuous line: second (saturation) phase.

. The lowest blue curve is hardly different from the ***r***=0 case, which would produce a **bi-phasic function**: linear change in the colostrum, followed by a constant ($y_{End}= y_\lambda$)) human milk concentration:

$$\textbf{(2.)} \qquad y(t)= \begin{cases} y_0 + a \cdot t & (0 \le t < \lambda) \\ y_\lambda & (\lambda \le t) \end{cases}$$

$$where\ y_\lambda = y_0 + a \cdot \lambda\,, \quad r \geq 0, \quad \lambda \geq 0.$$

The upper curve is hardly different from the λ=0, ***a***=0 case, which would represent **the single-phase pure saturation model**, with three parameters.

$$\textbf{(3.)} \qquad y(t)= y_0 \cdot e^{-r \cdot t} + y_{End}(1 - e^{-r \cdot t}) \qquad (0 \le t, r \ge 0)$$

F-test decided whether the full two-phase model with four parameters (y_0, ***a***, ***r***, y_{End}) were needed to describe a dataset, or any of the parameters could be a fixed value, to decrease the dimension of the model. Note that value set of λ was considered binary: it is either 6 (default) or 0. In the latter case, the first phase is embedded in the saturation model, and the result is a single-phase, pure saturation model, with three-parameters.

Again, F-test decided if the single-phase pure saturation model was sufficient to fit a particular dataset or the λ=6 case was significant with a slope a.

The secondary model could be used to quantify the effect of various factors, mother history and other characteristics on the fitted parameters (e.g. y_{End}). Similarly, F-test was used to decide whether one or two of the parameters can be considered identical, for a pair of datasets.

The non-linear regression algorithm was built in a bespoke Microsoft Excel Add-In, written in Visual Basic, implementing the standard Levenberg-Marquardt method *(Levenberg, 1944) (Marquardt, 1963)*. The Data Analysis Add-In of Excel was used to carry out linear regression and ANOVA procedures, with 5% significance level.

3. RESULTS

As a result of this research, we created the MilkyBase database, which stores milk composition records in linked Excel spreadsheets, and then used those data to perform mathematical modelling.

The generated database is an Excel workbook that makes data recording easy for nutrition researchers with non-computing backgrounds. The hierarchical organisation of fields provides programmability of statistical and graphical methods for data analysis. This is also supported by the fact that the database and its description, as well as the Excel macros created for it, are all available for free and free of charge and can be modified as required.

3.1. NOVELTY OF MILKYBASE DATABASE

The main novelty of MilkyBase is that it focuses both on the conditions under which the measured data are generated and the effect of these conditions on the human milk composition and focuses on the dynamics and uncertainty characteristics of these data, which are entered in the explanatory and response variables.

The novelties are the follows:

3.1.1. Developing the MilkyBase ontology with the explanatory variables (i.e., the conditions) and the response variables (i.e., the components) in mind

Our database development is based on the principle that a record is viewed as a representation of the various explanatory conditions under which observations were made, representing the composition of human milk as a response variable. We define the variables that describe the conditions under which the data were generated as Explanatory Variables, while the measured data for the breast milk components are considered Response Variables. Explanatory variables describe the details of the mother, child and birth (mother's age, diet, child's weight, week of gestation, etc.) and the so-called "history" of breast milk (geographical location, measurement and storage methods, etc.), while response variables are the quantitative data on breast milk components (nutrients, bioactive components, etc.).

The fixed values of these response variables can be in so-called "extended numeric" format or dynamic, time-varying, where the change of the components over time is stored

in a table and the indicator that points to the table is the input value of the variable. The numeric values of explanatory variables (condition fields) can also be recorded in a similar way to response variables, but not necessarily only numeric values. They may also be Boolean values or category lists.

3.1.2. Dynamic variables, using the technique of "pointers to a table"

The time-dependent explanatory and response variables [time, value] are represented by data tables, while the respective entry in the main tab of the database is just a pointer to this table. The derived parameters of the trajectories that fit these time-dependent data, such as velocity or steady-state level, are possible scalar representations of the trajectories. In other words, the temporal profile of a variable is described by a few key parameters (primary model), while the variation of these parameters as a function of the conditions (terms) affecting the response can be described by secondary models. This data structure also allowed predictive modelling of breast milk components.

The multivariate dynamic responses captured in a single record facilitate the visualisation of time-varying data and thus their comparison and analysis, for example to detect patterns and outliers in the data; to identify data gaps or possible errors.

3.1.3. Extended numerical variables provide the possibility to measure uncertainties

To incorporate the uncertainties of measurements to the database, we introduced an extended definition for the concept of *"numerical field*" of the database. Its default format is that of an ordinary real number, a certain centroid value of the available relevant data. This can also be supplied with a quantification of the spread of those data. This is commonly either their standard deviation or their quartile range. . An additional second part, separated by a semi-colon, may also be recorded about a prediction (or estimate) of the real mean. . This can also be supplied with uncertainty quantification, which is commonly either the standard error of the estimate, or its 95% confidence interval.

If the entered data represent intervals, then a stochastic interval-analysis can be used in subsequent calculations, which is more powerful than calculations with deterministic values, as not only quantitative conclusions can be derived but the confidence in those conclusions can also be quantified.

3.1.4. Tree structure, which is the general data structure

In our database, a "Component" is considered to be a molecule or group of molecules. Both the molecule C20:4n-6 ("Arachinodic acid") and "fatty acid" are components, but while the first is a molecule, the second is a molecular group, of which the molecule C20:4n-6 is a component. This type of grouping follows a hierarchical tree structure. Thus, not only the concentration of a particular molecule, but also the concentration of any molecule group from the level next to the mother molecule root can be recorded as the concentration of any component.

In the MilkyBase database, not only do the components form a hierarchical tree structure, but the whole database is characterized by this structure.

3.1.5. Data can be recorded in direct and indirect (derived) form

Besides the measured concentrations of components, many authors publish only the transformed or derived values for a component, i.e. either measured concentrations (g/L) or ratios compared to the group containing the component.

In order to handle such cases explicitly, the numerical value for the component (measured data) is referred to as the direct ("measured") response, whereas the value for the ratio is referred to as the indirect ("derived") response.

Our database provides the possibility to record all these formats, so a variable name can include the ":" character to be as close as possible to their biochemical notation, as well as the special characters "/" and "+" as flag codes for indirect variables.

The hierarchical classification and the ability to record data in an indirect form provides the possibility to analyse not only the concentration of a given molecule, but also the groups containing the molecule.

3.2. RECOGNITION OF PATTERNS OF THE MOLECULAR COMPOSITION OF HUMAN MILK (PREDICTIVE HUMAN MILK MODELLING)

In order to "predictive modelling of human milk", i.e. to identify patterns in the molecular composition of human milk, we focused on two parts:

1. first, to create a primary model that can describe the time trajectories of human milk components,
2. then to show how the parameters of the primary model depend on parameters such as geographic location

3.2.1. Primary model

To get ideas for the structure of possible primary models, we needed detailed data on the temporal variation of HMC-s, measured for individual mothers. Ideally, we should collect such trajectories for a large number of individual mothers, but such datasets are rare. John et al *(John és mtsai., 2019)* published such a dataset; it has also been deposited in MilkyBase.

First, we analyzed the data on measurements for **individual mothers**, focusing on the cases where the largest number of measurements were taken for individual mothers. Our findings showed that these individual trajectories are simple descent curves, i.e. they all decrease linearly over the interval of observation, and the individual trajectories can be described by a monotonic convergent decreasing mathematical function (saturation model).

Following this, individual measurement data for **all mothers** reported in the publication were analysed. Figure 4 shows the total protein concentrations, as a function of postpartum time, measured for individual mothers. As 545 measurements were made for 177 mothers, the average number of samples donated by one mother was around 3. From these, one cannot expect to identify the individual trajectories, but the average concentrations, as a function of time, can be well fitted by the above pure saturation model. As the rate of the exponential convergence here is based on a population, we call it population convergence rate as opposed to the individual convergence rate for a HMC of an individual mother.

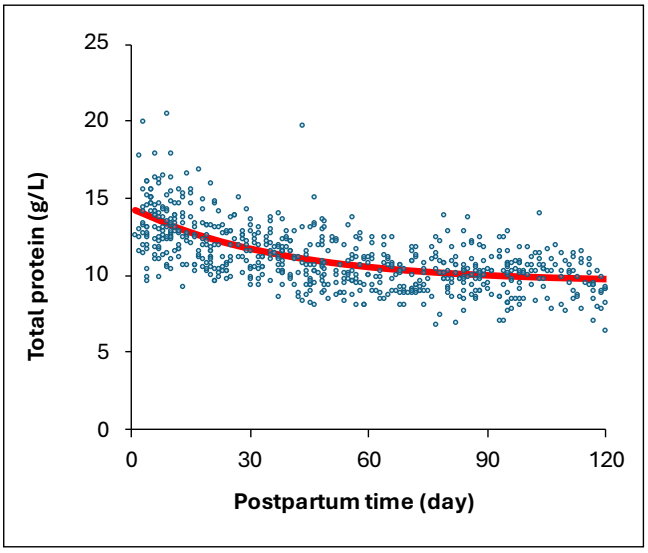

Figure 3. HM protein trajectories based on John et al. *(John és mtsai., 2019)*

It is important to see that, because of the non-linear model, the population convergence rate is not the arithmetical average of the individual rates, but a function of their distribution. However, the pure saturation function is still a good model for their typical (average) behaviour.

The fitted model can be described as:

$$y(t) = 14.245 \cdot e^{-0.031 \cdot t} + 9.691 \cdot (1 - e^{-0.031 \cdot t}) \qquad (0 \leq t)$$

The regression results , similar to microbiological "half-time",can be interpreted as follows: The rate of the exponential convergence is 0.031/day, which is equivalent to ca 22 days “half-time”; i.e. the remaining distance to the final level halves in every ca. 3 weeks. The relative error of the rate estimate is 15% .The standard error of fit is 1.5 g/L while the final saturation level was estimated as 9.7 ± 0.19 g/L. Considering that the cohort of this study was as homogeneous as can reasonably be expected, the results show, that the **total variation in the protein concentration is primarily due to the inherent biological differences between individual mothers**(cross-sectional) **and to less extent due to the temporal variation**(longitudinal)**.**

To further test our hypothesis, we used trajectories of the total HM protein concentrations from a set of MilkyBase records where the **cohort was from the EU**. Figure 5 shows the trajectories of total protein concentration in breast milk as a function

of time since birth, based on data reported in 7 publications. The data points placed on one trajectory are average protein concentrations derived from the cohort.

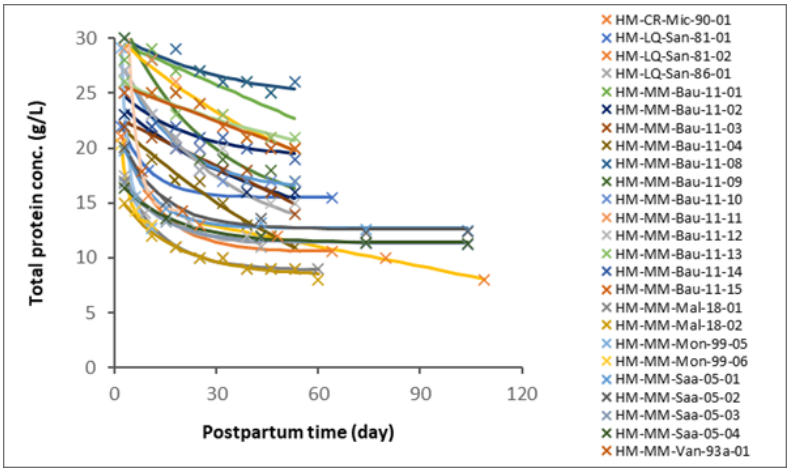

Figure 4. Temporal trajectories of total protein concentration in HM, published in seven papers and stored in the MilkyBase database.. (*(Pacza és mtsai., 2022)*

The data points are average concentrations, typically from >10 mothers. Conditions like gestation age, delivery method, geographical location, etc. all contributed to the total variation. The pure saturation model, with three parameters, was used to fit the trajectories.

The difference between the protein concentrations of **two randomly chosen mother's milk, at the same time after birth**, would be highly likely 3-4 times **bigger** than the difference between the protein concentrations **of two random samples**, one colostrum and one mature, **from the same mother**.

Figure 4 and Figure 5 demonstrates that the fit of the pure saturation model at population level is robust (all the three parameters were estimated by less than 20% relative error).

Analogously, it would be important to know whether a **pattern recognized for a group of molecules** like proteins, fatty acids, oligosaccharides, minerals, vitamins, is **equally valid for the individual molecules** of that group. Data are available for such specific molecules, too, though rarely for individual mothers. Samuel et al *(Samuel és mtsai., 2022)* reported on the population trajectories of selected molecules (Figure 6.). Some of the trajectories followed the pure saturation model but some did so only after a rapidly changing initial.

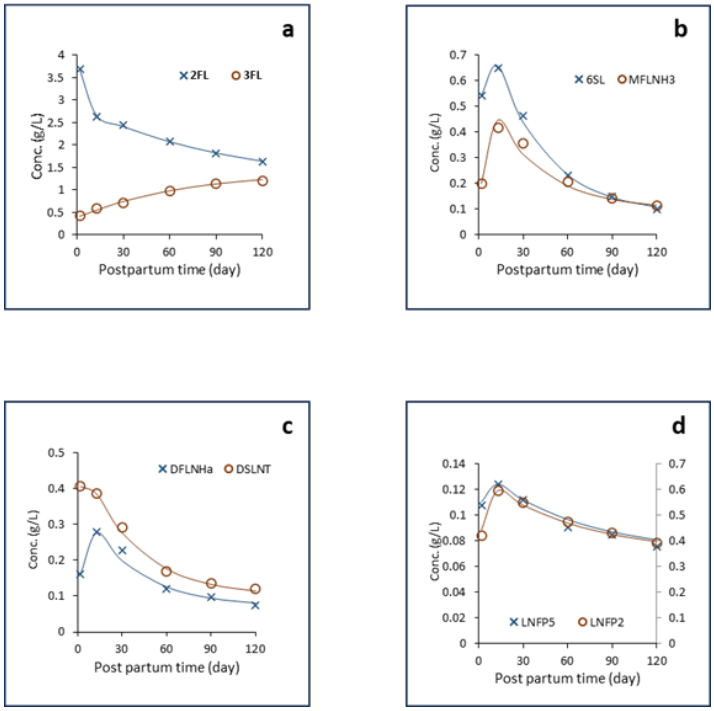

Figure 5. Measured temporal trajectories of various human milk oligosaccharides described by our two-phase primary model.

Very few data are available to vindicate this pre-saturation initial phase. The only such records in MilkyBase were from Liu et al. *(Liu és mtsai., 2019)*. The dataset convincingly demonstrated (see Figure 7) that the total fatty acid concentration linearly increased in the colostrum before the trajectory entered the second, the pure saturation phase. The average of the fitted slopes, in this initial phase) was 3.5 g/L/day, with ca 12% relative standard error. As a rule of thumb, we can say that the fat content ca. doubled during the first 6 days.

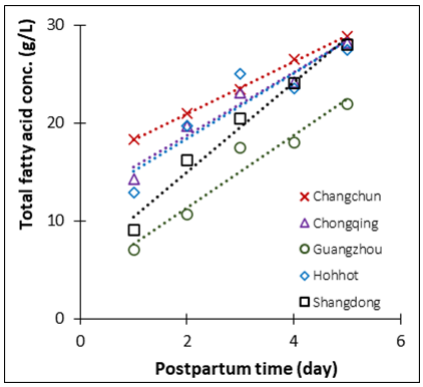

Figure 6. Fast linear increase of the total fatty acid concentration in the colostrum, in five Chinese cities, as reported by Liu et al *(Liu és mtsai., 2019)*

Therefore, for a generic model, we assume an initial, linear “take-off” period prior to the saturation model. We call this as **two-phase saturation model**, with five parameters, as opposed to the **pure saturation model**, with three parameters (i.e. no initial linear phase). If the two phases have the same trend, the initial phase can well be embedded in the pure saturation model. However, the existence of the initial linear phase is surely a good assumption if the two phases have different trends. A physiological explanation for the two phases may be that for some molecules, a sudden change in their production is triggered by the delivery.

We fitted all the data found in Samuel et al *(Samuel és mtsai., 2022)* (Figure 4). A single-phase three-parameter saturation model was fitted to the post-colostrum data by non-linear regression, then the only point from the colostrum was combined with the fitted value at the first post-colostrum point. This way, the slope defined by the first two points demonstrates a bound for the slope of the initial linear phase.

3.2.2. Secondary modell (Geographical differences)

Recall that primary models describe temporal trajectories of human milk components in more-or-less constant conditions, while secondary models are about the effect of those conditions on the parameters of the primary model. The main reason why the two kinds of models should not be merged is mechanistic: it is the rate, with time, how a component changes (not their level directly), that is determined by the conditions. To reduce complexity, it is the parameter set of the primary model, that can be used as a replacement for the whole temporal trajectory.

In what follows, we make comparisons, how a secondary parameter, the final concentration level, is affected by the geographical location. We take extra care, by putting down the plus/minus standard errors in the fitted curves, to indicate the confidence in our findings.

We analysed how the third parameter of the primary model, the final concentration level (secondary parameter), is affected by geography. Since the geographical region is one of the explanatory variables in the MilkyBase database and the time trajectories of the different breast milk components are dynamic response variables, it was easy to investigate the dependence on geographical location.

We took the two-phase saturation curves fitted to the data of Samuel et al *(Samuel és mtsai., 2022)* as a reference for comparison. The reason for this was that here the authors collected the samples at about the same times for individual mothers and the published concentrations are the averages of large samples, therefore their standard errors are small. Figure 8, created from MilkyBase, summarizes the difference between the human milk concentrations produced by EU and Chinese cohorts. The typical (average) fat concentration for the EU cohort is 25-30% higher than that for the Chinese cohort, all along the observation time.

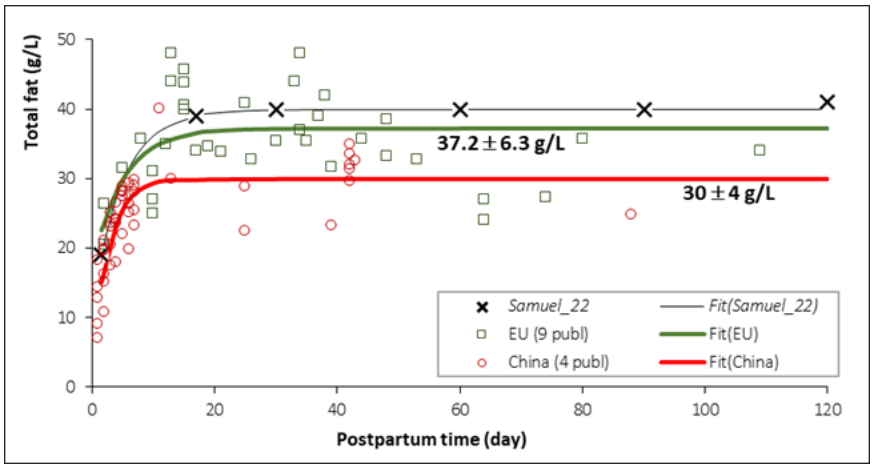

Figure 7. Human Milk total fat concentration trajectories derived from MilkyBase records on EU and Chinese cohorts

The base model (black continuous line) fits average fat concentrations, each of which was produced by hundreds of mothers at about the same time. Notice that they follow a smooth pattern, even if the individual trajectories are more stochastic. Ordinary ANOVA confirmed (though it is visible from the plot, too) that deviations of the raw data caused by the individual mothers' biological differences is significantly higher than the variation caused by the geographical location, i.e. whether the cohort was from EU or China.

The question is whether this difference **this difference in total fatty acid concentration also applies to individual fatty acid molecules**. This can be answered by using molecule-specific data from the MilkyBase database.

The most abundant fatty acid molecules are the oleic (C18:1n-9) and linoleic (C18:2n-6) acids, their sum providing more than half of the total fatty acid content of

human milk. Figure 9 shows the data available from MilkyBase for oleic and the linoleic acid, respectively. As can be seen, the concentrations from the Chinese cohorts tend to have lower level of oleic acid than EU cohorts do, but the situation is the opposite with linoleic acid, especially in the first two weeks.

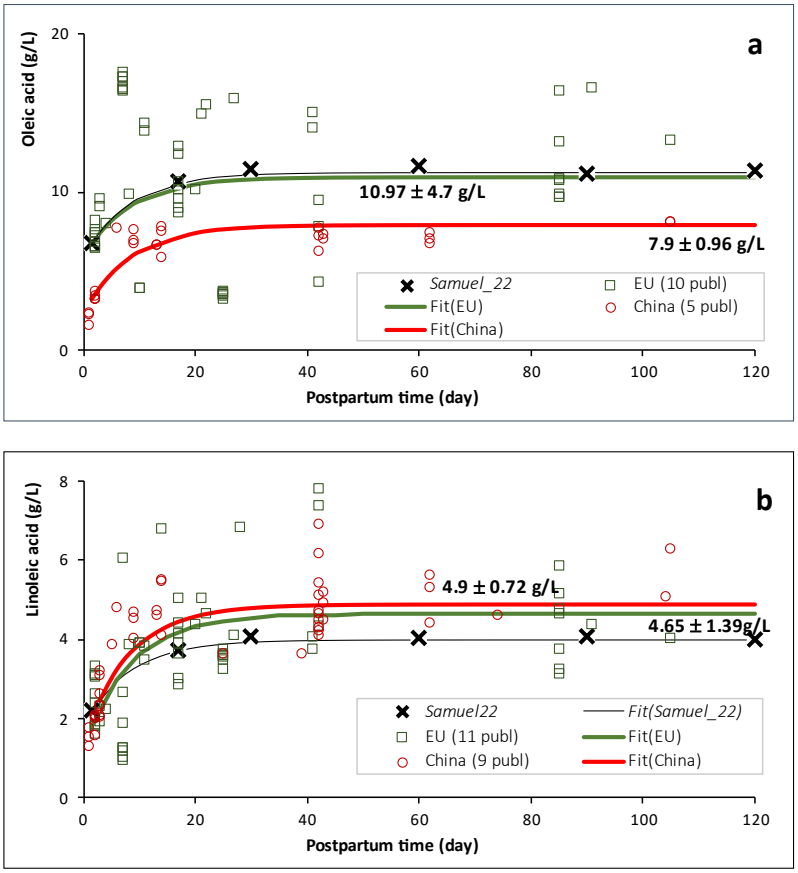

Figure 8. The trajectories of oleic acid (C18:1 n-9) and linoleic acid (C18:2 n-6), in the first 120 days of HM.

In MilkyBase, data are also available for eicosadienoic acid (C20:2n-6) and docosahexaenoic acid (DHA; C22:6n-3) which shows a similar trend.

Therefore, the geography-generated differences in the total fatty acid concentrations does not necessarily show the same patterns when individual fatty acid molecules are compared. The difference in total fatty acid concentrations due to geographic location is therefore not distributed proportionally between the different fatty acid molecules.

4. NEW SCIENTIFIC RESULTS

MilkyBase database novelties

1. We have created the MilkyBase database, based on scientific publications on human milk composition, which stores and organises records on human milk composition. We implemented the following novelties:

   a) An **ontology** was created that considers compositional data as a response to factors that interact with breast milk composition, which are recorded in **explanatory and response variables**. The records represent a 'cause and effect' relationship, i.e. the representation of various explanatory conditions under which observations were made, mapping to breast milk composition as a response variable.

   b) The structure of the database was designed to be able to track changes in breast milk components and composition over time, i.e. to record not only static but also **dynamic** (time-dependent) states. The time-dependent explanatory and response variables [time, value] are represented by data tables, while the respective entry in the main tab of the database is just a pointer to this table.

   c) We have implemented **extended numeric variables**, which allow the recording and measurement of uncertainties and the possibility to record data in **direct and indirect (derived) forms**, thus facilitating a wider range of data collection and analysis.

   d) The structure of the database is defined by grouping the elements in a **hierarchical tree structure**. The generated structure allows the creation of a graph of breast milk components and supports multilevel analyses..

Recognition patterns in the molecular composition of human

Mathematical modelling was used to characterise the components of human milk, and the following results were obtained.

2) In the primary modelling, **a single-phase simple saturation model** was created

$$y(t) = y_0 \cdot e^{-r \cdot t} + y_{End}(1 - e^{-r \cdot t}) \qquad (0 \leq t, r \geq 0)$$

The model has three parameters, the initial concentration y_0 and the final concentration y_{End}, and the rate of exponential convergence to the final concentration (r).

3) Based on the data recorded in the MilkyBase database, the following results were obtained for the **total protein concentration in human milk** using the single-phase simple saturation model:

   a) We found that **the fit** of the generated simple saturation model is quite **robust** at the population level (all three parameters were estimated with relative errors of less than 20%) and **flexible** enough to fit protein trajectories at **both the individual and population level**.

   b) We found that there is bigger variation in total protein concentration due to biological differences between individual mothers (so-called cross-sectional variation - standard deviation of about 5 g/L), and much less variation as a function of time (longitudinal variation - standard deviation of about 1-2 g/L).

4) A **general two-phase saturation model** (of which the single-phase simple saturation model is a special case) is developed to describe the time trajectories of the breast milk components, where the initial linear period is followed by an saturation phase

$$y(t) = \begin{cases} y_0 + a \cdot t & (0 \le t < \lambda) \\ y_\lambda \cdot e^{-r\cdot(t-\lambda)} + y_{End}(1 - e^{-r\cdot(t-\lambda)}) & (\lambda \le t) \end{cases}$$

where $y_\lambda = y_0 + a\cdot\lambda$, $0 \le r$, $0 \le \lambda$.

The model has 5 parameters, the initial concentration (y_0) and final concentration (y_{End}), the rate of change of the concentration of the breast milk component in the initial phase (a), the saturation rate (r) and the duration of the initial (colostrum) phase (λ).

The optional initial - colostrum - linear phase has two parameters: the concentration of the component on the day of birth and the rate of change of concentration. These phases justify the traditional breast milk temporal change distribution (colostrum - transitional - mature).

5) Fitting the two-phase model to the data of individual molecules recorded in the MilkyBase database we obtained the following results:

   a) The two-phase model showed a good fit for the fatty acid molecules, oligosaccharides as well as minerals studied.

   b) It was found that the trajectories of C18:0, C18:1n9, C18:3n3 fatty acid molecules increase rapidly up to the final concentration level, while C20:2n6, C22:1n9 fatty acid molecules, zinc and selenium all show a decreasing trend. Furthermore, the concentrations of C22:6n3, C16:1n7 and C20:0 fatty acid molecules and phosphorus in human milk first increase linearly and then converge to the final level following an exponentially decreasing trajectory.

6) Secondary modelling was used to evaluate the geographical effect on the final concentration level of total fatty acid in human milk, and the following results were obtained:

   a) For both Europe and China, the cross-sectional variation in total fat content of breast milk due to stochastic biological differences between mothers (about 15 g/L) is larger than the longitudinal or geographic variation (6.3 g/L for European cohorts and 4 g/L for Chinese cohorts).

   b) We found that the difference in total fatty acid concentrations caused by geographic location is not distributed proportionally among the different fatty acid molecules.

5. PRACTICAL APPLICABILITY OF THE RESULTS

Our primary goal was to define an ontology to help explore the relationship between the composition of breast milk and various factors. A secondary goal is to make this ontology useful for research on other foods, where users can store their own data in a similar format.

Practical results of the developed ontology and the used modelling methods are as follows:

1. 1. Researchers and nutritionists can use MilkyBase to identify patterns in the composition of milk based on a variety of factors. These patterns can help understand

how maternal conditions, birth or other environmental influences can affect the composition of breast milk.

2. The hierarchical organisation of MilkyBase fields helps to better use statistical methods to analyse the data, providing the opportunity to explore relationships and correlations within the data set.

3. The database can help identify " under-explored " areas, pointing the way for further research projects.

4. MilkyBase serves as a platform where users can contribute their own data in a standardised format, helping to standardise food composition data formats, make databases compatible and realise their full potential.

5. The unified structure provided by our ontology simplifies comparisons within and between datasets and allows for quick searching of the database based on predefined key parameters. Ease of use encourages researchers to capture the data they create in this format, facilitating control, collaboration and data sharing.

6. The introduction of mathematical modelling of human milk components opens up the possibility of identifying their hidden patterns on an objective basis. Predictions, can be of great benefit for experimental design and data interpretation, as well as for the selection of research and innovation areas. The field can be further developed by integrating machine learning techniques and managing complex and large data sets. Future challenges include developing more accurate and efficient computational models that can handle the increasing complexity and scale of scientific problems. With the rise of artificial intelligence, there is an opportunity for these programmes to make better use of measurements and the accumulated collective knowledge to support decision-making.

7. Furthermore, based on the results of our model, we suggest that research designs should include non-evidistant sampling times for breast milk research, with more frequent sampling in the first two weeks after delivery. This research approach could more effectively characterize rapid changes in colostrum period and provide a more accurate picture of the temporal dynamics of breast milk.

6. BIBLIOGRAPHY

1. *Agostoni, C. - Braegger, C. - Decsi, T. - Kolacek, S. - Koletzko, B. - Michaelsen, K. F. - Mihatsch, W. - Moreno, L. A. - Puntis, J. - Shamir, R. - Szajewska, H. - Turck, D. - van Goudoever, J. and Nutrition, E. C. o.*: 2009. Breast-feeding: A Commentary by the ESPGHAN Committee on Nutrition. *Journal of Pediatric Gastroenterology and Nutrition*. 49. (1). 112-125. 10.1097/MPG.0b013e31819f1e05

2. *Ballard, O. and Morrow, A. L.*: 2013. Human Milk Composition. *Pediatric Clinics of North America*. 60. (1). 49-74. 10.1016/j.pcl.2012.10.002

3. *Carr, L. E. - Virmani, M. D. - Rosa, F. - Munblit, D. - Matazel, K. S. - Elolimy, A. A. and Yeruva, L.*: 2021. Role of Human Milk Bioactives on Infants' Gut and Immune Health. *Front Immunol*. 12. 604080. https://doi.org/10.3389/fimmu.2021.604080

4. *Christian, P. - Smith, E. R. - Lee, S. E. - Vargas, A. J. - Bremer, A. A. and Raiten, D. J.*: 2021. The need to study human milk as a biological system. *The American Journal of Clinical Nutrition*. 113. (5). 1063-1072. https://doi.org/10.1093/ajcn/nqab075

5. *De Weerth, C. - Aatsinki, A.-K. - Azad, M. B. - Bartol, F. F. - Bode, L. - Collado, M. C. - Dettmer, A. M. - Field, C. J. - Guilfoyle, M. - Hinde, K. - Korosi, A. - Lustermans, H. - Mohd Shukri, N. H. - Moore, S. E. - Pundir, S. - Rodriguez, J. M. - Slupsky, C. M. - Turner, S. - Van Goudoever, J. B. - Ziomkiewicz, A. and Beijers, R.*: 2022. Human milk: From complex tailored nutrition to bioactive impact on child cognition and behavior. *Critical Reviews in Food Science and Nutrition*. 63. (26). 1-38. 10.1080/10408398.2022.2053058

6. *Hooton, F. - Menichetti, G. and Barabási, A.-L.*: 2020. Exploring food contents in scientific literature with FoodMine. *Scientific Reports*. 10. (1). 16191. 10.1038/s41598-020-73105-0

7. *John, A. - Sun, R. - Maillart, L. - Schaefer, A. - Hamilton Spence, E. and Perrin, M. T.*: 2019. Macronutrient variability in human milk from donors to a milk bank: Implications for feeding preterm infants. *PLOS ONE*. 14. (1). e0210610. 10.1371/journal.pone.0210610

8. *Levenberg, K. J. Q. o. A. M.*: 1944. A METHOD FOR THE SOLUTION OF CERTAIN NON – LINEAR PROBLEMS IN LEAST SQUARES. 2. 164-168.

9. *Liu, Y. - Liu, X. and Wang, L.*: 2019. The investigation of fatty acid composition of breast milk and its relationship with dietary fatty acid intake in 5 regions of China. *Medicine*. 98. (24). e15855. 10.1097/MD.0000000000015855

10. *Marquardt, D. W.*: 1963. An Algorithm for Least-Squares Estimation of Nonlinear Parameters. 11. (2). 431-441. 10.1137/0111030

11. *Pacza, T. - Martins, M. L. - Rockaya, M. - Müller, K. - Chatterjee, A. - Barabási, A.-L. and Baranyi, J.*: 2022. MilkyBase, a database of human milk composition as a function of maternal-, infant- and measurement conditions. https://figshare.com/s/c44b92932fc1a5785cd3

12. *Perrella, S. - Gridneva, Z. - Lai, C. T. - Stinson, L. - George, A. - Bilston-John, S. and Geddes, D.*: 2021. Human milk composition promotes optimal infant growth, development and health. *Seminars in Perinatology*. 45. (2). 151380. https://doi.org/10.1016/j.semperi.2020.151380

13. *Rossum, C. - Büchner, F. and Hoekstra, J.*: 2005. Quantification of health effects of breastfeeding - Review of the literature and model simulation. *Annals of Nutrition and Metabolism*. 51.

14. *Samuel, T. M. - Zhou, Q. - Giuffrida, F. - Munblit, D. - Verhasselt, V. and Thakkar, S. K.*: 2020. Nutritional and Non-nutritional Composition of Human Milk Is Modulated by Maternal, Infant, and Methodological Factors. *Frontiers in Nutrition*. 7. 576133. https://doi.org/10.3389/fnut.2020.576133

15. *Samuel, T. M. - Thielecke, F. - Lavalle, L. - Chen, C. - Fogel, P. - Giuffrida, F. - Dubascoux, S. - Martinez-Costa, C. - Haaland, K. - Marchini, G. - Agosti, M. - Rakza, T. - Costeira, M. J. - Picaud, J. C. - Billeaud, C. and Thakkar, S. K.*: 2022. Mode of Neonatal Delivery Influences the Nutrient Composition of Human Milk: Results From a Multicenter European Cohort of Lactating Women. *Front Nutr*. 9. 834394. 10.3389/fnut.2022.834394

16. *Sánchez, C. - Franco, L. - Regal, P. - Lamas, A. - Cepeda, A. and Fente, C.*: 2021. Breast Milk: A Source of Functional Compounds with Potential Application in Nutrition and Therapy. *Nutrients*. 13. (3). 1026. 10.3390/nu13031026

17. *Shenhav, L. and Azad, M. B.*: 2022. Using Community Ecology Theory and Computational Microbiome Methods To Study Human Milk as a Biological System. *mSystems*. 7. (1). e01132-01121. 10.1128/msystems.01132-21

7. LIST OF PUBLICATIONS RELATED TO THE DISSERTETION

UNIVERSITY of DEBRECEN

UNIVERSITY AND NATIONAL LIBRARY
UNIVERSITY OF DEBRECEN
H-4002 Egyetem tér 1, Debrecen
Phone: +3652/410-443, email: publikaciok@lib.unideb.hu

Registry number: DEENK/567/2024.PL
Subject: PhD Publication List

Candidate: Tünde Pacza
Doctoral School: Doctoral School of Nutrition and Food Sciences
MTMT ID: 10084787

List of publications related to the dissertation

Foreign language scientific articles in international journals (4)

1. Baranyi, J., Csorba, S., Farkas, Z., **Pacza, T.**, Jóźwiak, Á.: Internal dynamics of patent reference networks using the Bray-Curtis dissimilarity measure.
   J Big Data. 11 (1), 1-10, 2024. EISSN: 2196-1115.
   DOI: http://dx.doi.org/10.1186/s40537-024-00883-z
   IF: 8.6 (2023)

2. Baranyi, J.*, **Pacza, T.***, Martins, M. L., Thakkar, S. K., Samuel, T. M.: Modelling the temporal trajectories of human milk components.
   BMC Pregnancy Childbirth. 24 (1), 1-13, 2024. EISSN: 1471-2393.
   DOI: http://dx.doi.org/10.1186/s12884-024-06896-z
   * These authors contributed equally to this work.
   IF: 2.8 (2023)

3. Martins, M. L., **Pacza, T.**, Müller, K. E., Baranyi, J.: A computational approach to nutrition science reveals the dynamics of the protein content of human milk.
   Innovative Food Science & Emerging Technologies. 82, 1-5, 2022. ISSN: 1466-8564.
   DOI: http://dx.doi.org/10.1016/j.ifset.2022.103167
   IF: 6.6

4. **Pacza, T.**, Martins, M. L., Rockaya, M., Müller, K. E., Chatterjee, A., Barabási, A. L., Baranyi, J.: MilkyBase, a database of human milk composition as a function of maternal-, infant- and measurement conditions.
   Sci Data. 9 (1), 1-7, 2022. EISSN: 2052-4463.
   DOI: http://dx.doi.org/10.1038/s41597-022-01663-1
   IF: 9.8

Foreign language abstracts (1)

5. **Pacza, T.**, Martins, M. L., Müller, K. E., Baranyi, J.: MilkyBase- A Database for Molecular-Level Mapping of the Composition of the Human Milk.
   In: International Milk Genomic Consortium : IMGC HYBRID Symposium 2023, IMGC, Cork, 1, 2023.

UNIVERSITY AND NATIONAL LIBRARY
UNIVERSITY OF DEBRECEN
H-4002 Egyetem tér 1, Debrecen
Phone: +3652/410-443, email: publikaciok@lib.unideb.hu

List of other publications

Foreign language scientific articles in Hungarian journals (1)

6. Mposula, Z., **Pacza, T.**, Szepesi, J., Máthé, E.: Lifestyle and socio-economic inequalities in diabetes prevalence in Madadeni Township, South Africa.
Acta Med. Sociol. 14 (37), 5-21, 2023. ISSN: 2062-0284.
DOI: http://dx.doi.org/10.19055/ams.2023.12/15/1

Total IF of journals (all publications): 27,8
Total IF of journals (publications related to the dissertation): 27,8

The Candidate's publication data submitted to the iDEa Tudóstér have been validated by DEENK on the basis of the Journal Citation Report (Impact Factor) database.

14 November, 2024

DEBRECENI EGYETEM
Egyetemi és Nemzeti Könyvtár