COMMENT

INVITED COMMENTARY

Definitions of necrotizing enterocolitis: What are we defining and is machine learning the answer?

Camilia R. Martin[1,2]✉

Pediatric Research (2022) 91:488–489; https://doi.org/10.1038/s41390-021-01687-0

The quest to adequately define, diagnose, and manage human disease is recorded as early as 2600 BC as evidenced by the Egyptian Edwin Smith Papyrus.[1] This process continues to be iteratively explored and updated as new approaches (e.g., the scientific method and evidence-based medicine) and new information become available with advancing technologies. Disease classifications can serve many purposes. It can be used to diagnose, predict disease risk, optimize therapeutics, assess morbidity and/or mortality, determine quality of life and health care utilization, and predict long-term medical needs. At the same time, ideally, disease classification should not overdiagnose or overtreat. All these elements are important when considering necrotizing enterocolitis (NEC), the infants it afflicts, and the families it impacts.

It is often discussed that Dr. Martin Bell proposed the first "definition" of NEC in 1978.[2] However, this is not a definition of the disease but rather, in Dr. Bell's words, a "clinical staging" of an entity that was already labeled as NEC. How the diagnosis was made is not fully discussed. In 1986, Drs. Walsh and Kliegman expanded Bell's staging criteria to include additional stages of severity and therapeutic suggestions with true cases defined as the presence of pneumatosis intestinalis or intrahepatic portal venous gas.[3] With this, the term "definite NEC" was assigned to stage IIA and greater and became widely used as a diagnostic definition. However, the modified Bell's staging leaves just enough ambiguity in the certainty of an NEC diagnosis. The persistence of stage 1 in the algorithm raises the question of whether the presence of pneumatosis or intrahepatic air is a required element for diagnosis.[4] The lack of high interobserver reliability of detecting pneumatosis on radiograph or ultrasound complicates this matter further.[5] Other conditions can exhibit pneumatosis or pneumoperitoneum but are not NEC, notably spontaneous intestinal perforation (SIP) for the latter. Although there have been improvements in separating SIP from NEC, other diagnoses can present with pneumatosis, pneumoperitoneum, or severe gastrointestinal symptoms with bloody stools that can be difficult to distinguish from NEC and include intestinal obstruction, vascular compromise, and protein-induced enterocolitis.[6,7] Furthermore, the modified Bell's criteria does not address (nor claims to) the question that nags us the most, why does an infant get NEC? What is a reliable set of risk factors that are modifiable and are subject to practice change and early therapies to minimize the risk of disease? To resolve these issues, it rests on us to determine a robust definition of disease presence and here we enter a continuous imperfect circular argument.

Six definitions have since been proposed in the literature to chip away at our clinical uncertainties and were recently analyzed in a thorough review by Patel et al.[8] The more contemporary definitions include the Vermont Oxford Network definition, the Centers for Disease Control and Prevention definition, the gestational age-specific case definition of NEC (UK), the two out of three rule, the Stanford NEC score, and the International Neonatal Consortium NEC workgroup definition. With now a handful of models to potentially use for clinical and research purposes, should the dominant position of the modified Bell's criteria be replaced?

Lueschow et al. compared the sensitivity, specificity, and accuracy of each definition to make the diagnosis of NEC or no NEC using standard statistics and various machine learning (ML) modeling algorithms.[9] The single-center, retrospective cohort was identified by screening International Classification of Diseases, Ninth Revision (ICD-9) codes for "concern for NEC" and or a "concern for intestinal perforation". The final study cohort consisted of 219 infants and 220 patient events. Their outcome of NEC for which all models would be tested was uniquely defined for this study using the authors' own classification system that included a combination of physical exam findings, laboratory evidence of inflammation, and the presence of pneumatosis intestinalis or portal venous gas. Thus, the performance of each definition was dependent on the performance of the authors' NEC definition that was uniquely adopted for this study but not previously tested or validated.

Ultimately none of the models performed well for both sensitivity and specificity. Although ML optimized model metrics over standard statistics, this was not universally true. Using the top nine clinical features collectively across all models as a new definition, decision tree modeling did not outperform the originally tested definitions, suggesting fewer rather than a larger number of features might be preferred. However, if these very features played a role in the definition of the disease, they are no

[1] Department of Neonatology Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA. [2] Division of Translational Research Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA. ✉email: cmartin1@bidmc.harvard.edu

Received: 19 July 2021 Accepted: 26 July 2021
Published online: 31 August 2021

longer independent variables, and continuing to treat them as such may impact ML model performance.

The authors readily discuss the limitations of their study including single-center, retrospective cohort, reliance on ICD codes for subject identification,[10] lack of healthy controls, missing data, and the absence of potential variables or features of NEC including feeding type and the microbiome. By way of these limitations, the study by Lueschow et al. also reinforces that any definition or ML model is only as good as the data and the definitions applied. If there is not an accepted standardization of how these variables are handled there will always be disparate performances and comparisons among the definitions of NEC developed now and in the future. These lessons are not unique. IBM sought to use its artificial intelligence program to improve treatment options for cancer patients. Despite access to major medical centers and vast data, the project was never realized "frustrated by the complexity, messiness and gaps in the genetic data…".[11] IBM has since abandoned several multimillion-dollar projects in this area. Stating this fact is not to discourage us but rather to reinforce the challenge in harnessing medical data across disparate health care systems and the careful rigor that needs to be applied.

Reviewing these limitations begins to inform what needs to be systematically overcome to adequately apply ML for NEC moving forward. Cooperation between medical centers is vital to study a rare disease such as NEC and amass enough individualized data for model training, testing, and validation. Large collaboratives sharing raw data are needed with agreements on handling data variables, the intent of the prediction and/or definition model, and outcome determinations. The outcomes of interest should be informed by critical stakeholders including families. It is likely that more than one model will be needed depending on the purpose it is trying to achieve. Lessons from prior published literature should be understood as well as emerging recommendations for best practices.[12]

Electronic monitoring systems and medical records in the NICU collect vast amounts of data continuously throughout the day. Yet, only a few elements are relied upon consistently to deliver ongoing medical care. The promise of harnessing the data is understandable but, in the end, it is still the human user who makes judgments on navigating disparate electronic medical record systems and determining the comprehensiveness and characterization of the independent and dependent variables. Computerized models rely on these imperfect systems. Predictive models of NEC and its short- and long-term outcomes need to be carefully defined on their purpose and then repeatedly validated over time. The best chance to achieve reproducibility of models with clinical significance no matter the site of care is to agree on multidisciplinary collaboratives in sharing of data, data handling, and outcome definitions. Well-vetted ML models tested across multiple centers will be needed to achieve generalizability and population-level significance.

REFERENCES

1. Brawanski, A. On the myth of the Edwin Smith papyrus: is it magic or science? *Acta Neurochir.* **154**, 2285–2291 (2012).
2. Bell, M. J. et al. Neonatal necrotizing enterocolitis. Therapeutic decisions based upon clinical staging. *Ann. Surg.* **187**, 1–7 (1978).
3. Walsh, M. C. & Kliegman, R. M. Necrotizing enterocolitis: treatment based on staging criteria. *Pediatr. Clin. N. Am.* **33**, 179–201 (1986).
4. Kliegman, R. M. & Fanaroff, A. A. Neonatal necrotizing enterocolitis in the absence of pneumatosis intestinalis. *Am. J. Dis. Child* **136**, 618–620 (1982).
5. Di Napoli, A. et al. Inter-observer reliability of radiological signs of necrotising enterocolitis in a population of high-risk newborns. *Paediatr. Perinat. Epidemiol.* **18**, 80–87 (2004).
6. Liu, H. & Turner, T. W. S. Allergic colitis with pneumatosis intestinalis in an infant. *Pediatr. Emerg. Care* **34**, e14–e15 (2018).
7. Robinson, A. E., Grossman, H. & Brumley, G. W. Pneumatosis intestinalis in the neonate. *Am. J. Roentgenol.* **120**, 333–341 (1974).
8. Patel, R. M., Ferguson, J., McElroy, S. J., Khashu, M. & Caplan, M. S. Defining necrotizing enterocolitis: current difficulties and future opportunities. *Pediatr. Res.* **88**, 10–15 (2020).
9. Lueschow, S. R., Boly, T. J., Jasper, E., Patel, R. M. & McElroy, S. J. A critical evaluation of current definitions of necrotizing enterocolitis. *Pediatr. Res.* https://doi.org/10.1038/s41390-021-01570-y (2021).
10. Beam, K. S., Lee, M., Hirst, K., Beam, A. & Parad, R. B. Specificity of International Classification of Diseases codes for bronchopulmonary dysplasia: an investigation using electronic health record data and a large insurance database. *J. Perinatol.* **41**, 764–771 (2021).
11. Lohr, S. Whatever happened to IBM's Watson? *The Boston Globe* (7 July 2021).
12. Roberts, M. et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat. Mach. Intell.* **3**, 199–217 (2021).

COMPETING INTERESTS

Dr. Martin serves on the advisory boards of The NEC Society, LactaLogics, Inc., and Plakous Therapeutics, Inc. Dr. Martin has consulted with Mead Johnson Nutrition and has a research grant from Mead Johnson Foundation. None of these entities had a role in the writing of this article.

ADDITIONAL INFORMATION

Correspondence and requests for materials should be addressed to C.R.M.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.