TuftsMedicine
Tufts Medical Center

LEAPS

Learning Ecosystems Accelerator for
Patient-centered, Sustainable innovation

WHITE PAPER

A practical approach for defining outcomes and thresholds for predictive healthcare algorithm development using real-world data

March 20, 2023

Tufts Medical Center Institute for Clinical Research and Health Policy
Center for Biomedical System Design

Key Takeaways

- As the healthcare system evolves towards value-based care, predictive algorithms can play a critical role, but their findings must be perceived as meaningful, substantial, and actionable by those outside the data science community.
- Both clinically specific and system impact metrics are feasible and important to address the decision-making needs of payers, as well as patients, providers, and the development team itself.
- A practical, multi-stakeholder, fit-for-purpose metric identification process that is applicable to real-world evidence (RWE) can be executed in only a few months, as was demonstrated by the NEWDIGS LEAPS Project with the development of the METRICS process.
- The METRICS process can be an important tool to ensure alignment and the practical success of machine learning predictive algorithms to improve patient outcomes and value from therapeutic regimens.
- Establishing up-front thresholds for change in the chosen metrics so as to determine targets of meaningfulness to guide the predictive model development process and help ensure its future clinical adoption and reimbursement is a key differentiator of the METRICS process.

About LEAPS

The LEAPS Project seeks to modernize how we plan, produce, and use real world evidence (RWE) in order to optimize drug therapy regimens for patients.

LEAPS seeks to improve patient outcomes in economically sustainable ways through new patient-centered learning healthcare system designs, RWE platform infrastructures, and alignment of incentives across stakeholders.

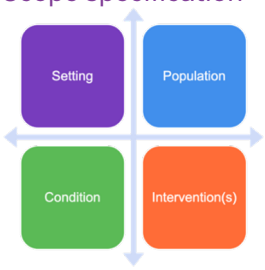
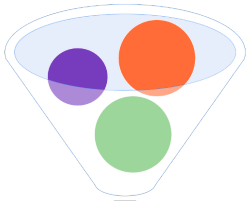
RATIONALE AND BACKGROUND

The Metrics for Evaluation Thresholds & Reimbursement for Incentive Correlation across Stakeholders (METRICS) process framework outlined in this document (see Figure 1 on page 3) was developed to support the identification of outcome targets for predictive modeling and payment of drug therapy regimens, through a defined multi-stakeholder process. A draft process was created and applied in the context of a case study focused on immune checkpoint inhibitors (ICIs) treatment of advanced/metastatic non-small-cell lung cancer (NSCLC). The draft process was then refined and generalized for application in other settings.

The advanced/metastatic NSCLC case study was selected by the NEWDIGS LEAPS Project community in late 2021 as the first in a series of case studies geared towards developing generalizable principles for Downstream System Innovation. NEWDIGS is focused on evolving the post-market biomedical innovation value chain into a collaborative downstream system in order to more efficiently connect and coordinate real-world evidence (RWE) generation, targeted treatment decisions, and reimbursement in ways that are both patient-centered and economically sustainable.

Many efforts to define core outcomes sets have been undertaken in a wide variety of clinical settings. The Core Outcome Measures in Effectiveness Trials (COMET) Initiative database lists more than 1,000 core outcomes set studies, with over 100 cancer studies.[1] These efforts typically involve multiple parties (often focused on clinicians and researchers) proceeding through a stepwise, formal Delphi process, and result in detailed lists of preferred outcome measures for clinical trials.[2] Our purpose was to complete a practical version of this multi-stakeholder consensus process for use by data science teams building predictive models intended for common clinical use and payer reimbursement. The process needed to be applicable to RWE, move quickly enough for the results to have some relevance to an active research team (our ongoing NSCLC use case), and result in a small number of prioritized endpoints for algorithm prediction (ideally, 1 or 2) along with thresholds for the magnitude of effect which would be considered important by decision makers.

Figure 1: METRICS Process Framework

Domain	Proposed LEAPS Process
Scope specification 	**SCOPE:** Align the setting(s) in which the outcomes sets are to be applied • The health condition(s) covered by the outcomes sets • The population(s) covered by the outcomes sets • The intervention(s) covered by the outcomes sets
Stakeholders involved 	**INVITE:** Apply LEAPS stakeholder mapping process to identify those who will use the outcomes sets in practice, RWD analysis, or coverage decisions
Consensus process 	**GATHER** initial list of outcomes considering views of all stakeholders • Collection process should include multi-stakeholder meeting(s), review of existing literature (both trials and RWE studies)
	**FILTER** initial list of metrics for feasibility / importance / duplication • Describe (implicit and explicit) criteria used to create "short list" • Note all measures ranked as important by stakeholders, but not included for practicality / lack of data. Consider proxy measures for important but technically infeasible outcomes.
	**PRIORITIZE** A scoring process and consensus definition is used, described, and refined based on continuous learning process • Scoring and consensus process: modified Delphi with ranking of each option or cumulative voting (multiple votes to be distributed among the options as desired)
	Establish **THRESHOLDS** for action • Thresholds for decision-making (e.g., incremental difference needed) to be sought for each stakeholder category • Care is taken to avoid ambiguity of language used in the list of outcomes

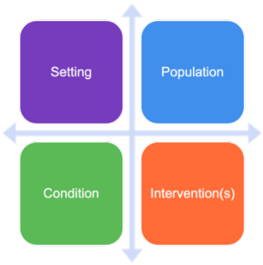
PROCESS

The task established for the METRICS team during multi-stakeholder Design Labs for the LEAPS Project was to collaboratively answer the questions, "what should we measure?" and "how substantial a difference (positive or negative) is needed to spur a change in decision-making?" Two types of measures were considered: Clinical Outcomes specific to the therapeutic context (in this instance the NSCLC use case), and system-wide Impact Metrics.

The team felt it was important that the decision-making process regarding measure selection and threshold setting be made explicit and reproducible, so that the LEAPS community and others can learn from the experience and apply those lessons to future pilots. Subteams were thus established for Clinical Outcomes, Impact Metrics, and Process Documentation.

A multi-stakeholder process was envisioned, using elements of published best practices for consensus methods, and focusing on practicality. As a starting point, we referred to the Core Outcomes Standards (COS-STAD) developed to improve reporting of core outcomes efforts in general and as applied oncology.[3,4] We retained essential steps of a more typical consensus process – defining the setting and stakeholders, gathering possible outcome measures, and applying rounds of filtering and prioritization – while maintaining a rapid cycle of review, decision, and application (approximately 5 months in total). The process was also modified by adding a step to define thresholds for action, as this was felt to be critical to make the results useful to analysts creating a predictive model.

Domain 1. Scope



The scope of the ICIs in the NSCLC use case was outlined at a multi-stakeholder meeting in June 2022 and refined through team discussion. The US health care system was the relevant setting for the initial use case, with adult advanced/metastatic NSCLC patients, regardless of insurance coverage, US region, or practice setting as the population of interest. Immune checkpoint inhibitors were proposed as the intervention(s) covered by the use case. However, feedback from the LEAPS community implied expansion to comparator treatments as well.

Domain 2. Invite

The METRICS team had the benefit of being situated within the existing LEAPS Project community, and patient, clinician/provider, payer, drug developer, and analyst representatives with interest in oncology topics were invited to participate. The LEAPS Project utilizes a stakeholder mapping process, which is a key component of the MIT NEWDIGS system design toolkit. Specifically, the LEAPS Stakeholder Mapping provides a structured process for identifying personnel and organizations with functional expertise and authority whose perspectives are essential to understand for success, including:

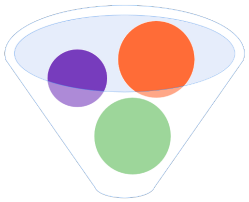
- End-User Value
- Feasibility of Implementation, Scaling, and Sustainability

Domain 3. Consensus process

Gather

The Clinical Outcomes subteam gathered outcome measures specific to advanced/metastatic NSCLC, including dozens from clinical trials. The Impact Metrics team underwent a similar process, including a broad assessment leveraging RWE studies, quality initiatives, and feedback from discussion in the prior Design Lab, both within and outside the oncology setting. Both teams considered the rationale for including various metrics as well as their feasibility, including potential RWE sources and how outcomes would be measured.

Filter



METRICS subteams worked to develop short lists of NSCLC-specific clinical outcomes and more generalizable, system-wide impact measures. Upon review, 8 candidate measures for each category were presented to the full METRICS team for filtering on importance and feasibility (Appendix A). Voting took place at a virtual meeting with 11 participants. Measures which received fewer than 2 votes for importance or feasibility were removed from the short list for further discussion.

Prioritize

At least one round of group voting or ranking is conducted as part of accepted consensus processes.[5] In addition to the filtering votes conducted above, the team collaboratively prioritized the short list options using a cumulative voting process (each participant received 5 votes to distribute as they saw fit).[6] For our NSCLC use case, this round of voting was held in-person at a LEAPS Project Design Lab in November 2022. An alternate method would be ranking or scoring of each option by each voter; this was thought to be somewhat cumbersome to accurately tally and report out in real time.[7] The prioritization voting was conducted following a discussion of pros and cons of each option, how best to define and operationalize the outcomes measures, and which outcomes and metrics were most likely to be useful for decision-making.

Determine thresholds for action

At the same November 2022 Design Lab, a collaborative exercise in threshold-setting was conducted with both METRICS team members and other LEAPS participants. This step was intended to determine at what magnitude sub-population differences identified by a predictive algorithm would be sufficiently large to advance the algorithm towards clinical practice. Specifically, the team discussed what result in a subpopulation, compared to the total ICIs population, would impact a:

- Coverage decision by a payer?
- Treatment recommendation by a provider?
- Treatment decision by a patient?

The discussion was framed with a hypothetical continuum of machine learning algorithm predictions for overall survival for an ICI-treated subpopulation in advanced/metastatic NSCLC (see Figure 2 on page 6).

FINDINGS IN NSCLC USE CASE

The preferred clinical outcome measure was Overall Survival and the preferred impact metric was Time to Effective Treatment. Interestingly, the first-place choice in each category was preferred overwhelmingly, with more than twice the votes of the next listed. Overall survival was ranked more feasible to measure with RWE and more meaningful for patients than some of the surrogate endpoints often used in cancer trials. The longer period of observation needed compared to, for example, objective response was not considered a barrier to adoption in the advanced/metastatic NSCLC setting, whereas for other more indolent cancers it may be less appropriate.

For impact metrics, participants assessed the available outcomes as involving more tradeoffs. The group coalesced around the time to effective treatment measure (from diagnosis to first receipt of a therapy which works, defined by continuing treatment) in part because of significant discomfort with the Total Cost of Care metric. Total cost was

thought to vary too much by stakeholder and was too difficult to interpret. Specifically, a decrease in total cost of care could represent efficient/optimal use of resources, on the one hand, or the use of sub-par or older treatment options that may not represent the best care for a patient, or rapid decline and death, or some mixture of these. This complexity challenged our ability to establish a meaningful threshold for what a positive or negative change would look like.

RELEVANCE TO DOWNSTREAM SYSTEM INNOVATION

Several observations on the outcomes selection process emerged from the group discussion which seem to apply beyond the advanced/metastatic NSCLC setting. These points should be considered in development of better outcomes prediction capacity using RWE:

- **Standard of care comparison:** Most participants agreed on the necessity of comparing results for a subpopulation to a (matched/adjusted) "standard of care" benchmark, not only to the whole treated population.
- **Asymmetric thresholds:** As thresholds for meaningful change were elicited using various scenarios, it was clear that reactions differed for positive and negative differences (yellow arrows in Figure 2): on the "less effective than median" side, decisions would not change until the result for an ICI-treated subgroup was down at or below the previous standard of care; in contrast, for the plus side, the group felt that improvements of 30% or even 100% would be needed to change decisions on the part of a provider, patient, or payer.
- **Magnitude:** It was notable that multiple participants expressed that the magnitude of the proposed thresholds for decision-making was greater than the size of statistically significant differences observed in typical RWE analyses.

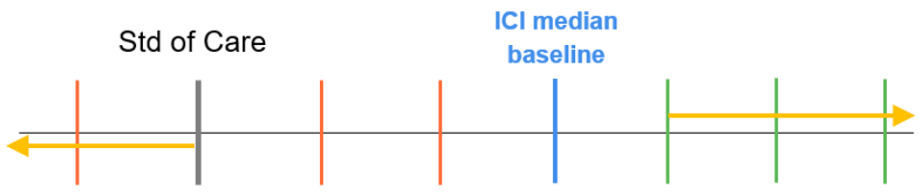
Some participants voiced that "everyone should have the option" for treatment even if an algorithm predicted no benefit or even negative average benefit because some fraction of patients still do respond and may elect to take the chance in the absence of other viable treatment options. Payers, especially, have a higher level concern that any policy must be applied equitably and consistently across eligible treatment populations. These concerns align with other multi-stakeholder groups who have identified equity and the "value of hope" as components of value that can be missed by conventional cost-effectiveness analyses. [8,9,10,11]

When the group considered negative (unfavorable) results, a key question would inevitably be raised by the observation that a given subpopulation was not benefiting from treatment: Is this finding driven by biology (genetics, age, physiology) or by access (social determinants of health, adherence)? The root cause would need to be assessed before the finding would be actionable—raising the question of whether the RWE outcomes tracking platform would become merely a way of driving future research. This was not felt to be as problematic with subpopulations appearing to benefit more from treatment, who could be encouraged and incentivized to seek out treatment. This type of effort fits in well with existing programs for access to care.

Figure 2: Sample Thresholds – Clinical Outcomes

Median Overal Survival (OS)

What result in a subpopulation, compared to the total ICIs population, would impact a:
- Coverage decision by payer?
- Treatment recommendation by a provider?
- Treatment decision by a patient?



Benefit to SOC	-2 mo.	+0 mo.	+2 mo.	+4 mo.	+6 mo.	+8 mo.	+10 mo.	+12 mo.
Benefit to ICI Baseline	-8 mo.	-6 mo.	-4 mo.	-2 mo.	+0 mo.	+2 mo.	+4 mo.	+6 mo.
% of benefit	*-133%*	*-100%*	*-66%*	*-33%*	*0%*	*+33%*	*+66%*	*+100%*

CONCLUSION: INSIGHTS FOR FUTURE TEAMS

The METRICS process as designed and demonstrated by the NEWDIGS LEAPS consortium for aiding the development of a machine learning predictive model of checkpoint inhibitors for advanced/metastatic NSCLC suggests the following learnings for future teams:

- That a multi-stakeholder, fit-for-purpose metric identification process can be executed in only a few months.
- That both clinically specific and system impact metrics are feasible and important to address the decision-making needs of payers as well as patients, providers, and the development team itself.
- That establishing up-front thresholds for change in the chosen metrics can establish targets of meaningfulness to guide the predictive model development process and help ensure its future clinical adoption and reimbursement.

As the healthcare system evolves towards value-based care, predictive algorithms can play a critical role, but their findings must be perceived by those outside the data science community as meaningful, substantial, and actionable. The METRICS process can be an important tool to ensure that alignment and so enable the practical success of machine learning predictive algorithms to improve patient outcomes and value from therapeutic regimens.

ACKNOWLEDGEMENTS

We want to express our appreciation to all the participants who helped create and refine this METRICS Framework including the 100+ Design Lab participants at the June 2022 and November 2022 LEAPS Design Labs as well as the leaders and members of the LEAPS METRICS team and subteams. Without their expertise and assistance throughout all aspects of the development of the generalizable METRICS process and its application to the advanced/metastatic NSCLC case study, we would not have been able to achieve the goals and objectives we set forth.

REFERENCES

1 Core Outcome Measures in Effectiveness Trials (COMET) Initiative database, available at: https://comet-initiative.org/Studies. When queried on February 7, 2023, the database returned over 1000 core outcomes set studies, with 139 in cancer.

2 Iorio A, Skinner MW, Clearfield E, et al. Core outcome set for gene therapy in haemophilia: Results of the coreHEM multistakeholder project. Haemophilia 2018; 24: 167-72.

3 Kirkham JJ, Davis K, Altman DG, Blazeby JM, Clarke M, Tunis S, et al. Core Outcome Set-STAndards for Development: The COS-STAD recommendations. PLoS Med. 2017; 14(11):e1002447.

4 Gargon E, Williamson PR, Blazeby JM, Kirkham JJ. Improvement was needed in the standards of development for cancer core outcome sets. J Clin Epidemiol 2019; 112: 36-44.

5 Nair R, Aggarwal R, Khanna D. Methods of formal consensus in classification/diagnostic criteria and guideline development. Semin Arthritis Rheum 2011; 41: 95-105.

6 Stanford Encyclopedia of Philosophy: Voting Methods. Revised June 2019. Available at: https://plato.stanford.edu/entries/voting-methods. Accessed February 7, 2023.

7 Zimmerman S, Sloane PD, Wretman CJ, et al. Recommendations for medical and mental health care in assisted living based on an expert Delphi consensus panel: a consensus statement. JAMA Network Open 2022; 5:e2233872.

8 Lakdawalla DN, Doshi JA, Garrison LP, Phelps CE, Basu A, Danzon PM. Defining elements of value in healthcare – a health economics approach: an ISPOR Special Task Force Report. Value Health 2018; 21: 131-9.

9 Drummond M, Torbica A, Tarricone R. Should health technology assessment be more patient centric? If so, how? Eur J Health Econ 2020; 21: 1117-20.

10 Peasgood T, Mukuria C, Rowen D, Tsuchiya A, Wailoo A. Should We Consider Including a Value for "Hope" as an Additional Benefit Within Health Technology Assessment? Value Health 2022; 25:1619-23.

11 Frank LB, Concannon TW. Inclusion In Health Technology Assessments: The First Step Toward Equity. Health Affairs Forefront, November 2021. Available at: https://www.healthaffairs.org/do/10.1377/forefront.20211104.341669. Accessed March 13, 2023.

APPENDIX A

METRICS CONSIDERED FOR ADVANCED/METASTATIC NSCLC USE CASE AND PRIORITIZATION

Clinical Outcome Measures – Initial Candidates Presented by Subteam

- Median Overall Survival (OS)
- Time to Treatment Failure
- Median Progression Free Survival (PFS)
- Duration of treatment*
- Proportion of participants surviving at a timepoint (e.g., 12 months) after the initiation of treatment*
- Objective response rate (ORR) according to RECIST1*
- Number of Participants with Dose Limiting Toxicities*
- Quality of Life Core Questionnaire*

Impact Metrics – Initial Candidates Presented by Subteam

- Time to treatment / "effective treatment"
- Total cost of care
- Time on ineffective treatment (less = better)*
- Expenditure on ineffective treatment (lower = better)*
- Administrative burden / steps to treatment (fewer = better)*
- "Intact research system" that rewards true breakthroughs*
- Screening*
- Palliative Care*

The METRICS team members voted on the two short lists of measures above using the following questions:

1. Select the Clinical Outcomes/Impact measure that you think is **most important** for the NSCLC Case Study
2. Select the Clinical Outcomes/Impact measure that you think is **most feasible** to implement in the NSCLC Case Study

**Measures which received fewer than 2 votes for importance or feasibility by METRICS team members in anonymous voting process and removed from short lists for further discussion.*

Priority-ranked "Short Lists" of both Clinical Outcome Measures and Impact Metrics as determined by METRICS team session participants at the November 2022 Design Lab:

Clinical Outcome Measures	Priority
Median Overall Survival (OS) The length of time from either the date of diagnosis or the start of treatment for a disease, such as cancer, that half of the patients in a group of patients diagnosed with the disease are still alive	**1**
Time to Treatment Failure Time to treatment failure (TTF) is defined as a composite endpoint measuring time from treatment initiation to discontinuation of treatment for any reason, including disease progression, treatment toxicity, and death.	**2**
Median Progression Free Survival (PFS) The length of time during and after the treatment of a disease, such as cancer, that a patient lives with the disease but it does not get worse.	**3**

Impact Metrics	Priority
Time to treatment/"effective treatment" Time from diagnosis or symptom onset to effective first treatment.	**1**
Total cost of care Total cost of an episode of care for a patient	**2**