Innovative Food Science and Emerging Technologies 82 (2022) 103167

Contents lists available at ScienceDirect

Innovative Food Science and Emerging Technologies

journal homepage: www.elsevier.com/locate/ifset

A computational approach to nutrition science reveals the dynamics of the protein content of human milk

Mayara L. Martins [a], Tünde Pacza [a], Katalin Müller [b,c], József Baranyi [a,*]

[a] Institute of Nutrition, Doctoral School of Nutrition and Food Science, University of Debrecen, Böszörményi út 138, 4032 Debrecen, Hungary
[b] Doctoral School of Clinical Medicine, University of Debrecen, 4032 Debrecen, Egyetem tér 1, Hungary
[c] Heim Pál National Paediatric Institute, Üllői út 86, 1089 Budapest, Hungary

ARTICLE INFO

Keywords:
Food composition
Food database
Computational nutrition
Data science
Human milk

ABSTRACT

To study the computational aspects of collecting available data in a systematically organized database is becoming a matter of urgency in Nutrition & Food Science. Indeed, major projects on developing big datasets have attempted to fill this gap, but so far with limitations on important facets of food composition such as its temporal variation and uncertainty quantification. The need for methodological data processing, from data acquisition, digital storage, statistics and visualization, via pattern recognition and modelling to prediction and optimization is key to make objective and knowledge-based decisions on scientific and technological issues for food industry, academy and regulation. This study aims to demonstrate the use of a recently developed database on the composition of human milk, the first and easily the most complex food in one's life. We show that the purpose-built ontology of the database, with novelties like considering the food composition as a temporal and stochastic response, can help to recognize patterns in the variation of its protein content.

Industrial relevance text: This study highlights the need (i) for introducing ISO-like standards how to digitize food composition data; (ii) for computational methods to explore and utilize such databases to their full potentials.

1. Introduction

The increase in the amount of digitally processed information in the world went through a big explosion twenty years ago (Hilbert, 2020). Today, the question is less whether data are available on a certain issue, but more on how to make sense of the deluge of data on the issue. Since the early 2000s, plenty of journals have emerged and become visible indicators of this new research interest (Raban & Gordon, 2020). Well-known publishers have launched initiatives in this area, e.g., Database (2022), Giga Science (2022), Data in Brief (2022) or Scientific Data (2022). By publishing papers with a focus on digitally shared resources, the academic community has increased their collective knowledge, epitomized by their wiki-philosophy. This process is key to make objective and knowledge-based decisions on scientific and technological issues, for academy, industry, and regulation.

Raban and Gordon (2020) ranked the top 10 broad scientific areas which benefit from Big Data. Intriguingly, Agriculture and Food Science did not make this group, though intuitively one feels that the potential for this is at least at that level where for example Environmental Sciences are, which made No.7 on it. Though, as Kapsokefalou et al. (2019) pointed out, there are examples, where food industry stakeholders make great use of relevant databases through data extraction and/or data validation, still no doubt that Big Data still represents a relatively unused potential for Agriculture and Food Science.

One specific example, where databases are not used to their potentials, is the biochemical composition of foods. Though national food composition databases, such as the FoodData Central created by United States Department of Agriculture (United States Department of Agriculture (USDA), 2022), are commonly used by industrial and non-commercial users (Kapsokefalou et al., 2019), the data have not been utilized efficiently enough to effectively aid decision making related, for example, to public health. It would be beneficial if these databases could provide detailed information on the thousands of biochemical compounds in various food items, as well as on their effects on health

Abbreviations: AI, Artificial intelligence; EuroFIR AISBL, European Food Information Resource Association Internationale Sans but Lucratif; HM, Human milk; INFOODS/FAO, International Network of Food Data Systems/Food and Agriculture Organization; ISO, International Organization for Standardization; USDA, United States Department of Agriculture; R&D, Research & Development.

* Corresponding author.
E-mail addresses: mayara.lopes.martins@agr.unideb.hu (M.L. Martins), pacza.tunde@unideb.hu (T. Pacza), baranyi.jozsef@med.unideb.hu (J. Baranyi).

https://doi.org/10.1016/j.ifset.2022.103167
Received 19 June 2022; Received in revised form 27 September 2022; Accepted 5 October 2022
Available online 8 October 2022

(Barabási, Menichetti, & Loscalzo, 2020).

A good candidate to be covered by such novel database is Human Milk (HM), the first food that humans come across. Once we overcome the challenges related to the database ontology, so that it can accommodate the variability and complexity of HM composition (nutrients, cells, antigens, etc.), the database could serve as a prototype for other similar databases on other food items. The methodology, if well communicated, may present an opportunity to establish ISO-like standards how to make such databases as compatible as possible.

HM is a species-specific food, a unique biological system. It comprised from nutrients and non-nutritional components which are constantly interacting with one another and changing during lactation period (Christian et al., 2021). This is why we consider the way of recording its temporal changes as one of the most important novelties in the database, called MilkyBase, that we recently developed (Pacza et al., 2022). For us, the target is the temporal profiles of the molecules; they are the central entries. In what follows, we demonstrate the advantages of this approach to defining an ontology: how records representing conditions and respective responses, in many cases as temporal profiles, enabled us to reveal dynamic patterns in the protein content of HM.

2. Material and methods

MilkyBase (Pacza et al., 2022) digitizes published data on HM composition as a response to various mother / infant characteristics, as well as to environmental and history conditions. Its novelties are (i) the focus on dynamic conditions and responses; (ii) the quantification of the uncertainties in the entries and (iii) the technique that both conditions and responses are defined within a tree structure, enabling users to perform probabilistic estimations analogous to interval arithmetics.

2.1. Embedding the variables in a tree-structure

The possible value-sets of the various variables (explanatory and response fields) were organized hierarchically in a tree-structured scheme where the root symbolises the food matrix itself. The root is connected to the component-nodes by branches until the leaves (specific molecules) are reached. Each node represents a group of molecules, starting from the broadest one, then with increasing resolution, ending at the molecular level (Fig. 1).

This way, estimating the HM composition is analogous to calculations via the above-mentioned interval-arithmetics supplied with probability distributions. For example, if the concentration of a molecule (a leaf in terms of the tree-structure, like immunoglobulin A) is to be studied but information is available on the total immunoglobulin only, then the (random) proportion of the molecule in question is to be estimated from other data first, then the (also random) concentration of the molecule itself can be estimated by convoluting the respective distributions.

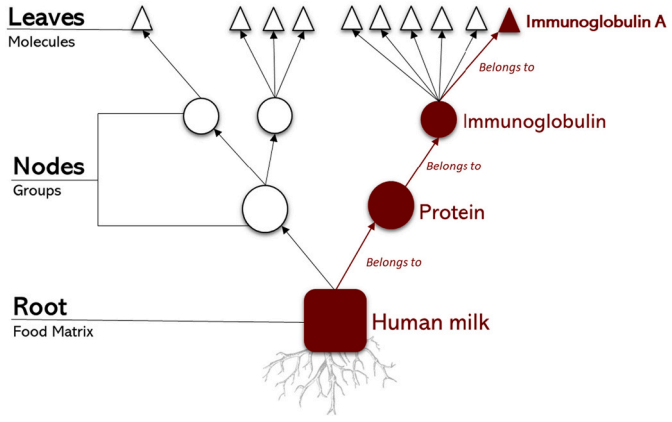


Fig. 1. The components of Human Milk form a tree-structure. For instance, Immunoglobulin A is represented by a single leaf in the tree (molecule-level). Hierarchically, Immunoglobulin A belongs to Immunoglobulins which belong to Proteins. All of them (leaf and nodes) belong to HM (root).

2.2. Quantifying 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 minimax 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 (Table 1). This format demands at least one single numerical value, then all the rest above are optional.

Usually, the first part is the arithmetical average of observations accompanied by the respective spread indicators. This average is typically but not necessarily used as a prediction / estimation of the real mean. Which, in turn, can also be supplied with uncertainty quantification, like the standard error of the estimation of the mean, or its 95% confidence interval. Note that, generally, taking the average is indeed an estimation of the true mean but not necessarily the best one and sometimes (rightfully) the two centroids in the two parts of the field are not the same. Besides, the standard deviation of the data should never be smaller than the standard error of the estimation of the mean, which gives an opportunity to find (shockingly frequent) anomalies in publications, whether the authors speak about standard deviation of the data or the standard error of the estimation of the real mean.

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.

2.3. Dynamic profiles as entries

Time-dependent conditions and responses are represented by [time,

Table 1
Possible entries satisfying the definition of "extended numerical field". It consists of one or two parts separated by a semi-colon. The first part relates to the raw data, the second to prediction. The "@" sign can be read: "from the interval…", followed by two numbers in parentheses defining the interval.

Description	Database entry
Measured/calculated *value*	x_1
Measured/calculated *value* with ***standard deviation***	$x_1 \pm y_1$
Measured/calculated *value* is in the indicated ***interval***	$x_1@[y_1, z_1]$
Measured/calculated value; its ***prediction***	x_1; x_2
Measured/calculated value with ***standard deviation***; its ***prediction***	$x_1 \pm y_1$; x_2
Measured/calculated value is in the indicated ***interval***; its ***prediction***	$x_1@[y_1, z_1]$; x_2
Measured/calculated value; its ***prediction*** with its ***standard error***	x_1; $x_2 \pm y_2$
Measured/calculated value with ***standard deviation***; its ***prediction*** with its ***standard error***	$x_1 \pm y_1$; $x_2 \pm y_2$
Measured/calculated value is in the indicated ***interval***; its ***prediction*** with its ***standard error***	x_1 $@[y_1, z_1]$; $x_2 \pm y_2$
Measured/calculated value; its ***prediction*** with ***95% confidence interval***	x_1; $x_2@[y_2, z_2]$
Measured/calculated value with ***standard deviation***; its ***prediction*** with ***95% conf. Interval***	$x_1 \pm y_1$; $x_2@[y_2, z_2]$
Measured/calculated value is in the indicated ***interval***; its ***prediction*** with ***95% conf. Interval***	$x_1@[y_1, z_1]$; $x_2@[y_2, z_2]$

value] tables and the respective entry in the database is a pointer to this table. Derived parameters of such temporal profiles (curves), such as rate, total change or steady state level, are possible scalar representatives of the profiles. This structure has been inspired by the so-called "primary – secondary model" approach that has become the basis of predictive microbiology, both in terms of mathematical modelling and data storage (see www.combase.cc). That is, the temporal profile of a variable is described by a few key parameters (primary model) and the variation of these parameters, as a function of the conditions under which the response was produced, is described by secondary models.

3. Results

The most time-consuming part of digitizing publications is the integration of the authors' data in the wanted ontology. As mathematical modelling is a sort of "art of omitting the unnecessary" (Baranyi, 2005), so is database building. Digitizing published data is a tedious selection and interpretation process, a good candidate for the application of *artificial intelligence* (AI). Until such AI tools are available, the decision on "*what to record, into what fields, possibly via what transformation*" depends on the purpose of the database and it can be even subjective.

3.1. Utilizing the tree-structure ontology

An example for the pitfalls one can run into is the interpretation of the "concentration" of a component. According to our definition, a component can be a molecule (compound) or a group of molecules. Our reference unit is g/L, i.e. mass in a litre of HM. The approximation that 1 L milk is considered 1000 g of mass, makes the "g / 100g" concentrations directly convertible. However, when the authors measure the concentration of total protein as well as a defined *group of proteins* one must take exceptional care whether the authors mean, by the concentration of that specific *protein group*, as a proportion of the *total protein* or that of the *milk*. This information is sometimes so much hidden or unclarified that only (or not even) experts can make it clear. To avoid errors, it is vital that the person recording the data is familiar with the paper's subject.

Sometimes the tree-structure ontology helps to identify such errors. Namely, if the mentioned specific protein is a proportion of the total protein and the concentration of the total protein can be estimated from other publications, then the "g / L-milk" concentration can also be estimated. Such use of the tree-ontology, of course, was applied to other components too, not to proteins only.

We estimate that up to 1–5% of the hundreds of papers we digitized contained contradictory quantitative information, discoverable by either comparing the description in the "Material and Methods" with the tables anf graphs in the "Results" section of the paper or comparing the resultant record with others' data and finding clear outliers. In obvious cases, we made corrections in the digitized version but, when the mistake was not obvious, we stored the data as had been published.

3.2. Quantifying uncertainties

Another typical example for misinterpretation is the confusion regarding the ± error term. Statistics-minded authors (Barde & Barde, 2012; Nagele, 2003) warned that the *standard deviation* (quantifying the scatter of the statistical sample around its average) is frequently mistaken with the *standard error* of the *estimation* of the real mean of the data. As our method separates the raw data and estimation / prediction, it is a must for us to decide where to put the published error terms. Sometimes, as above, only a comparison with the rest of the data is the only support to make this decision, as the terminology is frequently wrong in publications, as pointed out by the authors above.

An example for this is shown on Fig. 2. Measurements on the alpha-tocopherol (Vitamin E) content of HM were collected from 6 publications. In two publications, producing the first and the next four datapoints in the figure, respectively, we could find the *standard deviation* of the datapoints in the papers, so we plotted them, too, as error bars. These reported standard deviation values of the second paper were much smaller than that of the first point. Going back to the original paper, it turned out that yes, the authors should have called their error term as the standard error of the mean.

3.3. Advantage of focusing on (time, value) tables as default entries

For a case study, we present here the dynamics of the protein content of HM, based on 21 publications (**Table S1**). The observed (time, concentration) pairs published in these papers represent the time elapsed from the infant's birth, in days; while the concentration is meant for the total protein, in g/(L milk). Fig. 3A presents the original data of **Table S1** and Fig. 3B shows the data on the logarithmic time-scale.

It emerges that during the months 2–8 (60–250 days), the decrease of the protein concentration versus the logarithm of time is close to linear (Fig. 3B). This can give the idea of describing and predicting the dynamics of the protein concentration as a function of the logarithm of the post-partum time rather than just the time.

Fig. 4 shows an example for this. Four different authors / papers report on close-to-parallel trajectories (temporal variation of protein content) on the log-time scale, which has a good agreement with the overall trend Fig. 3 describes.

As mentioned, the variability of the protein concentration looks much bigger at the beginning than in months 2–8. This gives the idea that the mother's history affects the protein concentration less and less as time elapses from birth.

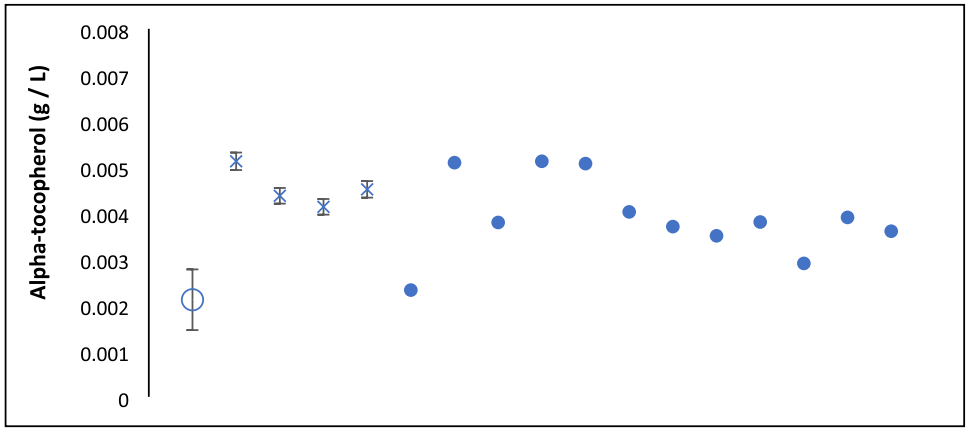
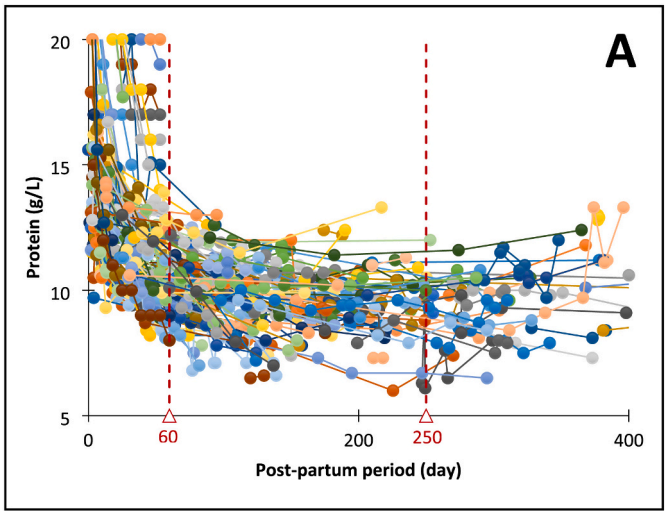


Fig. 2. Alpha-tocopherol (Vitamin E) content of human milk based on six publications. The first 1 + 4 = 5 points, from 2 publications (with symbols empty circle and stars), were supplied with error terms, too, represented by error bars here. The error bar for the first record (key HM-TP-Eli-11-01 in MilkyBase) is much longer than those of the other four points (record keys: from HM-LQ-Lim-20-01 to HM-LQ-Lim-20-4). The original publications double checked, the Standard Deviation and the Standard Error were mistaken for each other in the latter dataset.



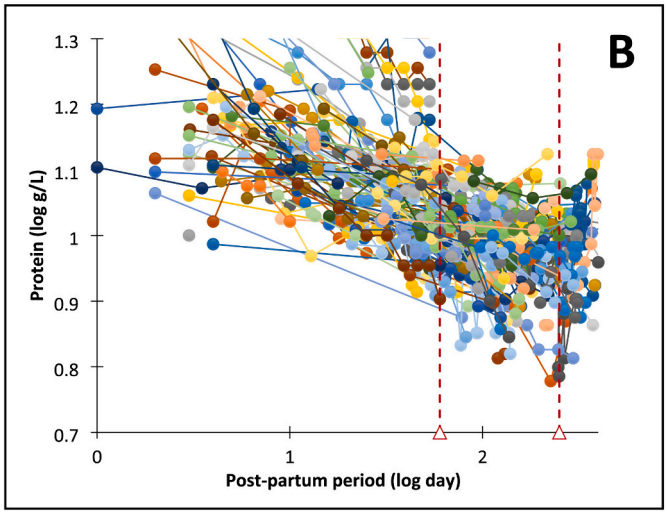


Fig. 3. Dynamics of the protein concentration of HM on the arithmetical (Fig. 3A) and on the logarithmic (Fig. 3B) time-scale. Visually, the days 60 and 250 look like major milestones in the temporal variation of HM composition.

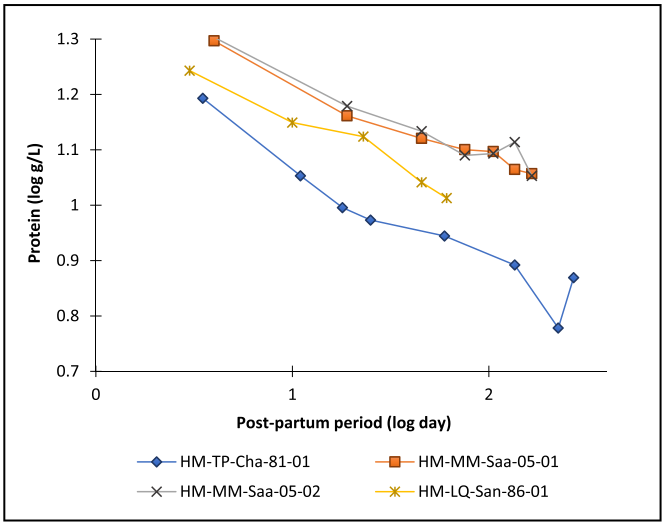


Fig. 4. Published temporal variations of the protein concentration of HM from different publications. Their trend is close to linear on the log-time scale.

4. Discussion

When the matter is extracting or validating data regarding food items and their composition, R&D researchers commonly utilize national food databases (Kapsokefalou et al., 2019). An example of promising projects of food composition datasets is the European Food Information Resource Association Internationale *Sans but Lucratif, the* EuroFIR AISBL (2022) Other examples are the International Network of Food Data Systems provided by the Food Agriculture Organization, known as INFOODS/FAO (2022) and the FoodData Central of the USDA (2022)). While FoodData Central (USDA, 2022) provides data only on American food samples, EuroFIR AISBL (2022) and INFOODS/FAO (2022) appear to be more sophisticated by combining datasets from different national-based databases worldwide.

Overall, these resources carry general data on food items, such as energy, macronutrients (protein, carbohydrate, fat) and their derivates (amino acids, saccharides, fatty acids), minerals, and vitamins. Phytochemicals and other non-nutritional compounds are included, though to less extent, in EuroFIR AISBL (2022) and INFOODS/FAO (2022). As a pioneer of a different, specialized ontology, MilkyBase focuses on the special characteristics of HM (Pacza et al., 2022). While it contains the usual studied nutrients, as INFOODS/FAO (2022), EuroFIR AISBL (2022) and FoodData Central (USDA, 2022), it gives the same weight of importance to the non-nutritional compounds, e.g., oligosaccharides, immunoglobulins, phytochemicals. Besides, MilkyBase prioritises the temporal course for both types of components, which is key to find patterns in the dynamics of HM (Pacza et al., 2022). Mapping the variation over time in food components has not yet been covered by any of the food composition databases (Aleta et al., 2022; Kapsokefalou et al., 2019).

Pooling biological data in a digitized format, with the intention to describe causal interactions, is still a gap in nutrition sciences (Touré, Flobak, Niarakis, Vercruysse, & Kuiper, 2020). The ontology of MilkyBase follows a condition → response mapping, where the condition-fields contain those data that influence the composition of HM, a (multi-variate) response variable (Pacza et al., 2022). The conditions can be environmental factors such as storage temperature of the milk samples or internal influences, like the mother's health or gestational age; the delivery mode, or other characteristics of the mother-infant pair. The way the data are recorded makes the ontology of MilkyBase more efficient than other databases to detect causal effects of the conditions on the milk composition (Pacza et al., 2022).

While combining databases in a common platform remains key to make them comparable, more detailed descriptions of the measurement conditions are needed, to guarantee better accuracy when validating the data (Touré et al., 2020). In MilkyBase, we put special attention on providing information on the uncertainty of data (Pacza et al., 2022). A typical (but not the only) example for this is the standard deviation assigned to observations and estimations. However, when collecting these in the scientific literature, we repeatedly observed that the standard deviation of the data was confused with the standard error of their mean in agreement with Barde and Barde (2012) and Nagele (2003). Confusing terminology is not restricted to statistical concepts, as the various interpretations of the concentration of a particular protein showed.

One of the purposes of our initiative is to emphasize the importance of using the same definitions during quantifications in nutrition sciences, especially in its computational and quantitative areas. MilkyBase, a free resource, can serve as a template to build composition databases for other foods, too, constructed with similar rigour, therefore helping the spread of compatible ontology and uniform terminology.

CRediT authorship contribution statement

Mayara L. Martins: Methodology, Validation, Investigation, Visualization, Formal analysis, Writing – original draft. **Tünde Pacza:**

Methodology, Validation, Formal analysis, Investigation, Visualization. **Katalin Müller:** Conceptualization, Methodology, Supervision. **József Baranyi:** Conceptualization, Methodology, Software, Formal analysis, Supervision, Project administration, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The MilkyBase database is publicly available via Pacza et al (2022)

Acknowledgments

MM has been supported by the Stipendium Hungaricum Scholarship Programme of the Ministry of Foreign Affairs and Trade of Hungary, via the Tempus Public Foundation.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ifset.2022.103167.

References

Aleta, A., Brighenti, F., Jolliet, O., Meijaard, E., Shamir, R., Moreno, Y., & Rasetti, M. (2022). A need for a paradigm shift in healthy nutrition research. *Frontiers in Nutrition, 9*. https://doi.org/10.3389/fnut.2022.881465

Barabási, A.-L., Menichetti, G., & Loscalzo, J. (2020). The unmapped chemical complexity of our diet. *Nature Food, 33–37*. https://doi.org/10.1038/s43016-019-0005-1

Baranyi, J. (2005). Quantitative microbial ecology of food: Evolution of mathematical modelling in food microbiology. *Acta Alimentaria, 335–337*. https://doi.org/10.1556/AAlim.34.2005.4.1

Barde, M., & Barde, P. (2012). What to use to express the variability of data: Standard deviation or standard error of mean? *Perspectives in Clinical Research, 113–116*. https://doi.org/10.4103/2229-3485.100662

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

European Food Information Resource (EuroFIR). (2022). Food data: List of EuroFIR databases. Central. Available online https://www.eurofir.org/food-information/food-composition-databases/.

Hilbert, M. (2020). Digital technology and social change: The digital transformation of society from a historical perspective. *Dialogues in Clinical Neuroscience, 22*(2), 189–194. https://doi.org/10.31887/DCNS.2020.22.2/mhilbert

International Network of Food Data Systems, Food and Agriculture Organization (INFOODS/FAO). (2022). Available online https://www.fao.org/infoods/infoods/en/ (accessed on 15 May 2022).

Kapsokefalou, M., Roe, M., Turrini, A., Costa, H., Martinez-Victoria, E., Marletta, L., … Finglas, P. (2019). Food composition at present: New challenges. *Nutrients, 1714*. https://doi.org/10.3390/nu11081714

Nagele, P. (2003). Misuse of standard error of the mean (sem) when reporting variability of a sample. A critical evaluation of four anaesthesia journals. *British Journal of Anaesthesia*, 514–516. https://doi.org/10.1093/bja/aeg087

Pacza, T., Martins, M. L., Rockaya, M., Müller, K., Chatterjee, A., Barabási, A.-L., & Baranyi, J. (2022). MilkyBase, a database of human milk composition as a function of maternal-, infant- and measurement conditions. *Sci Data, 9*, 557. https://doi.org/10.1038/s41597-022-01663-1

Raban, D., & Gordon, A. (2020). The evolution of data science and big data research: A bibliometric analysis. *Scientometrics, 1563–1581*. https://doi.org/10.1007/s11192-020-03371-2

Touré, V., Flobak, Å., Niarakis, A., Vercruysse, S., & Kuiper, M. (2020). The status of causality in biological databases: Data resources and data retrieval possibilities to support logical modeling. *Briefings in Bioinformatics*, bbaa390. https://doi.org/10.1093/bib/bbaa390

United States Department of Agriculture (USDA). (2022). FoodData Central. Available online: https://fdc.nal.usda.gov/ (accessed on 05 May 2022).

Database (Oxford Academic). About the journal: Database, the journal of Biological Database and Curation. https://academic.oup.com/database/pages/About. 2022. (Accessed 08 Oct 2022).

GigaScience (Oxford Academic). About the journal: GigaScience. https://academic.oup.com/gigascience/pages/About. 2022. (Accessed 08 Oct 2022).

Data in Brief (Elsevier). About the journal: Data in Brief. https://www.sciencedirect.com/journal/data-in-brief. 2022. (Accessed 10 Oct 2022).

Scientific Data (Springer Nature). Journal Information: Sci Data. https://www.nature.com/sdata/journal-information. 2022. (Accessed 08 Oct 2022).