# SHORT THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY (PHD)

## DEVELOPMENT AND APPLICATION OF A NEW FOOD COMPOSITION DATABASE TO REVEAL PATTERNS OF HUMAN MILK COMPOSITION

Mayara Lopes Martins

Dissertation Supervisor: Dr. József Baranyi
Dissertation Co-supervisor: Dr. Katalin E. Müller


UNIVERSITY OF DEBRECEN

DOCTORAL SCHOOL OF NUTRITION AND FOOD SCIENCES

Debrecen, 2023

**<< Development and application of a new food composition database to reveal patterns of human milk composition>>**

**By Mayara Lopes Martins (MSc)**

**Supervisor: Dr. József Baranyi, PhD**

**Co-supervisor: Dr. Katalin E. Müller, MD, PhD**

Doctoral School of Nutrition and Food Science, University of Debrecen

**Head of the Defense Committee: Prof Dr. Béla Juhász, Ph.D.**

Reviewers: Dr. János Major, Ph.D.

Prof. Dr. Erzsébet M. Karaffa, Ph.D.

Members of the Defense Committee: Prof. Dr. András Arató, Ph.D., D.Sc.

Dr Pálma Siposné Fehér, Ph.D.

The PhD defense takes place in the Lecture hall of Department of Internal Medicine, Building A,, Faculty of Medicine, University of Debrecen, on 19th of June, 2024, at 2 p.m.

## INTRODUCTION

In early infancy, nutrition is key to the modulation of the offspring's health and development driven by a unique plasticity period, in the first 1,000 days of life. This period is a window of opportunity when infant nutrition plays a crucial role in forming a basis on which health in adult life will depend.

It is well-recognized worldwide that the best food for infants over the first period of life is their own mother's breastmilk, abbreviated by HM (Human Milk) in this in this study. Exclusive breastfeeding is recommended until the 6th month of life and infant feeding should be continued, along with complementary food, for up to 2 years or beyond.

The uniqueness of HM is rooted in its distinctive composition that is different from any other food. Beyond delivering general nutritional requirements, in terms of both macro- and micronutrients, HM composition (HMC) comprises bioactive components such as cells, microbiota, hormones and antigens,

and, together with the nutritional elements, they cover all metabolic, developmental and growth needs of the infant.

HM is a living system that is shaped individually, depending on the infant (gestational age, size, sex, etc.) and maternal (diet, environment, etc.) characteristics, as well as own factors like expression, storage. HMC changes over time, driven by various factors of the mother-milk-infant triad, that makes HM a dynamic "system within a system".

Nutrition and food scientists face key questions on HMC such as (i) how maternal and infant characteristics influence the dynamics of HM, (ii) how to tailor HM to positively influence the infant's health and development and, (iii) in circumstances, when breastfeeding is not an option, how to optimize and personalize infant formulae that can be also specific to mother-infant pairs.

Three main points hinder the advancement of our current knowledge: (i) most studies focus on the impacts of single

factors, within the mother-milk-infant triad, on the individual components of HM; (ii) multi-omics methods, which are primary tools to analyse HMC, are used in cross-sectional rather than longitudinal sense and (iii) there is a notable lack of knowledge about advanced computational and statistical tools, and this hinders the adequate analysis of relevant data.

To explore HMC knowledge and advance the research field, a proper framework to capture and store quantitative data from published studies on HMC is a must. Ideally, such a database needs to comport the temporal variation of HMC, accompanied by information on maternal, infant and methodological factors affecting it. Once this structure exists, available HMC data can be stored and analysed in an analogous manner to a biological system.

At the moment, no database concentrates on the dynamics of HM. Therefore, the present study focuses on some basic principles, how to build a database on HMC (nutritional and

bioactive components) that complies with the above scientific and IT standards.

## OBJECTIVE

Our primary objective has been, to build a novel food composition database on HMC (nutritional and bioactive components), considering two elements of complexity: time-dependence and variability. The secondary objective is , to demonstrate the applicability of such database via its visual tools to find pattern in the total protein content of HM.

## METHODOLOGY

FCD-s carry a large amount of data, mostly focused on the composition of food as static data, neglecting the inevitable temporal variation of that composition and frequently omitting important information on the conditions, under which the data were acquired. MilkyBase was created to address this issue by data science methods.

The target of MilkyBase is the dynamics of HMC as a function of relevant factors in the mother-milk-infant triad. Our vision is that MilkyBase could contribute to a Big Data approach to advancing HM research. Big Data can be defined by several V-s: Volume, Velocity, Veracity Variety, Variability and Value. “Volume” indicates the big amount of generated data; “Velocity” refers to the fast access and processing of the data; “Veracity” implies that the generated data are verifiable, reliable and consistent. “Variety and Variability” denote the diversity of data due to its complexity and heterogeneity, and lastly, “Value” refers to the benefit provided by comprehensible data analysis using.

To make MilkyBase easily accessible for nutrition- and food-scientists, the database was created in Microsoft Excel, a popular software used by scientists of all fields, with user-friendly facets. Additionally, Microsoft Excel is accompanied by the Visual Basic for Application (VBA) programming language that was utilized to support the functionality of

MilkyBase. Macros written in VBA were used to validate the syntax and semantics of the database. It made users aware of possible errors in the data while populating MilkyBase.

VOLUME

To cover the volume aspect, a targeted search of relevant literature was performed to find large amount of quantitative data on HMC. The search focused on PubMed with the following MeSH terms and Boolean operators: (“human milk” OR “breast milk” OR “mothers’ milk”) AND (“nutrients” OR “components” OR “composition” OR “biochemical” OR “quantification” OR “bioactive”). The search prioritized (but was not limited to) English language literature.

In parallel, FoodMine was used to systematically evaluate title and abstract of published studies related to HMC in PubMed. As the volume, the goal was to get enough quantitative and longitudinal data on HMC to start building a prototype framework to comport them all and, progressively, while more

data were coming in, more adjustments were made until the final MilkyBase template was created.

## VELOCITY

The principle of velocity was not stressed in MilkyBase, mainly because the current volume of data is modest for a Big Data philosophy and does not affect the functionality of the database. As the number of records will reach tens of thousands, this will be an issue. Then, Microsoft Excel will be the transit area to transfer MilkyBase to a more sophisticated SQL system that would be more like a piece in a Big Data chain.

## VARIETY AND VARIABILITY

After literature search and before inputting data in MilkyBase, two key questions regarding the source of knowledge (scientific papers) needed to be asked, (i) Does it provide sufficiently large set of quantitative and longitudinal data on HMC? and (ii) What data are important and should be added? If the paper provides numerous temporal datapoints in addition to information on the

conditions under which the data were measured, then such paper was prioritized to be inputted in MilkyBase.

A record in MilkyBase can be divided into two types of fields: quantitative data on the HMC (response fields) and the information on the conditions (explanatory fields) under which the responses were measured. The latter ones could be related to the mother (BMI, diet, age, etc.), infant (gestational age, sex, weight, etc.) and milk sample (expression, pasteurization, storage temperature, etc.).

In terms Veracity, the original data were extracted from refereed scientific papers, so their verification was not an issue as that had happened during the refereeing process and it is not among the remits of Milkybase to possibly overrule the publisher's decision.

Variety was addressed in many ways. The default unit used for HMC was set as "gram *per* Liter" of milk, therefore, conversion had to be performed on the original data whenever other units

were used. Similarly, if values were given in mass of component *per* mass of HM, they were converted to g/L where 1 kilogram of milk was assumed to be 1 liter. However, an entry can be a derived quantity, too, like the ratio between two HMC-s. Such is the proportion of a (group of) molecule(s) compared to a bigger group, like the percentage of omega-3 within the Polyunsaturated Fatty Acid.

Variety demanded more efforts to validate the syntax of the records. As described before, checking the syntax is vital to maintain the structural integrity of the database, while the semantic check makes sure that the data are interpreted correctly. While the first one can be automated, the latter one must be mainly human-based work, necessitating nutritional expertise, which affects directly what data should be collected.

Besides, MilkyBase allows estimations on HMC even when quantitative data of a certain component is missing. That was facilitated by organizing all the data (response and explanatory

fields) in a hierarchical tree-structured order. For instance, in the response fields, the total HMC plays the role of the root of the tree. It has many nodes - the groups of components - until it reaches the leaves, i.e., the component itself at molecular level. So, if the information of a certain component is needed (leaf), but only data on its former group (node) was available, estimations (at certain confidence level) could be performed to find an expected value for that component.

MilkyBase contains information on the uncertainties of HMC measurements. A numerical entry has two parts, where the first represents the arithmetical average of measured data (including a single measurement), while the second part characterizes its uncertainty or spread. Alternatively, the second part can be an interval, such as the 95% confidence interval. This format requires a minimum of one single value, then the remaining info is optional. Therefore, if the entry represents an interval, then a (possibly stochastic) interval-analysis can be applied in succeeding calculations.

A centroid value on the concentration of the component is recorded (target data), along with the quantification of its spread or uncertainty (standard deviation or minimax boundaries). This information may be accompanied by estimations/predictions, supplied by their uncertainties, commonly the standard error of the mean, or its 95% confidence interval.

To capture HMC temporal variability, MilkyBase holds an independent location for temporal measurements of HMC (dynamic data). Namely, in a separate sheet, time-dependent values of the variables in question are represented by a table of "time-HMC" pairs. This feature in MilkyBase allows the plotting and use of dynamic data and analysing their parameters, such as rates, and their dependence on the conditions of the data generation. This template focusing on dynamic HMC data was inspired by the primary and secondary modelling methods of predictive microbiology, where primary model describes the temporal profile of a variable under constant conditions, characterized by a few (most importantly

rate-) parameters. The variations of the primary parameters are described by secondary models, as a function of the explanatory variables characterizing conditions, under which the primary temporal profiles were produced. The way HMC data were structured in MilkyBase follows this scheme.

## VALUE

MilkyBase brings three novelties in the FCD field: (i) the target is temporal data (both in response and explanatory fields); (ii) the quantification of the uncertainties and (iii) the tree structured organization of both the explanatory and response variables, allowing the users to execute probabilistic estimations equivalent to interval arithmetic.

## MILKYBASE STRUCTURE

Ultimately, MilkyBase is a flat database, which is a system of linked sheets in a single Microsoft Excel workbook. The core of MilkyBase is named Master sheet that contains records (Excel rows) identified by unique keys. The fields

(columns) can be divided into three groups: (i) administrative fields, with details on how the data were collected; (ii) explanatory fields related to the conditions, under which the data on factors of the mother-milk-infant triad were produced and (iii) response fields with the measured/reported data on one or more HMC.

*Numeric value*

Numeric values in MilkyBase can be integer (e.g., 6), fixed point format (e.g., 6.12 ), exponential format (6.12e5) and interval (e.g., [5.12e5, 9.9e6]. The Hashtag (#) is a special character, indicating a placeholder for an unknown number. Thus, [6, #] indicates a number greater than 6. In the Condition and Component sheets, the values can be combined with error margins like 6±2 or 6@[4,8]. Respectively, 6±2 or 6@[4,8] mean that either (i) the real value is between 4 and 8; or (ii) 6 is the centre with [4,8] as quantiles; or (iii) if 6 is an estimation then its 95% confidence interval is defined as [4,8]. Scientific notation is used when a number is less than 0.001.

In the Component and Condition sheets, both singular and dual numerical formats are permitted. The dual format comprises of two parts: the first refers to raw (measured) data, possibly with its scatter (standard deviation or inter-quantile range or minimax interval); the second part describes an estimation and the confidence in this estimation: either 95% confidence interval or standard error of the mean). So, the first part is descriptive, and the second part is predictive (or: an estimate).

For instance, as typical numeric entry X±Y; X@[x,y] can be described as: X, a centroid of the raw data (average, median, etc.). X in the second part (after the semicolon) is about its estimation, generally the observed mean. Y is the Standard Deviation of the measured data, but if it was presented in the second part, then it would represent the standard error of the mean. The interval represented by [x, y] represents the estimated value of interval, such as Confidence Interval (normally 95% CI), but if it was placed in the first part (before

semicolon), then it could also mean the raw minimax values or interquartile etc.

## RESULTS

The workflow of MilkyBase was set in five steps: (i) manual and automated literature search, (ii) selection of papers that gave quantitative and longitudinal data on HMC. Once such papers were selected, (iii) relevant information on the mother-milk-infant triad and on temporal data of HMC were spotted and, finally, (iv) the chosen data were inputted in their corrected location as explanatory or response fields.

By the time MilkyBase was in the process of being populated, the database structure underwent major modifications to be able to capture new information. It took around 150 records in the Master sheet to achieve a consistent structure. Currently, MilkyBase holds roughly 10,000 datapoints and around 600 components of HM (nutritional and bioactive).

MilkyBase can be downloaded and opened in any computer with Excel installed. A Fig Share repository holds the links to download MilkyBase, its user-manual and macros.

## FINAL FRAMEWORK

The heart of MilkyBase is the Master sheet. It has eight fields to be filled with data related to HMC. The fields carry either numerical values (possibly with data on uncertainty as defined above) or list of allowed category values. The numerical domains are intervals, the allowed category values, as an interpretation domain, are given in a separate sheet with a name that is the same as that of the field in question in the Master sheet. Such sheets are called “Definition sheets”, for obvious reasons. The interpretation intervals and the Definition sheets are the main resources for the syntax check programs accompanying MilkyBase.

Definition sheets are organized in a tree structure, which is utilized by the syntax check Excel VBA Add-ins accompanying the database. It is also useful to carry out estimations on specific

molecule content via Bayesian and/or interval arithmetics. Namely, if for example immunoglobulin-A and immunoglobulin-G were measured together, then a new variable handily named "immunoglobulinA+immunoglobulinG" can be introduced and the measured data will be the input for this new, composite variable. This operation corresponds to the merge of two numerical intervals. When the concentration of only one of the two immunoglobulin molecules is to be estimated, then this can be done with a Bayesian conditional probability if, from other, independent data, the proportion of that specific immunoglobulin concentration within the "immunoglobulinA+immunoglobulinG" composite is available.

Another useful feature of the tree-structured data-recording emerges when the proportion between two variables is available. For example the "X/Y" is the name of a variable that indicates the proportion of X compared to Y. This is important

because (especially with fatty acids), sometimes it is not clear what “a percentage of a molecule” means. For example, the proportions “omega6/PUFA” and “omega6/TotalFat” are obviously not the same, still the authors of the scanned article assumed that the meaning of the reported values is clear for the reader. In this case, other MilkyBase records (or simply the expertise of the data inputter should be used) to estimate the denominator, so the definition (and its interpretation interval) will become clear (hence the name “interval arithmetic”).

Beside the Master and definition sheets, MilkyBase includes a dedicated sheet for temporal data, named “DynVal” sheet. This is for “time-value” pairs, representing the temporal variation of the variable in question.

Temporal profiles are the main target entries of Milkybase, the ideal singular entities, because they are the variables of the dynamic mathematical models on which the predictive power of the database is based. In summary, MilkyBase holds 841 records in the Master sheet, from 141 different scientific

publications, adding up ca 10,000 of data points on HMC, including more than 7,000 dynamic datapoints (longitudinal data).

## FINDING INCONSISTENCIES IN PUBLICATIONS VIA MILKYBASE

Through the construction of MilkyBase, various errors were found in publications. One of them is the confusing use of the terms “standard deviation”, and “standard error”. The first quantifies the scatter of the sample around its measured mean, the second quantifies the expected error of the estimation of the real mean. As MilkyBase ontology separates the measured data from the estimates, it is straightforward to find such inconsistencies.

MilkyBase is structured in a way that allows simple checks, by visualization, via comparing suspicious data with other published data.

## DEMONSTRATING THE TEMPORAL VARIATION OF THE HM PROTEIN CONTENT

It is to be noted that the remit of this first version of MilkyBase was not a systematic review of the literature, much rather to collect and clean many of them to reach a critical mass for data analysis. Thus, it may not include all the relevant data that are available, and any scientific conclusion should be interpreted with caution. Protein content in HM was chosen to demonstrate the potential of MilkyBase due to the vast number of available longitudinal data. In total, 21 publications were inputted in MilkyBase on the temporal changes of protein levels in HM.

Based on these 21 publications, the time of collection (postpartum day) and the concentration of protein (g/L) were entered in MilkyBase, along with the correspondent explanatory fields (region, measurement analytical methods and mother-milk-infant conditions. It is noticeable that, between the $60^{th}$ and $250^{th}$ postpartum days (months 2 to 8), the drop in the

protein concentration versus the logarithm of time is close to linear. So, visually, the days 60 and 250 appear to be key markers in the temporal variation of protein in HMC.

## DISCUSSION

MilkyBase was constructed following the primary – secondary modelling structure taken over from predictive microbiology ideas. Its focus is the temporal variation (primary model) of a HM component as a response to various factors (secondary model) related to the mother-milk-infant triad. The database has been supplied with VBA macros to support syntax and semantic checks along with its user manual. Therefore, the present study created a brand-new FCD to cover gaps in HMC research as mentioned previously.

FCD-s carry fundamental data that allow stakeholders from private and public sectors to gather and utilize food composition information. National food databases are the most popular among users, including renown researchers of the

field. The main purpose of national FCD-s are data extraction and validation. However, frequently, vital details of the recorded information (origin, way of sample collection, etc) is missing probably due to difficulties in tracking the process how the FCD was populated. An alternative source of information for food composition may be extracting individual data from retailers and producers via their websites or private databases. However, this alternative has not been explored by researchers yet.

The most popular national and international FCD-s are the FoodData Central from the United States Department of Agriculture, the International Network of Food Data Systems, Food and Agriculture Organization created by FAO, and the European Food Information Resource Association Internationale *Sans but Lucratif*.

There are important limitations related to data quality in a conventional FCD. FoodData Central is based on HMC-

information from the 70s, without any details on the maternal or infant characteristics, sampling, storage, or the methodology applied to measure components. In contrast, MilkyBase does include such background along with details such as sample size, region and measurement methods.

A traditional FCD commonly focuses on the general energy, macro- (protein, fat, carbohydrate) and micronutrients (vitamins and minerals), with less emphasis on bioactive components, such as antioxidants. MilkyBase, on the other hand, comports data from macro-, micronutrients and bioactive components, at a similar priority level.

To date, the structure of current FCD-s does not easily accommodate dynamic data of food composition, possibly due to the complexity behind gathering and harmonizing different databases into a single template. Unlike other FCD-s, the very focus of MilkyBase is the temporal data.

Current FCD-s lack information on some factors influencing food composition, such as environmental aspects (temperature, harvest, soil, etc). It is known that from the beginning of food production, raw foods do change their composition in relation to agricultural practices and animals' diet, and that is similar to the role that factors related to the mother-milk-infant triad play in HMC.

Data related to the mother (age, gestational age, BMI, etc.), infant (weight at birth, allergy, etc.) and the sampling of the milk (storage, handling, etc.) are captured by MilkyBase in its explanatory fields. They are followed, in the same record, by HMC components in the response fields. A cornerstone of MilkyBase was to follow the structure "explanatory-response variables", in order to allow predictive modelling based on the data it stores.

To guarantee veracity and quality of collected data, MilkyBase can record uncertainty and spread (interval) values of HMC measurements. The standard deviation or minimax values along

with any published estimation derived from them (confidence interval or standard error of the mean) are examples of these. The (i) availability of uncertainty data in addition to facts that (ii) HMC is organized in a hierarchical tree structure and (iii) the focus is on dynamic data, are the main novelties of MilkyBase.

The tree-structure is a major novelty of the database; it is vital for checking both syntax and semantics, and it also played a crucial role in making use of uncertain information as well as finding outliers and errors in the published data. Although MilkyBase is currently far from being a big data project, (with appropriate support) it does have the potential to increase its volume to a big data level, when its value can be truly appreciated. To date, the capability of MilkyBase has been shown by brief demos, such as the HM protein dynamics.

In summary, there was a need for a new HMC database with computational tools in HM research and MilkyBase is an

answer to that need. It was developed with novelties that can be used to make more precise decisions in infant nutrition, with the potential of being used generally in the FCD field.

**SUMMARY**

The main barriers to advance the current knowledge in HMC research were (i) most of the studies are focused on studying the impact of single aspects of mother-milk-infant triad on individual components of HM, (ii) multi-omics methods are used in cross-sectional studies rather than longitudinal ones and (iii) there is a notable deficiency of knowledge on advanced computational and statistical tools to properly analysed data on the field.

At the moment, there is no FCD comports such complexity behind HMC, which behaves as a biological system. By following the low-hanging fruit principle, to unlock HMC knowledge and evolve the research field, the present

dissertation initiated a brand-new framework to hold and analyze HMC data in an innovative manner.

MilkyBase stores dynamic data of HMC and carries information on explanatory factors related to the mother-milk-infant triad. It brings together three novelties and valuable features for HM and FCD research field, (i) the target on temporal data (response and explanatory fields); (ii) the quantification of the uncertainties and (iii) the tree structured organization of explanatory and response variables, allowing users to execute probabilistic estimations equivalent to interval arithmetic.

Our database came to cover needs in HM science and proposes a new approach to data management in FCD research where dynamic data can be stored and analysed in order to make more precise decisions in nutrition and food science.

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/157/2023.PL
Subject: PhD Publication List

Candidate: Mayara Lopes Martins
Doctoral School: Doctoral School of Nutrition and Food Sciences
MTMT ID: 10084850

## List of publications related to the dissertation

1. **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.
   DOI: http://dx.doi.org/10.1016/j.ifset.2022.103167
   IF: 7.104 (2021)
2. 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.
   DOI: http://dx.doi.org/10.1038/s41597-022-01663-1
   IF: 8.501 (2021)

**Total IF of journals (all publications): 15,605**
**Total IF of journals (publications related to the dissertation): 15,605**

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.

15 May, 2023

DEBRECENI EGYETEM * Egyetemi és Nemzeti Könyvtár *