



## Bayesian Classification of Self-Reported Gastrointestinal Conditions among Mathematics Students

Ilham Arifin<sup>1\*</sup>

<sup>1</sup>Department of Mathematics, Universitas Sumatera Utara, Medan, Indonesia  
ilhamtheclever@gmail.com

### ABSTRACT

*This study examines a simple Bayesian classification procedure using questionnaire summaries from mathematics students at Universitas Sumatera Utara. Four self-reported categories were considered: dyspepsia, gastroesophageal reflux disease, gastritis, and Helicobacter pylori infection. The original charts contain category totals of 13, 24, 3, and 1, respectively, representing 41 questionnaire classifications. Fifteen symptom indicators were tabulated, and a worked example combined six reported symptoms under the naive conditional independence assumption. Recalculation using the manuscript's rounded inputs yielded normalized model probabilities of 4.52% for dyspepsia and 95.48% for gastroesophageal reflux disease, with zero weights for the other categories. These zeros resulted from unsmoothed zero-frequency estimates. A supplementary calculation using chart-derived counts produced 95.38% for gastroesophageal reflux disease; adding one to each binary symptom count reduced this value to 80.18%. The comparison shows that numerical confidence depends on estimation choices in sparse categories. The results demonstrate a reproducible probability calculation, but do not establish diagnostic accuracy. Self-reported labels, overlapping conditions and symptoms, incomplete recruitment information, and the absence of independent clinical validation limit interpretation. The procedure is therefore presented as an educational classification example requiring further data and validation before any clinical application.*

**Keywords:** Bayes' theorem; gastrointestinal symptoms; naive Bayes; student questionnaire; probability

Received: 5 July 2025

Revised: 5 August 2025

Published: 30 September 2025

### INTRODUCTION

Upper gastrointestinal complaints include symptoms that can occur in more than one condition. Gastroesophageal reflux disease (GERD) concerns reflux into the esophagus, whereas dyspepsia denotes a symptom complex that requires a distinct clinical assessment (Katz et al., 2022; Moayyedi et al., 2017). Helicobacter pylori infection is also related to gastric inflammation, so infection and gastritis should not be treated as inherently exclusive

clinical states (Malfertheiner et al., 2022). This distinction matters when a questionnaire requires respondents to select one label.

Student health knowledge provides a relevant context for discussing these terms. A study at Universitas Airlangga examined student knowledge of dyspepsia, gastritis, GERD, and antacids (Nurhaidah et al., 2021). Such work supports attention to terminology and health literacy, but does not establish disease prevalence among students at another institution. In the present study, the available observations concern mathematics students at Universitas Sumatera Utara and are interpreted as self-reported classifications.

Bayes’ theorem links prior probabilities with observed evidence. Previous work has applied it in a gastric-disease expert system (Murni & Riandari, 2018), while a recent review describes broader uses of Bayesian networks in diagnosis and prognosis (Polotskaya et al., 2024). These applications motivate computational exploration; they do not validate a new model automatically. Diagnostic interpretation also depends on the starting probabilities and the context in which evidence is collected (Maguire & Looi, 2025).

The present objective is to reproduce and examine a four-category naive Bayes calculation from the reported questionnaire summaries. The analysis distinguishes model probabilities from diagnostic accuracy, identifies inconsistencies between the tables and charts, and evaluates how rounded inputs and zero-frequency handling affect one worked example. Its immediate contribution is methodological and educational. Bayesian data-analysis teaching can connect probability concepts with substantive contexts (Dogucu et al., 2025); no measured community-service intervention or improvement in student health is claimed here.

METHODS

Study material and reporting scope. The source questionnaire asked mathematics students who reported previous stomach-related illness to identify a condition and associated symptoms. The four available labels are defined in Table 1. The embedded charts give category totals of 13, 24, 3, and 1; their sum is 41. These totals describe the available questionnaire classifications. They are not estimates of population prevalence, and clinical confirmation of the labels is not documented.

Table 1. Questionnaire categories and chart-derived totals

| Code | Self-reported label | Count | Proportion (%) |
|------|---------------------|-------|----------------|
| X1   | Dyspepsia           | 13    | 31.71          |
| X2   | GERD                | 24    | 58.54          |
| X3   | Gastritis           | 3     | 7.32           |
| X4   | H. pylori infection | 1     | 2.44           |

Note. Counts are taken from the original charts (total = 41); percentages are rounded independently.

Recruitment dates, sampling procedures, response rate, participant characteristics, questionnaire validation, consent procedures, and an ethics-review determination are not documented in the available study record. Individual response records were also unavailable for this analysis. Accordingly, the calculations use aggregate counts and probabilities and cannot assess missingness or within-person symptom dependence. Transparent reporting of these design details is needed for an observational study (von Elm et al., 2007).

Symptom coding. Table 2 retains the fifteen source indicators. Wording has been standardized without merging potentially overlapping items. In particular, Y1 denotes epigastric pain; heartburn is Y7. Y4 and Y15 both describe bloating, while Y1 and Y12 may overlap anatomically. These distinctions were not validated, so the codes should be regarded as questionnaire items rather than independent clinical measurements.

**Table 2. Symptom indicators retained from the questionnaire**

| Code | Symptom                   |
|------|---------------------------|
| Y1   | Epigastric pain           |
| Y2   | Nausea                    |
| Y3   | Vomiting                  |
| Y4   | Bloated stomach           |
| Y5   | Early satiety             |
| Y6   | Loss of appetite          |
| Y7   | Heartburn                 |
| Y8   | Bitter taste in the mouth |
| Y9   | Frequent belching         |
| Y10  | Dry cough                 |
| Y11  | Postprandial fullness     |
| Y12  | Upper abdominal pain      |
| Y13  | Weight loss               |
| Y14  | Gastric burning sensation |
| Y15  | Feeling bloated           |

Note. Similar items remain separate because participant-level responses and a validated coding rule are unavailable.

Probability specification. Let  $X_i$  denote category  $i$  and let  $S$  denote the selected set of present symptoms. The source uses the naive Bayes approximation in Equation (1), followed by normalization in Equation (2). Here  $\pi(i)$  is the category prior and  $\theta(j, i)$  is the probability of symptom  $j$  conditional on category  $i$ . This factorization assumes conditional independence of the selected indicators within a category (Hand & Yu, 2001; Friedman et al., 1997).

$$w(i) = \pi(i) \prod \theta(j, i), \quad j \in S \quad (1)$$

$$p(i) = \frac{w(i)}{w(1) + w(2) + w(3) + w(4)} \quad (2)$$

The worked example uses  $S = \{Y1, Y2, Y3, Y6, Y8, Y9\}$ : epigastric pain, nausea, vomiting, loss of appetite, bitter taste in the mouth, and frequent belching. The other nine indicators are unspecified and omitted, rather than treated as absent. A model using explicitly absent symptoms would also require their complementary likelihoods. The reported prior values sum to 0.999 because of rounding; dividing each by 0.999 leaves the normalized result unchanged.

The main calculation preserves the rounded probabilities in Table 3, except that the Y13 discrepancy is displayed explicitly. The GERD chart reports two occurrences of Y13 among 24 classifications, corresponding to 0.0833, whereas the original probability table lists 0.833. The chart-based value is used for the descriptive display. Y13 is not part of the worked example, so this discrepancy does not change its primary result.

**Table 3. Reported priors and conditional symptom probabilities**

| Input | X1    | X2     | X3    | X4    |
|-------|-------|--------|-------|-------|
| Prior | 0.317 | 0.585  | 0.073 | 0.024 |
| Y1    | 0.384 | 0.750  | 0.666 | 0.000 |
| Y2    | 0.769 | 0.791  | 0.333 | 1.000 |
| Y3    | 0.615 | 0.375  | 0.000 | 0.000 |
| Y4    | 0.461 | 0.583  | 0.333 | 0.000 |
| Y5    | 0.230 | 0.250  | 0.666 | 0.000 |
| Y6    | 0.384 | 0.166  | 0.333 | 0.000 |
| Y7    | 0.076 | 0.166  | 0.000 | 0.000 |
| Y8    | 0.076 | 0.375  | 0.333 | 0.000 |
| Y9    | 0.076 | 0.333  | 0.666 | 0.000 |
| Y10   | 0.076 | 0.000  | 0.000 | 0.000 |
| Y11   | 0.153 | 0.333  | 0.000 | 0.000 |
| Y12   | 0.461 | 0.333  | 0.333 | 0.000 |
| Y13   | 0.076 | 0.0833 | 0.333 | 0.000 |
| Y14   | 0.153 | 0.208  | 0.000 | 0.000 |
| Y15   | 0.384 | 0.208  | 0.333 | 0.000 |

Note. All values retain the source table's rounding except Y13 for X2: the chart gives  $2/24 = 0.0833$ ; the original table lists 0.833. This item is excluded from the worked example.

Reproducibility and sensitivity analysis. The six selected likelihoods and the prior can be placed in rows 2–8 of an Excel column. The unnormalized score is `=PRODUCT(B2:B8)`; if the four scores occupy B9:E9, the normalized value in B10 is `=B9/SUM($B$9:$E$9)`. Copying across the four columns reproduces the reported calculation. This is a deterministic calculation, not a stochastic simulation.

Two supplementary calculations assess numerical sensitivity using the counts stored in the original charts. The first uses  $\pi(i) = n(i)/41$  and  $\theta(j, i) = k(j, i)/n(i)$ . The second retains those priors but uses add-one smoothing for each binary symptom, as shown in Equation (3). The smoothing calculation is an explicitly selected sensitivity scenario; it is not an empirically validated replacement model. No new observations are introduced.

$$\theta(j, i) = \frac{k(j, i) + 1}{n(i) + 2} \quad (3)$$

No held-out sample, independent clinical reference standard, or participant-level predicted outcomes are available. Therefore, diagnostic accuracy, sensitivity, specificity, calibration, and clinical benefit are not estimated. The distinction between model development and evaluation follows the reporting aims of TRIPOD and TRIPOD+AI (Collins et al., 2015; Collins et al., 2024). A study claiming diagnostic accuracy would additionally need the reference-standard and participant-flow information addressed by STARD (Bossuyt et al., 2015).

RESULT AND DISCUSSION

Distribution of reported categories and symptoms. Table 1 shows that GERD accounts for 24 of the 41 questionnaire classifications (58.54%). Dyspepsia accounts for 13 (31.71%), gastritis for three (7.32%), and H. pylori infection for one (2.44%). The last two groups are too sparse to support stable estimates for fifteen symptoms. These proportions characterize the recorded labels, not clinically confirmed disease frequencies.

Figures 1–4 preserve the original chart counts, with compact symptom codes for readability. Figure 1 shows that nausea was recorded in 10 of the 13 dyspepsia classifications. Figure 2 shows 19 reports of nausea and 18 of epigastric pain in the GERD category. Counts cannot be compared as conditional frequencies without accounting for the different category totals.

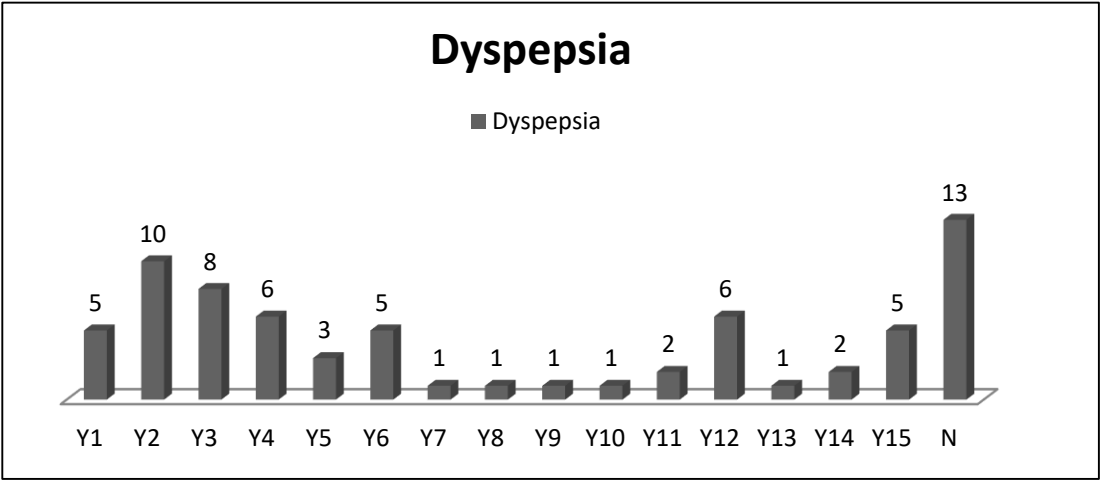


Figure 1. Dyspepsia (n = 13): original reported symptom counts.

Note. Y1–Y15 follow Table 2; N denotes the category total, not an additional symptom.

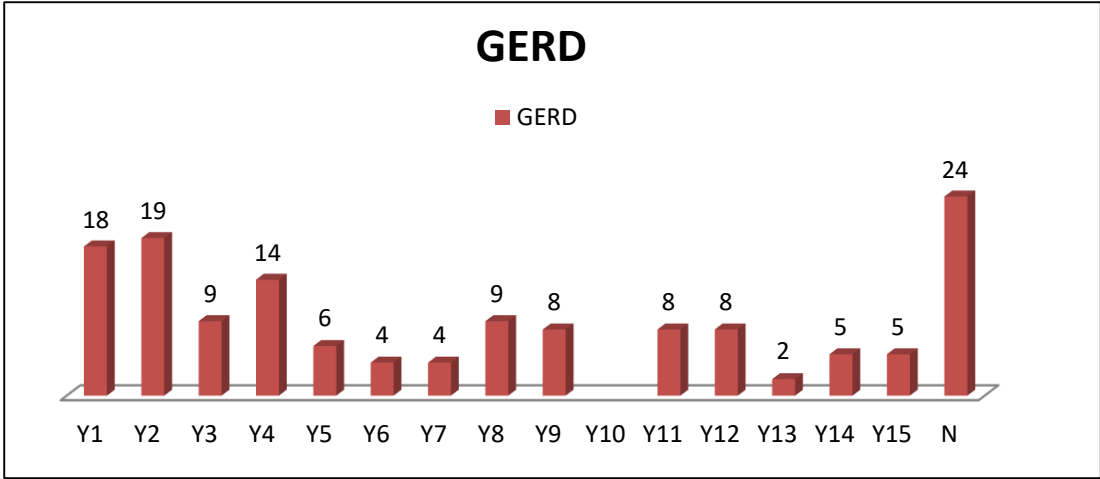


Figure 2. GERD (n = 24): original reported symptom counts.

Note. Y1–Y15 follow Table 2; N denotes the category total, not an additional symptom.

Figure 3 contains at most two observations for any individual symptom in the gastritis category. Figure 4 represents only one H. pylori classification, with nausea recorded. A zero bar in these small groups means that the symptom was not represented in the available aggregate counts; it is not evidence that the symptom is impossible in the condition.

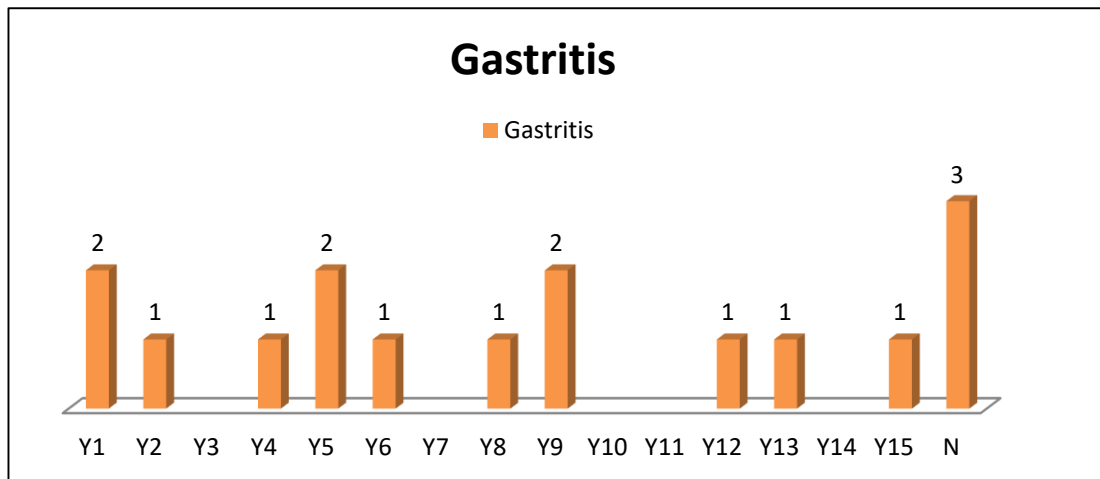


Figure 3. Gastritis (n = 3): original reported symptom counts.

Note. Y1–Y15 follow Table 2; N denotes the category total, not an additional symptom.

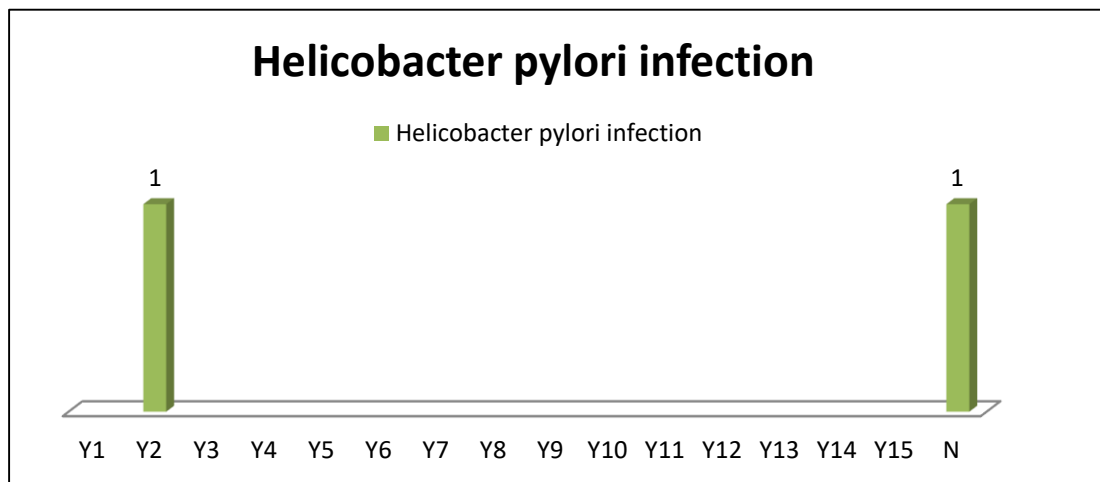


Figure 4. H. pylori infection (n = 1): original reported symptom counts.

Note. Y1–Y15 follow Table 2; N denotes the category total, not an additional symptom.

Worked example. Table 4 presents the six likelihoods used with the rounded priors. The product for dyspepsia is 0.000127688078, and that for GERD is 0.002697792047. The other two products are zero because at least one selected symptom has a zero conditional frequency. The denominator is 0.002825480125. Normalization gives 4.52% and 95.48% for dyspepsia and GERD, respectively.

**Table 4. Six-symptom calculation using the original rounded inputs**

| Input or result | X1          | X2          | X3    | X4    |
|-----------------|-------------|-------------|-------|-------|
| Prior           | 0.317       | 0.585       | 0.073 | 0.024 |
| Y1              | 0.384       | 0.750       | 0.666 | 0.000 |
| Y2              | 0.769       | 0.791       | 0.333 | 1.000 |
| Y3              | 0.615       | 0.375       | 0.000 | 0.000 |
| Y6              | 0.384       | 0.166       | 0.333 | 0.000 |
| Y8              | 0.076       | 0.375       | 0.333 | 0.000 |
| Y9              | 0.076       | 0.333       | 0.666 | 0.000 |
| Score           | 0.000127688 | 0.002697792 | 0     | 0     |
| Normalized (%)  | 4.52        | 95.48       | 0.00  | 0.00  |

Note. Scores are prior  $\times$  six likelihoods. Normalized values are model outputs, not measured diagnostic accuracy.

The 95.48% value is a model probability conditional on the selected symptoms, the four questionnaire labels, the rounded input estimates, and the independence assumption. It is neither the percentage of students correctly diagnosed nor a measured probability of clinical disease. The example has no verified outcome against which a prediction could be scored. The four-category normalization also excludes conditions outside the questionnaire and does not allow simultaneous diagnoses.

Sensitivity to estimation choices. Table 5 compares the original rounded calculation with the two supplementary scenarios. Using exact chart-derived proportions changes the GERD result slightly, from 95.48% to 95.38%. Add-one smoothing produces 80.18% for GERD and nonzero values for all four categories. Thus, the highest-scoring label is unchanged in this example, but the apparent certainty is strongly affected by sparse-data treatment.

**Table 5. Sensitivity of normalized model probabilities to estimation choices**

| Scenario                  | X1 (%) | X2 (%) | X3 (%) | X4 (%) |
|---------------------------|--------|--------|--------|--------|
| Original rounded inputs   | 4.52   | 95.48  | 0.00   | 0.00   |
| Exact chart proportions   | 4.62   | 95.38  | 0.00   | 0.00   |
| Add-one symptom smoothing | 9.82   | 80.18  | 8.34   | 1.66   |

Note. All scenarios use the same six present symptoms. The two supplementary scenarios use priors  $n_i/41$ ; smoothing changes only the symptom likelihoods. Percentages are rounded independently.

The sensitivity results should not be interpreted as evidence that 80.18% is a clinically correct estimate. The smoothing choice modifies the statistical model and is especially influential when a category contains one or three observations. Sample-size planning for prediction research depends on the model and intended precision, rather than a small total being adequate merely because a calculation is possible (Riley et al., 2020). Participant-level data and an appropriate validation design would be needed to compare estimation choices.

Dependence and calibration. Overlapping items can count related information more than once in a naive Bayes product. Research on the classifier shows that useful class assignment can sometimes persist despite violations of independence, but probability estimation and

classification performance are distinct questions (Domingos & Pazzani, 1997). Bayesian network classifiers offer ways to represent additional dependence (Friedman et al., 1997), although the present aggregate data cannot support fitting or validating such a model.

A large normalized score does not demonstrate calibrated risk. Calibration concerns agreement between predictions and observed outcomes (Van Calster et al., 2019). Evaluation should consider calibration and discrimination together and select measures appropriate to a multcategory outcome (Steyerberg et al., 2010). None can be established from the single arithmetic example. PROBAST also directs attention to participants, predictors, outcomes, and analysis when judging risk of bias in prediction-model studies (Wolff et al., 2019). The undocumented recruitment process, self-reported outcomes, and small categories remain material limitations here.

Educational and practical implications. The worked example can help explain prior probabilities, likelihood products, normalization, and the zero-frequency problem. Its usefulness in a teaching or community-education activity would require a separate evaluation of learning outcomes. Clinical usefulness is a further question: decision-curve analysis was developed to assess consequences and net benefit beyond conventional accuracy measures (Vickers & Elkin, 2006). No such benefit is measured in this study, and the spreadsheet should not be presented as a validated diagnostic service.

## CONCLUSION

The reported six-symptom example assigns the largest normalized model probability to GERD: 95.48% using the original rounded inputs. Recalculation from chart counts gives 95.38%, while add-one likelihood smoothing gives 80.18%. These differences show why a high numerical output should be interpreted together with the data size and estimation assumptions. The zero outputs for gastritis and *H. pylori* infection arise from unsmoothed frequencies and do not exclude either condition clinically.

The study provides a reproducible educational example of Bayesian classification. It does not establish diagnostic accuracy or clinical utility. Further work requires consistent symptom definitions, reconciliation of the Y13 source discrepancy, documented recruitment and ethical procedures, clinically appropriate outcome definitions, participant-level data, and independent validation. Any educational or community-service application should assess its own outcomes rather than assume benefit from the calculation alone.

## BIBLIOGRAPHY

- Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L., Lijmer, J. G., Moher, D., Rennie, D., de Vet, H. C. W., Kressel, H. Y., Rifai, N., Golub, R. M., Altman, D. G., Hooft, L., Korevaar, D. A., & Cohen, J. F. (2015). Stard 2015: An updated list of essential items for reporting diagnostic accuracy studies. *BMJ*, *351*, h5527. <https://doi.org/10.1136/bmj.h5527>
- Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., van Smeden, M., Boulesteix, A.-L., Camaradou, J. C., Celi, L. A., Denaxas, S., Denniston, A. K., Glocker, B., Golub, R. M., Harvey, H., Heinze, G., ..., Logullo, P. (2024). Tripod+ai statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, *385*, e078378. <https://doi.org/10.1136/bmj-2023-078378>

- Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. *BMJ*, 350, g7594. <https://doi.org/10.1136/bmj.g7594>
- Dogucu, M., Kazak, S., & Rosenberg, J. M. (2025). The design and implementation of a Bayesian data analysis lesson for pre-service mathematics and science teachers. *Journal of Statistics and Data Science Education*, 33(2), 177–188. <https://doi.org/10.1080/26939169.2024.2362148>
- Domingos, P., & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29(2-3), 103–130. <https://doi.org/10.1023/a:1007413511361>
- Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. *Machine Learning*, 29(2-3), 131–163. <https://doi.org/10.1023/a:1007465528199>
- Hand, D. J., & Yu, K. (2001). Idiot's Bayes—not so stupid after all? *International Statistical Review*, 69(3), 385–398. <https://doi.org/10.1111/j.1751-5823.2001.tb00465.x>
- Katz, P. O., Dunbar, K. B., Schnoll-Sussman, F. H., Greer, K. B., Yadlapati, R., & Spechler, S. J. (2022). ACG clinical guideline for the diagnosis and management of gastroesophageal reflux disease. *American Journal of Gastroenterology*, 117(1), 27–56. <https://doi.org/10.14309/ajg.0000000000001538>
- Maguire, P. A., & Looi, J. C. (2025). Diagnostic testing in psychiatry: Insights and examples from a Bayesian perspective. *Australasian Psychiatry*, 33(1), 162–167. <https://doi.org/10.1177/10398562241300887>
- Malfertheiner, P., Megraud, F., Rokkas, T., Gisbert, J. P., Liou, J.-M., Schulz, C., Gasbarrini, A., Hunt, R. H., Leja, M., O'Morain, C., Rugge, M., Suerbaum, S., Tilg, H., Sugano, K., & El-Omar, E. M. (2022). Management of Helicobacter pylori infection: The Maastricht VI/Florence consensus report. *Gut*, 71(9), 1724–1762. <https://doi.org/10.1136/gutjnl-2022-327745>
- Moayyedi, P. M., Lacy, B. E., Andrews, C. N., Enns, R. A., Howden, C. W., & Vakil, N. (2017). ACG and CAG clinical guideline: Management of dyspepsia. *American Journal of Gastroenterology*, 112(7), 988–1013. <https://doi.org/10.1038/ajg.2017.154>
- Murni, S., & Riandari, F. (2018). Penerapan metode teorema Bayes pada sistem pakar untuk mendiagnosa penyakit lambung. *Jurnal Teknologi dan Ilmu Komputer Prima (JUTIKOMP)*, 1(2), 19–25. <https://doi.org/10.34012/jutikomp.v1i2.226>
- Nurhaidah, F. S., Anugrah, S. D., Putri, A. F., Tukloy, W. D. R., Khairunnisa, S., Primadani, L. H., Wahyudi, T., Aisyia, A., Kamaruzzaman, A. R., Shofa, K. N., & Nita, Y. (2021). Pengetahuan mahasiswa Universitas Airlangga mengenai dispepsia, gastritis, dan GERD beserta antasida sebagai pengobatannya. *Jurnal Farmasi Komunitas*, 8(2), 57–64. <https://doi.org/10.20473/jfk.v8i2.24116>
- Polotskaya, K., Muñoz-Valencia, C. S., Rabasa, A., Quesada-Rico, J. A., Orozco-Beltrán, D., & Barber, X. (2024). Bayesian networks for the diagnosis and prognosis of diseases: A scoping review. *Machine Learning and Knowledge Extraction*, 6(2), 1243–1262. <https://doi.org/10.3390/make6020058>

- Riley, R. D., Ensor, J., Snell, K. I. E., Harrell, F. E., Martin, G. P., Reitsma, J. B., Moons, K. G. M., Collins, G., & van Smeden, M. (2020). Calculating the sample size required for developing a clinical prediction model. *BMJ*, 368, m441. <https://doi.org/10.1136/bmj.m441>
- Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J., & Kattan, M. W. (2010). Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology*, 21(1), 128–138. <https://doi.org/10.1097/ede.0b013e3181c30fb2>
- Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: The Achilles heel of predictive analytics. *BMC Medicine*, 17(1), Article 230. <https://doi.org/10.1186/s12916-019-1466-7>
- Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making*, 26(6), 565–574. <https://doi.org/10.1177/0272989x06295361>
- von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gøtzsche, P. C., & Vandenbroucke, J. P. (2007). The strengthening the reporting of observational studies in epidemiology (STROBE) statement: Guidelines for reporting observational studies. *PLoS Medicine*, 4(10), e296. <https://doi.org/10.1371/journal.pmed.0040296>
- Wolff, R. F., Moons, K. G. M., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., Reitsma, J. B., Kleijnen, J., & Mallett, S. (2019). Probast: A tool to assess the risk of bias and applicability of prediction model studies. *Annals of Internal Medicine*, 170(1), 51–58. <https://doi.org/10.7326/m18-1376>