

Reducing the data demand in (clinical) studies

Myra Spiliopoulou

ELLIIT Symposium, Halmstad, Sept 23, 2026

Handout version

Some figures have been replaced by pointers and text (marked in orange).



FACULTY OF
COMPUTER SCIENCE



Predicting treatment outcome

EXAMPLE: Asthma Control Questionnaire (ACQ)^{1 2}

At visit 1

Question 1

During the last 4 weeks, how much of the time has your asthma kept you from getting as much done at work, school or home?

All of the time Most of the time **Some of the time** A little of the time None of the time

Question 2

During the last 4 weeks, how often have you had shortness of breath?

More than once a day **Once a day** 3 to 5 times a week Once or twice a week Not at all

Question 3

During the last 4 weeks, how often have your asthma symptoms (wheezing, coughing, shortness of breath, chest tightness or pain) woken you up at night or earlier than usual in the morning?

4 or more nights a week **2 to 3 nights a week** Once a week Once or Twice Not at all

Question 4

During the last 4 weeks, how often have you used your rescue inhaler or nebuliser medication (such as Salbutamol)?

3 or more times per day Once or twice per day 2 or 3 times per week Once a week or less **Not at all**

Question 5

How would you rate your asthma control during the last 4 weeks?

Not Controlled at all Poorly Controlled **Somewhat Controlled** Well Controlled Completely Controlled

← Treatment →

At visit 2

Question 1

During the last 4 weeks, how much of the time has your asthma kept you from getting as much done at work, school or home?

All of the time Most of the time Some of the time **A little of the time** None of the time

Question 2

During the last 4 weeks, how often have you had shortness of breath?

More than once a day **Once a day** 3 to 5 times a week Once or twice a week Not at all

Question 3

During the last 4 weeks, how often have your asthma symptoms (wheezing, coughing, shortness of breath, chest tightness or pain) woken you up at night or earlier than usual in the morning?

4 or more nights a week 2 to 3 nights a week Once a week **Once or Twice** Not at all

Question 4

During the last 4 weeks, how often have you used your rescue inhaler or nebuliser medication (such as Salbutamol)?

3 or more times per day Once or twice per day 2 or 3 times per week **Once a week or less** Not at all

Question 5

How would you rate your asthma control during the last 4 weeks?

Not Controlled at all Poorly Controlled **Somewhat Controlled** Well Controlled Completely Controlled

Treatment outcome is a change of a *standardized quantity* measured before and after treatment.

If the quantity changed in the desirable direction, then *the patient responded to treatment* (aka: positive response).

¹ American Thoracic Society, access: June 28, 2026 at www.thoracic.org/assemblies/assemblies/srn/questionnaires/acq

² www.asthmacontroltest.com/quiz/adult-quiz/

Claim: The more features we seek, the less data we harvest.

Proof by Example

[Puga et al., 2023]

Patients with chronic tinnitus, outpatient clinic of University Hospital of Regensburg, from January 3, 2016 till May 28, 2020.

Table: Transposition of the first three columns from [Puga et al., 2023, Table 2] 'Number of patient records per questionnaire at t_0 vs at t_0 and t_1 ' ... ordered on the number of questionnaires at t_0 descending'

Questionnaire	THI	MDI	TFI	TQ	TBF12	All ^a
t_0	1067	981	798	746	700	479
t_0 and t_1	123	109	87	70	35	4

^a"All" refers to patients that completed all the questionnaires

The absence of a feature translates to missingness.

The acquisition of a feature incurs cost.

- ▶ the realm of imputation
- ▶ the realm of synthetic data generation
- ▶ the realm of cost-aware learning

Missingness in digital health questionnaires

Track Your Tinnitus app

08:21 26%

Progress 0%

Tinnitus diary

We will now ask you a few questions about how your tinnitus is right now - and how the tinnitus was today. Please answer the questions as best you can.

Questions about this moment:

How loud is your tinnitus at the moment?

inaudible very loud

How burdensome do you find your tinnitus at the moment?

not burdensome very burdensome

How tense does your jaw feel right now?

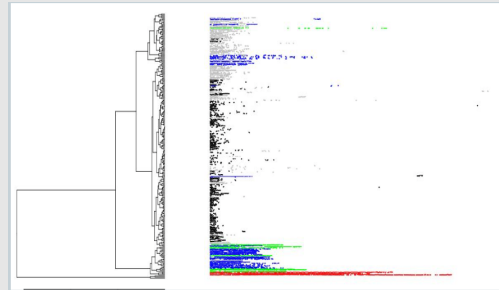
not at all tense very tense

How tense does your neck feel right now?

not at all tense very tense

Back Next

Dataset of the old TYT app [Unnikrishnan, 2017]



Missingness in temporal sequences

- ▶ Imputation
- ▶ SDG
- ▶ cost-aware learning

The realm of imputation

- ✓ Venerable domain of research
- ✓ MICE
- ✓ Deep learning algorithms

? How to evaluate whether an algorithm is good for the data at hand?

Data Analytics Challenge on Missing Data Imputation

[Luo, 2022], using the MIMIC III dataset [Johnson et al., 2016]

Simulating missingness on MIMIC-III

In this Figure 1 from [Luo, 2022] we see the flowchart of selecting admissions from MIMIC III for masking in DACMI; the figure contains the exclusion criteria and the numbers of retained admissions, time points and test results.

Masking

For each admission a and each (one of the 13) labtest t , one of the results was picked at random and masked.

Evaluating on *normalized Root Mean Square Deviation*

$$nRMSD(t) = \sqrt{\frac{\sum_{a,j} I(a,t,j) \left(\frac{|X_{a,t,j} - Y_{a,t,j}|}{\max_j(Y_{a,t,j}) - \min_j(Y_{a,t,j})} \right)^2}{\sum_{a,j} I(a,t,j)}}$$

where $X_{a,t,j}$ is the imputed value for labtest t of admission a at moment j , $Y_{a,t,j}$ is the true value if it existed, whereupon $I_{a,t,j} = 1$ (Masking), else $I_{a,t,j} = 0$.

Natively missing data were also imputed.

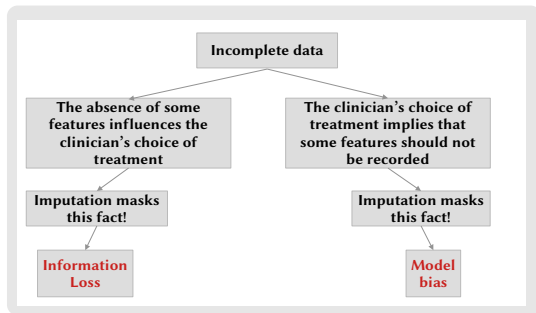
- ✓ not counted in $nRSMD()$
- ! missing because of treatment choice
- ? should they have been imputed ?

⇒ When to impute?

[Berrevoets et al., 2023]

Missingness linked to treatment outcome

[Berrevoets et al., 2023]



The concept of 'Mixed Confounded Missingness'

'... MCM [is] a new missingness mechanism where some missingness determines treatment selection and other missingness is determined by treatment selection.'

Selective Imputation: Impute $\rightarrow \tilde{X}$ but not $\tilde{X} \rightarrow$

Figure 3 from [Berrevoets et al., 2023] shows the DAG for 'selective imputation' with their new concept of MCM, explaining when imputation is permitted and when not

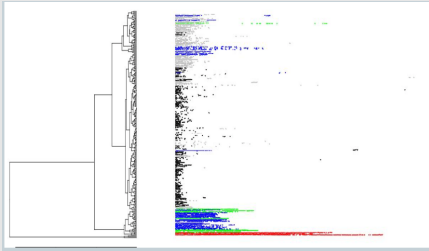
Departing from the realm of imputation

Imputation demands knowledge about the data generating process.

More effort is needed to reconstruct how and when clinical decisions are made, and to incorporate this knowledge into the imputation validation process.

Besides ...

We should not impute when we have more emptiness than data.



The realm of Synthetic Time Series Generators

A choice of recent surveys to begin with:

- ▶ Brophy E, Wang Z, She Q, Ward T **(2023)** Generative adversarial networks in time series: A systematic literature review. ACM Computing Surveys, 55(10), 1-31.
- ▶ Viana D, Teixeira R, Baptista J, Pinto T **(2024)** Synthetic data generation models for time series: A literature review. In Proc. of 2024 IEEE Int. Conf. on Electrical, Computer and Energy Technologies (ICECET), 1-6.
- ▶ Loni M, Poursalim F, Asadi M, Gharehbaghi A **(2025)** A review on generative AI models for synthetic medical text, time series, and longitudinal data. npj Digit. Med., 8(1), 281.
- ▶ Miletic M, Sariyar M. **2025/2026)** Synthetic data generation methods for longitudinal and time series health data: a systematic review. BMC Medical Informatics and Decision Making 26, 30.
- ▶ ... and more in arxiv / as preprints



Evaluating STSGs [Stenger et al., 2024]

Evaluation is key – Human assessment & Machine-computed measures

[Stenger et al., 2024]

Real time series and their synthetic counterparts (Fig. 1)

Figure 1 from [Stenger et al., 2024] has two rows and three columns. The first row shows one real univariate time series, and two generated ones, one good and one bad. The second row shows one multivariate time series with three dimensions, and again one good and one bad synthetic time series generated from it.

Machine-computed workflows include **TRTS & TSTR**

[56] Esteban C, Hyland SL, Rätsch G (2017) Real-valued (Medical) time series generation with recurrent conditional GANs. arXiv preprint. DOI : 10.4855/ARXIV.1706.02633

Machine-computed measures

- at the sample-level (p.32–40) →
- at the distribution-level (p.8–32) ✓ →

Agreement rate (Fig 2)

Figure 2 of [Stenger et al., 2024] show the workflow for computing the agreement between two models. The agreement rate is one of the simplest distribution-level measures for STSG.

Classifier 2 Sample Test (Fig 5)

Figure 5 of [Stenger et al., 2024] shows the workflow for C2ST, which is ultimately for the comparison between two classifiers

Num of statistically Different Bins (Fig 16)

Figure 16 of [Stenger et al., 2024] contains an example of NBD. This measure is on how much the STSG concentrates on particular modes of the data distribution.

Evaluation measures at distribution level and at sample level

[Stenger et al., 2024]

Does the generator just *memorize* (and reproduce) noisy versions of the training data?

Distribution-level measures

- ★ *Distribution of reconstruction errors (DRE)* by Esteban et al [56] ^a
- ★ *Memorization-informed Frechet inception distance (MiFID)* by Bai et al [75]: a 'memorization distance' between the real and the synthetic dataset

^aStenger et al point out that 'While others do this by calculating the dNN_K (for $K = 1$) for all synthetic samples, Esteban et al. [56] suggest a supposedly less computationally expensive approach.' [Stenger et al., 2024, p.17]

Sample-level measures

- * *Authenticity* by Alaa et al [87] is 'the rate at which a generator produces samples that appear novel': the likelihood that the generator delivers a novel sample is the complement of the likelihood that it delivers a memorized one ^a
- * *Nearest neighbor in training (NNT)* by Lin et al [34] computes the 'degree to which the generator is memorizing the data' by computing the K nearest neighbours among the real data for each (of many) randomly selected synthetic sample. ^b

^aAlaa et al use the term 'faithfulness' for the synthetic samples towards the real ones.

^bLin et al [34] use the term 'fidelity'.

Fidelity

'refers to an **individual synthesized** time series to *preserve characteristics, patterns, and noise levels* present in the underlying real time series, often associated with notions of resemblance or typicality' [Stenger et al., 2024, p. 41]

All the measures

[Stenger et al., 2024]

Table 2 shows all the distribution-level measures presented in [Stenger et al., 2024]; Table 3 shows all sample-level measures they presented

All the measures, clustered

[Stenger et al., 2024]

- ▶ Figure 22 depicts the taxonomy of measures from [Stenger et al., 2024]: they have built several clusters under the general concept of 'quality'; *fidelity* is one very important cluster with many measures, most but not all of them at sample level
- ▶ In Table 4 of [Stenger et al., 2024], the authors count how often each of the measures has been reused.
- ▶ Table 5 of [Stenger et al., 2024] shows the role of length and dimensionality for the measures; some measures have no restriction (n.r.) in the sense of allowing time series of different lengths and/or more than one dimensions, but many measures have restrictions.

Closing on STSG evaluation

Insights, good and rich

- ✓ STSG evaluation concerns more than quality towards a downstream task. Some simple workflows have emerged (eg TRTS, TSTR), many distance functions are there to choose upon.
- ✓ Many measures available to assess how realistic is a synthetic sample and to prevent memorization of the training data.
- ✓ Many measures available to evaluate fidelity of individual samples.

Certain properties of time series are challenging:

- ▶ Length – some methods demand constant length
- ▶ Dimensionality – some methods are for univariate time series only
- ▶ Dimensionality – many dimensions $\Rightarrow$ different resolutions $\Rightarrow$ cf. Length
- ▶ Volume – many training data are needed $\Rightarrow$ cf. Claim on slide 4

Can we refrain from some dimensions, chop all time series to the same length, and secure more samples, so that we can use these good methods?

The realm of cost-awareness

- ⦿ The acquisition of features incurs cost - money, energy, time, burden, ...
- ⦿ and so does the recording of modalities with sensors / wearables.
- ⦿ Sensors also incur cost.



WHEN should we minimize costs?

- ▶ PROACTIVELY: find a minimal amount of features / modalities ^a
- ▶ DYNAMICALLY: acquire only the features / modalities needed
- ▶ Coarsening

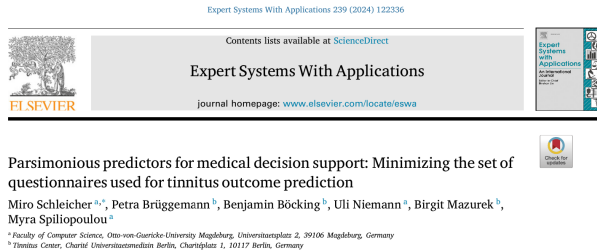
^aFor theoretical guarantees, cf. [He et al., 2024]

Using the least-cost questionnaire(s)

[Schleicher et al., 2024]

Parsimonious prediction of treatment outcome

- ▶ Assess the cost of each questionnaire
- ▶ Identify questionnaires that predict the score of other questionnaires



Cost-ignorant and cost-aware models

Figure 1 of [Schleicher et al., 2024] shows the workflow for building cost-ignorant and cost-aware models from the questionnaires

Table 8 of [Schleicher et al., 2024] shows to what extent some pairs of questionnaires (three pairs shown) predict other questionnaires

- ✓ The scores of some questionnaires are predictable
- ! but not for each patient alike.

Building models for different budgets

[Momeni et al., 2021]

- ▶ **Cost** is energy consumption – by the sensor, to compute a feature, to calculate the bio-parameters needed in the computation
- ⇒ Cost depends on sensors and bio-parameters that are already selected
- ⇒ Build models for different budgets and
 - ▶ Choose for each time window the best model, given performance in the previous window

Building a set of 'constrained' features

Figure 3 of [Momeni et al., 2021] shows the workflow of assessing the cost of features. Their Figure 6 shows the contribution of each feature to the downstream task, and the cost of this feature.

Example scenario

Figure 1 of [Momeni et al., 2021] depicts their example scenario of onbody wearables with sensors

Toggleing to a model of higher confidence and energy consumption^a

Table III of [Momeni et al., 2021] has the title 'Self-aware Stress Detection Models', shows the self-aware configurations S2 and S3, and depicts performance quantities and energy consumption.

^a...but all models must have been built, so that CAFS can pick one

Building a patient-specific featureset in P-CAFE

[Kashani et al., 2025] using MIMIC III

This slide was skipped due to lack of time. It presents the method P-CAFE which iteratively chooses the features needed for each patient to do prediction. In doing so, P-CAFE simulates the iterative information acquisition done by a clinician (cf. their Figure 1). Figure 2 of [Kashani et al., 2025] shows the workflow, and Table 2 shows the performance of P-CAFE and that of other algorithms for the task of mortality prediction on MIMIC III. The feature costs for the considered features are depicted in their Table 8.

Cost vs Contribution of Modalities over time

[Varghese et al., 2026b] using WESAD [Schmidt et al., 2018]

- ★ **Cost** encompasses resource demand – memory, battery, storage, computation power, ...
- ★ **Contribution** is affected by the sampling rate

Insights from a static investigation [Varghese et al., 2026a] →

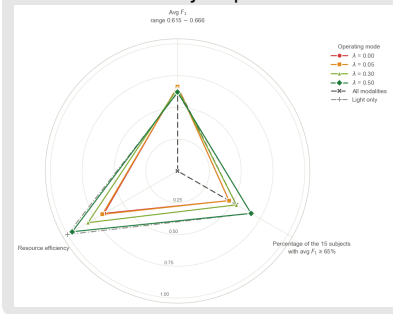
Figure 3 from [Varghese et al., 2026a] shows the contribution of modalities to prediction, for two sampling rates (4 sec and 8 sec), and points out that for one of the sampling rates two of the modalities have a negative influence on prediction quality.

Adding modalities whenever the confidence drops

Figure 2 from [Varghese et al., 2026a] shows the progressive inference policy, used at deployment time



Resource efficiency vs performance



Contribution of Modalities to Multiple Targets

[Manam et al., 2026] using AI4Food [Romero-Tapiador et al., 2023]

Contribution is affected by the sampling rate

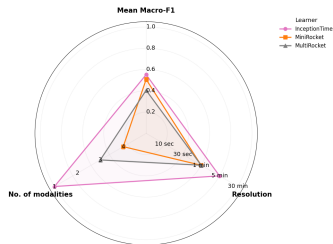
Modalities and their sampling rates (from Table II)

Targets	Anxiety, Depression, Stress
Modalities	Heart rate, Glucose, Oxygen saturation, skin Temperature
Combinations	H; HG; HGO; HGOT
Resolution	10 sec; 30 sec; 1 min; 5 min; 30 min
Learner	RidgeClassifierCV (ROCKET family); Softmax output layer (InceptionTime)

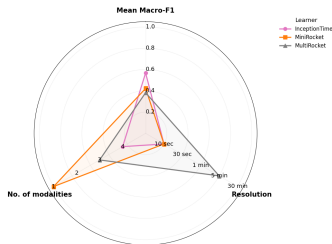
How do temporal representations capture the effects of sampling rate?

Figure 1 from [Manam et al., 2026] shows the workflow for the extraction of the temporal representations used for the prediction task. Three algorithms are used to this purpose: InceptionTime, Multi-Rocket, Mini-Rocket.

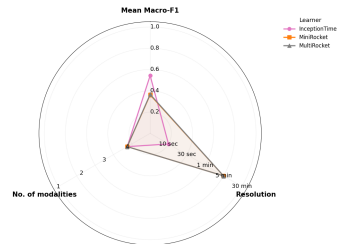
Anxiety - Best Pareto-Optimal Configuration



Stress - Best Pareto-Optimal Configuration



Depression - Best Pareto-Optimal Configuration



Closing

How to deal with emptiness?

- Imputation, STSG
- ⊙ Cost-awareness – proactive vs dynamic

Where to go from here?

- ▶ Minimize the set of subjects needed to train a model
- ▶ Minimize the amount of data per subject



Stopping a learning experiment in less than 240 trials

- ▶ Convergence before trial 100 [Abolfazli et al., 2020]
Figure 5 from [Abolfazli et al., 2020] shows how the cumulative mean sensitivity and specificity curves converge for the different configurations of the auditory category learning experiment.
- ▶ Convergence per subject [Jamaludeen et al., 2022]
This excerpt of Fig 2 from [Jamaludeen et al., 2022] shows the curves of the Granger causalities for four experiment participants; each curve shows how the contribution of one stimulus dimension to prediction evolves over time for each of the participants.

Less of data, rounds & subjects [Varghese et al., 2026c]

- ▶ Refraining from some features
Figure 3 from [Varghese et al., 2026c] shows the contribution of each feature to prediction for each of the two models. The contribution of some features is very low.
- ▶ Reducing the number of subjects *and* of rounds
Figure 5 from [Varghese et al., 2026c] shows the prediction quality of each of the two models when increasing the number of participants and the number of rounds considered for training. The quality curves become flatter as the numbers increase.

... **and push the border of prediction quality.**

The KMD team



Prof. Dr. Myra Spiliopoulou
Lab Head & Professor of Business Information Systems

Research Interests: Medical Mining, temporal data analysis, active learning and Knowledge Discovery in changing environments



Dr. Christian Beyer
Postdoctoral Researcher

Research Interests: Entity-centric learning, prediction & classification, AI in teaching, stream mining, active learning, active feature acquisition



Dr. Miro Schleicher
Postdoctoral Researcher

Research Interests: Medical Mining, Compliance, Adherence and Dropout, Time Series and Trajectories, Prediction and Classification



Malk Büttner
PhD Researcher

Research Interests: Time series & trajectories, predictive maintenance, active learning & active feature acquisition



Syed Muhammad Hamza Zaidi
PhD Researcher

Research Interests: Graph Neural Networks, Sensor Networks, AI based Medical Devices, Explainable AI.
TACTIC



Sarun Varghese
PhD Researcher

Research Interests: Multi-Modal Learning, Human-AI Interaction, Transformers, Cost-aware learning
AICOLABS

We are indepted to our inspiring partners:

- ▶ **UNITI:** *Unification of treatments and Interventions for Tinnitus patients (2020-2024)*, funding from Horizon 2020, Funding Nr. 848261
- ▶ **RheumaMining** (2024-2027) Biomarker und digitales Monitoring zur Prävention der Krankheitsprogression von Rheuma, co-funded by the Federal state of Saxony-Anhalt and the EU (EFRE), Funding Nr. ZS/2023/12/182327
- ▶ **TACTIC:** *Towards Co-Evolution in Human-Technology Interfaces (2024-2027)*, co-funded by the Federal state of Saxony-Anhalt and the EU (ESF+), Funding Nr. ZS/2023/12/182067
- ▶ **AICOLABS:** *AI CoWorking Lab Network (2025-2027)*, co-funded by the Federal state of Saxony-Anhalt and the EU (ESF+), Funding Nr. ZS/2024/01/183362

Thank you!

Questions ?

Bibliography I

 Abolfazli, A., Brechmann, A., Wolff, S., and Spiliopoulou, M. (2020). Machine learning identifies the dynamics and influencing factors in an auditory category learning experiment. *Scientific Reports*, 10(1):1–12.

 Berrevoets, J., Imrie, F., Kyono, T., Jordon, J., and van der Schaar, M. (2023). To impute or not to impute? missing data in treatment effect estimation. In Ruiz, F., Dy, J., and van de Meent, J.-W., editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 3568–3590. PMLR.

 He, Y., Cheng, R., Balasubramaniam, G., Tsai, Y.-H. H., and Zhao, H. (2024). Efficient modality selection in multimodal learning. *Journal of Machine Learning Research*, 25(47):1–39.

 Jamaludeen, N., Unnikrishnan, V., Brechmann, A., and Spiliopoulou, M. (2022). Discovering instantaneous granger causalities in non-stationary categorical time series data. In *20th Int. Conf. on Artificial Intelligence in Medicine (AIME)*. Springer.

 Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., and Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):1–9.

 Kashani, N., Cohen, M., and Shaham, U. (2025). P-cafe: Personalized cost-aware incremental feature selection for electronic health records. In *Proceedings of 28th European Conference on Artificial Intelligence (ECAI 2025)*, Bologna. Version used: arXiv preprint arXiv:2508.08646.

 Luo, Y. (2022). Evaluating the state of the art in missing data imputation for clinical data. *Briefings in Bioinformatics*, 23(1):bbab489.

 Manam, M., Schleicher, M., and Spiliopoulou, M. (2026). Analyzing the impact of temporal resolution on the contribution of wearables to chronic disease monitoring. In *Proc. of 13th IEEE Int. Conf. on Data Science and Advanced Analytics (DSAA 2026)*.

Bibliography II

-  Momeni, N., Valdés, A. A., Rodrigues, J., Sandi, C., and Atienza, D. (2021). Cafs: cost-aware features selection method for multimodal stress monitoring on wearable devices. *IEEE Transactions on Biomedical Engineering*, 69(3):1072–1084.
-  Puga, C., Niemann, U., Schlee, W., and Spiliopoulou, M. (2023). A cost-based multi-layer network approach for the discovery of patient phenotypes. *International Journal of Data Science and Analytics (JD SA)*.
-  Romero-Tapiador, S., Lacruz-Pleguezuelos, B., Tolosana, R., Freixer, G., Daza, R., Fernández-Díaz, C. M., Aguilar-Aguilar, E., Fernández-Cabezas, J., Cruz-Gil, S., Molina, S., et al. (2023). Ai4fooddb: a database for personalized e-health nutrition and lifestyle through wearable devices and artificial intelligence. *Database*, 2023:baad049.
-  Schleicher, M., Brüggemann, P., Böcking, B., Niemann, U., Mazurek, B., and Spiliopoulou, M. (2024). Parsimonious predictors for medical decision support: Minimizing the set of questionnaires used for tinnitus outcome prediction. *Expert Systems with Applications*, 239:122336.
-  Schmidt, P., Reiss, A., Duerichen, R., Marberger, C., and Van Laerhoven, K. (2018). Introducing WESAD, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM international conference on multimodal interaction*, pages 400–408.
-  Stenger, M., Leppich, R., Foster, I., Kounev, S., and Bauer, A. (2024). Evaluation is key: a survey on evaluation measures for synthetic time series. *Journal of big data*, 11(1):66.
-  Unnikrishnan, V. (2017). Analysis of patient evolution on time series of different lengths. Faculty of Computer Science, Otto-von-Guericke Univ. Magdeburg. Master Thesis.

Bibliography III



Varghese, S., Jonnala, K. S., and Spiliopoulou, M. (2026a). Loom net: An importance-aware cnn–transformer for multimodal physiological signals on wearables. In *Proceedings of the 39th IEEE International Symposium on Computer-Based Medical Systems (CBMS)*, Limassol, Cyprus. To appear.



Varghese, S., Jonnala, K. S., and Spiliopoulou, M. (2026b). Multimodal sensing via cost–aware progressive inference for stress detection in wearables. In *Proceedings of the 27th IEEE International Conference on Mobile Data Management (MDM), Industry Track*, Athens, Greece. To appear in the Industry Track.



Varghese, S., Nguyen, T. T., Jonnala, K. S., Lakshmanan, R., Spiliopoulou, M., and Ragni, M. (2026c). How much data is needed for predictive models of consumer choice? In *Proceedings of the 24th International Conference on Cognitive Modeling (ICCM)*, page To Appear.