Early View

Editorial

The perils of premature phenotyping in COVID: a call for caution

Lieuwe D. J. Bos, Pratik Sinha, Robert P. Dickson

Title: The perils of premature phenotyping in COVID: a call for caution

Author: Lieuwe D. J. Bos[1,2], Pratik Sinha[3,4], Robert P. Dickson[5,6,7]

Affiliations: 1. Intensive Care, Amsterdam University Medical Centers, location AMC, University of Amsterdam, Infection and Immunity, Amsterdam, The Netherlands. 2. Department of Respiratory Medicine, Amsterdam University Medical Centers, location AMC, University of Amsterdam, Infection and Immunity, Amsterdam, The Netherlands. 3. Department of Medicine, University of California, San Francisco San Francisco, California. 4. Department of Anesthesia, University of California, San Francisco San Francisco, California. 5. Division of Pulmonary and Critical Care Medicine, Department of Internal Medicine, University of Michigan Medical School, Ann Arbor, Michigan; 6. Department of Microbiology and Immunology, University of Michigan Medical School, Ann Arbor, Michigan; 7. Michigan Center for Integrative Research in Critical Care; Ann Arbor, Michigan

Corresponding Author Information:

Lieuwe D. J. Bos

l.d.bos@amsterdamumc.nl

Word count: 1579

Figures: 1

Tables: 1

References: 15

Body

The Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2) poses an unprecedented global healthcare challenge. Severe novel Coronavirus disease (COVID-19) pneumonia frequently causes hypoxemic respiratory failure, manifesting in the acute respiratory distress syndrome (ARDS). Recently, authors have proposed distinct clinical phenotypes of COVID-19 pneumonia in several influential, high-profile essays[1-3]. For example, in a recent Perspective in this journal [3], authors speculated that COVID-19 has five phenotypic presentations: three phenotypes based on severity of hypoxemia and need for supportive care (no hypoxemia, mild hypoxemia, and moderate hypoxemia), and two phenotypes of severely hypoxemic patients based on additional physiologic and clinical features. Aligned with other recent efforts to phenotype COVID patients [1, 2], the authors subtyped patients into a supposedly prevalent phenotype with normal compliance, low lung weight, and predominant perfusion abnormalities ("L"-phenotype), and a less-prevalent phenotype with more typical features of ARDS such as profound consolidation and low compliance ("H"-phenotype). The authors advocate for distinct management strategies for these purported phenotypes, include permitting increased tidal volumes and restricted positive end-expiratory pressure in the "L" phenotype patients.

The urge to phenotype patients with COVID-19 pneumonia is understandable and relatable. Outside of critical care medicine, the past decade has been characterized by major advances in precision medicine, promising tailored therapies based on individual patients' physiological and biological characteristics. The emergence of a novel disease without effective treatment incentivizes heuristic-based identification of subsets of patients who may respond similarly to a particular intervention. Yet this temptation to define phenotypes based on early clinical experience should be resisted. By prematurely

phenotyping patients, we risk causing considerable harm and generating more static than signal. In this Perspective, we provide four arguments against premature phenotyping, discuss the features of responsible phenotyping, and recommend a path forward in advancing our understanding of the true heterogeneity underlying patients with COVID-19.

The first - and simplest - argument against premature phenotyping is that our initial intuitions are often wrong. As a vivid example, a prominent essay [2] recently asserted without qualification that *"soon after onset of respiratory distress from COVID-19, patients initially retain relatively good compliance despite very poor oxygenation."* This claim, while not supported by references cited, formed the basis for extended discussions of the pathophysiology and tailored management of patients with this purported "L phenotype" of COVID-19 (discussed above). Yet subsequent cohort studies [4, 5] have demonstrated that lung compliance in COVID-19 patients is in fact quite low, entirely congruent with non-COVID-19 ARDS cohorts [6-8], and normally distributed along a continuum rather than existing as discrete phenotypes. Further, purported radiographic and physiologic features of these phenotypes (e.g. dense airspace filling on CT scans paired with decreased compliance in the "H" phenotype) have subsequently been shown to be entirely uncorrelated with each other[9]. Identification of clinical phenotypes - and speculation regarding their underlying biology - should be deferred until after careful, objective inspection of adequately sized cohorts. Human intuitions are simply too fallible – and clinical experience too contingent and heterogenous – to reliably identify phenotypes without sufficient data.

A related argument against premature phenotyping is that it exacerbates our inherent susceptibility to cognitive biases. Once we are informed of clinical categories (however

false they may be), our brains treat them as real and begin selectively filtering our observations. As an example, following dissemination of the since-disproven claim that COVID-19 patients have preserved lung compliance, the myth was reinforced by common cognitive traps. The Baader-Meinhof phenomenon (also called the "frequency illusion") ensured that once clinicians were prompted to notice COVID-19 patients with near-normal lung compliance, they began noticing them everywhere (when in fact their frequency was no higher than in non-COVID ARDS) [6-8]. Similarly, clinicians could dismiss low-compliance COVID-19 cases by unintentionally committing the "no true Scotsman" fallacy: by dismissing away purported exceptions on an *ad hoc* basis, claiming that low-compliance COVID-19 cases must be atypical, as "real COVID-19" has near-normal respiratory mechanics. If we do not insist on data-driven phenotypes, our cognitive biases guarantee that we'll end up with phenotype-driven data.

A third argument against premature phenotyping is that it distracts us from sound, evidence-based practices. Clinical outcomes in ARDS have improved markedly in recent decades [10], driven not by blockbuster drug discoveries, but rather by incremental improvements in the delivery of supportive care. These slow but cumulative advances have been built on hard-won lessons from rigorous randomized controlled trials. By their design, these trials have "lumped" heterogenous ARDS patients together under a syndrome-based definition. Despite this, these trials have provided the field with an extensive literature informing evidence-supported therapies. By presumptuously splitting COVID-19 patients into false phenotypes – and by recommending "tailored management" based on untested physiologic intuitions – authors have advocated for abandonment of what remains our most effective tool against COVID-19: meticulous, evidence-driven critical care delivery.

A final argument against premature phenotyping is that it worsens the already-unfavorable ratio of signal and noise in the ICU. At the bedside, critical care physicians must filter, process, and interpret a tremendous stream of data generated by every patient: physiological, biochemical, radiographic, etc.. Clinicians must synthesize these findings with the published literature, which is similarly daunting: more than 10,000 PubMed-indexed manuscripts on COVID-19 were published in the first four months of 2020. This deluge of information threatens the most overlooked and precious resources in the ICU: clinicians' attention, time, and bandwidth. By needlessly clouding the clinical picture, false phenotypes consume time on rounds and distract us from more immediate concerns. As a field, our research prioritization has been similarly clouded: investigators' time and resources are squandered trying to explain the biology underlying clinical phenomena that, upon inspection of patient data, simply do not exist.

So what does responsible phenotyping look like? As with any scientific experiment, there needs to be an explicit purpose as to why we seek phenotypes. In medical science, this mandate ultimately converges on improving patient outcomes (although gleaming novel biological and clinical insights is an equally important motivating factor as they may be critical to achieving this goal). To that end, our field has recent examples of empirically-derived phenotypes that have been successfully used to identify treatment-responsive and/or biologically distinct subgroups. In asthma for example, using a data-driven unbiased clustering approach, two distinct phenotypes of asthma were identified based on interleukin-13 inducible gene-expression. [11] The phenotype signature specific to high gene-expression was later shown, in a randomised-controlled trial (RCT), to be responsive to a monoclonal antibody that specifically inhibits interleukin-13 activity [12]. In ARDS, again using unbiased clustering methods, two phenotypes have been identified with distinct biological and clinical characteristics, consistent across five RCTs,

and with markedly different clinical outcomes [13]. Importantly, in three of these RCTs, divergent treatment responses were observed to randomised interventions. Further, simpler models have been recently described that offer the potential for the clinical application of these phenotypes.[14]

These data-driven approaches to clustering are not impervious to errors and misuse. These are powerful tools and - independent of the validity of the research question or study design - clusters will inevitably emerge. It is, therefore, incumbent on the investigators to demonstrate the validity and utility of the identified phenotypes. In the absence of ground truth, the conditions that optimally surrogate for validity are 1) robustness, 2) consistency and 3) reproducibility in data external from the population from which they were derived. In almost all algorithms, the phenotypes identified are highly subject to the predictor variables. Taking critically-ill COVID-19 patients as a specific example, it is imperative to acknowledge that we are studying complex biological systems in which inter-connected pathways share non-linear associations. Seeking univariate solutions in such populations, therefore, seems unlikely to yield meaningful subgroups other than for prognostication [15]. Further, univariate solutions, particularly when sought prematurely, can be more susceptible to the central limit theorem. This mathematical theorem states that given a sufficiently large sample size, the distribution of means of a variable will converge to a normal distribution, suggesting that continuous variables that appear bimodal with limited data, will become normally distributed over time. Thus both in terms of biologic plausibility and mathematical principles, we are unlikely to derive useful phenotypes if we anchor on simplistic, one-dimensional features of disease.

To avoid these pitfalls, predictor variables in multivariate models should be selected with the research question in mind and be highly informative in terms of effectively splitting the population. Moreover, the use of these complex data-science algorithms, intended to overcome cognitive bias, will be limited to theorising unless they are accompanied by a measurement system that can identify the phenotypes rapidly and consistently. Regardless of the motivation or approach used, phenotyping in critical care is typically a data-hungry exercise, and in studies currently purporting COVID-19 phenotypes, the requisite quantity and quality of data are regrettably lacking. Ultimately, however, the true success of phenotypes in diseases will be judged by the identification of actionable interventions. In critical care, although many examples of heterogeneous treatment effects in phenotypes have been described in secondary analyses, their efficacy will need testing via randomized controlled trials (RCTs). The mere identification of disease phenotypes – whether derived prematurely or responsibly – should not itself change clinical practice, but instead inform prospective, "phenotype-aware" trials.

In summary, the COVID-19 pandemic has posed novel challenges for clinicians and researchers. While we share the ultimate goal of tailoring therapies to the specific pathophysiology of each patient's condition, it is imperative that we first objectively collect, collate, and interpret sufficient data to "type" and understand the disease comprehensively. By prematurely phenotyping patients with COVID-19, we expose ourselves and our patients to considerable and preventable risk.

1

Table

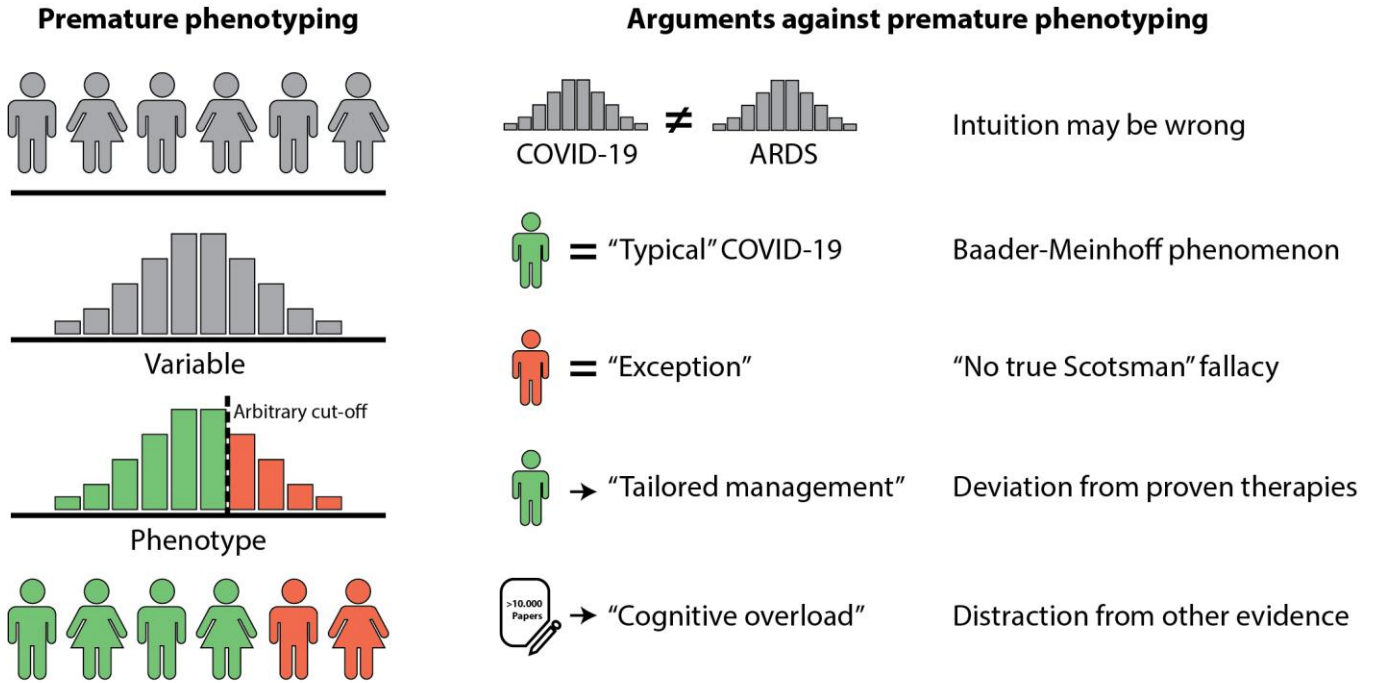|	Premature phenotyping	Responsible phenotyping
Timing	Identification of phenotypes precede systematic collection of data	Systematic collection of data precedes identification of phenotypes
Derivation	"Top-down": theory informs data collection and interpretation	"Bottom-up": unbiased data collection informs subsequent theory
Size of derivation cohort	Derived from limited, single-center experiences	Derived from large, multi-center cohorts
Validation	Contingent on the specific population and biases of the observer	Robust across multiple cohorts, independent of observer
Internal coherence	Features are inconsistently associated with each other across patients (e.g. physiology and imaging)	Phenotype-defining features, by design, are internally consistent in their relationship to each other
Relation to underlying biology	Reflects a simplistic or incorrect model of underlying pathophysiology	Reflects meaningful differences in underlying pathophysiology
Complexity	Anchored on one or two clinically apparent variables	Incorporates high-dimensional data with non-linear interactions between variables
Discreteness, continuity	Purported phenotypes are merely extremes of a normally distributed continuum	Phenotypes represent distinct "clusters," meaningfully divergent from each other and their attributes cannot be explained by a single variable
Impact on practice	Immediately used to justify changes in practice	Prospectively tested in phenotype-informed clinical trials

Figure

FIGURE 1 The perils of premature phenotyping. By focusing on extremes of a normally distributed continuum, we risk creating arbitrary "phenotypes" that are not representative of meaningful underlying differences in pathophysiology. Premature phenotyping is often based on erroneous initial impressions, and contributes to cognitive biases including the Baader-Meinhoff phenomenon (the "frequency bias") and the "No true Scotsman" fallacy (excluding incompatible observations via an ad hoc purity test). Premature phenotyping can compromise the delivery of care by inspiring deviation from evidence-based practices as well as contributing needlessly to the cognitive load of clinicians.

References

1. Gattinoni L, Chiumello D, Caironi P, Busana M, Romitti F, Brazzi L, Camporota L. COVID-19 pneumonia: different respiratory treatments for different phenotypes? *Intensive Care Med* 2020.

2. Marini JJ, Gattinoni L. Management of COVID-19 Respiratory Distress. *JAMA* 2020.

3. Rello J, Storti E, Belliato M, Serrano R. Clinical phenotypes of SARS-CoV-2: Implications for clinicians and researchers. *Eur Respir J* 2020.

4. Bhatraju PK, Ghassemieh BJ, Nichols M, Kim R, Jerome KR, Nalla AK, Greninger AL, Pipavath S, Wurfel MM, Evans L, Kritek PA, West TE, Luks A, Gerbino A, Dale CR, Goldman JD, O'Mahony S, Mikacenic C. Covid-19 in Critically Ill Patients in the Seattle Region - Case Series. *N Engl J Med* 2020.

5. Ziehr DR, Alladina J, Petri CR, Maley JH, Moskowitz A, Medoff BD, Hibbert KA, Thompson BT, Hardin CC. Respiratory Pathophysiology of Mechanically Ventilated Patients with COVID-19: A Cohort Study. *Am J Respir Crit Care Med* 2020.

6. Guerin C, Reignier J, Richard JC, Beuret P, Gacouin A, Boulain T, Mercier E, Badet M, Mercat A, Baudin O, Clavel M, Chatellier D, Jaber S, Rosselli S, Mancebo J, Sirodot M, Hilbert G, Bengler C, Richecoeur J, Gainnier M, Bayle F, Bourdin G, Leray V, Girard R, Baboi L, Ayzac L. Prone positioning in severe acute respiratory distress syndrome. *N Engl J Med* 2013: 368(23): 2159-2168.

7. Bellani G, Laffey JG, Pham T, Fan E, Brochard L, Esteban A, Gattinoni L, van Haren F, Larsson A, McAuley DF, Ranieri M, Rubenfeld G, Thompson BT, Wrigge H, Slutsky AS, Pesenti A. Epidemiology, Patterns of Care, and Mortality for Patients With Acute Respiratory Distress Syndrome in Intensive Care Units in 50 Countries. *Jama* 2016: 315(8): 788-800.

8. Goligher EC, Combes A, Brodie D, Ferguson ND, Pesenti AM, Ranieri VM, Slutsky AS. Determinants of the effect of extracorporeal carbon dioxide removal in the SUPERNOVA trial: implications for trial design. *Intensive Care Med* 2019: 45(9): 1219-1230.

9. Bos LDJ, Paulus F, Vlaar APJ, Beenen LFM, Schultz MJ. Subphenotyping ARDS in COVID-19 Patients: Consequences for Ventilator Management. *Annals of the American Thoracic Society* 2020.

10. Esteban A, Frutos-Vivar F, Muriel A, Ferguson ND, Penuelas O, Abraira V, Raymondos K, Rios F, Nin N, Apezteguia C, Violi DA, Thille AW, Brochard L, Gonzalez M, Villagomez AJ, Hurtado J, Davies AR, Du B, Maggiore SM, Pelosi P, Soto L, Tomicic V, D'Empaire G, Matamis D, Abroug F, Moreno RP, Soares MA, Arabi Y, Sandi F, Jibaja M, Amin P, Koh Y, Kuiper MA, Bulow HH, Zeggwagh AA, Anzueto A. Evolution of mortality over time in patients receiving mechanical ventilation. *Am J Respir Crit Care Med* 2013: 188(2): 220-230.

11. Woodruff PG, Modrek B, Choy DF, Jia G, Abbas AR, Ellwanger A, Koth LL, Arron JR, Fahy JV. T-helper type 2-driven inflammation defines major subphenotypes of asthma. *Am J Respir Crit Care Med* 2009: 180(5): 388-395.

12. Corren J, Lemanske RF, Hanania NA, Korenblat PE, Parsey MV, Arron JR, Harris JM, Scheerens H, Wu LC, Su Z, Mosesova S, Eisner MD, Bohen SP, Matthews JG. Lebrikizumab treatment in adults with asthma. *N Engl J Med* 2011: 365(12): 1088-1098.

13. Sinha P, Calfee CS. Peeking under the Hood of Acute Respiratory Distress Syndrome Phenotypes: Deeper Insights into Biological Heterogeneity. *Am J Respir Crit Care Med* 2019: 200(1): 4-6.

14. Sinha P, Delucchi KL, McAuley DF, O'Kane CM, Matthay MA, Calfee CS. Development and validation of parsimonious algorithms to classify acute respiratory distress syndrome phenotypes: a secondary analysis of randomised controlled trials. *Lancet Respir Med* 2020: 8(3): 247-257.

15. Lessov-Schlaggar CN, Rubin JB, Schlaggar BL. The Fallacy of Univariate Solutions to Complex Systems Problems. *Front Neurosci* 2016: 10: 267.