embold HEALTH™

Embold Health Curation Methodology

2024

By Robert Gambrel & Hannah Krason

Table of Contents

2. PROGRAM GOALS

Embold's goal is to steer members to high-quality, high-value providers. Doing so leads to better clinical outcomes, rarer use of low-value services, and lower total cost of care.

Our program can support a variety of user preferences on how to balance provider quality against cost when evaluating providers. Our default recommendation is to treat them equally for both "low performers" and "high performers" by using our overall composite score. This score gives 50% weight to Clinical Performance and 50% weight to Cost Performance. We have found that this approach optimizes the marginal gains possible on both axes of evaluation. However, some customers may choose to balance these two domains differently. It goes without saying that down-weighting either component will lead to lower expected improvements in that area. Embold acknowledges that a benefits program is optimization exercise that includes customer preferences, member access issues, expected outcomes improvements, and feasibility of implementation and integration within the existing benefits ecosystem. We are happy to collaborate and make recommendations for customer-specific goals throughout the implementation process.

3. EMBOLD'S EVALUATION METHODOLOGY

Embold's provider analytics are uniquely designed to answer two questions:

1. **How different is the provider than average?**

   a. Examples:

      i. A provider has a 10% higher surgical complication rate than their peers.

      ii. A provider has a 15% lower patient-level adherence to recommended medications than their peers.

      iii. On average, a provider is 7% higher risk of undesirable practices across all clinical measures.

      iv. A provider costs 8% lower per member per month than their peers.

2. **How confident are we that difference is real?**

    a. Examples (each correspond to an item above):

        i. We are 90% confident that the provider's complication rate is between 4% and 16% higher than average.

        ii. We are 90% confident that the provider's medication adherence is 10% to 20% lower than average.

        iii. We are 90% confident that a provider's risk profile is between 2% and 10% worse than their peers.

        iv. We are 90% confident that the provider costs between 6% and 10% less than average.

When making provider recommendations, we rely on both aspects to make curation decisions. As described below, we generally require that multiple conditions are met to designate a provider as high or low performing:

1. A provider's risk-adjusted performance is substantially different than average.

    a. E.g., if a provider is 1% better than market peers, this isn't different enough to highlight.

2. We are statistically confident that a provider is truly different than average.

    a. E.g., if a provider performs well on a limited number of cases, but there is not enough data to establish a clear pattern for future cases, treat them as average.

An example of how we summarize this richness is shown in Figure 1. This shows a hypothetical provider's model-adjusted surgical complication rate on the blue panel. Our most-likely estimate is that the provider's complication rate is 2.5%, but there is some uncertainty around that, as there is under any provider evaluation method.

Because measures have different ranges, we standardize the provider's expected score against their peer group average: in this case, the peer group has a complication rate of 1.5%. This yields the green distribution on the upper right. Compared to peers, this provider is (2.5 / 1.5) = 1.66 times riskier for complications. Again, we have a confidence band around that. The bottom right panel shows that, based on the distribution, we are 90% confident the provider's risk profile is 1.24 to 2.09 times riskier than the peer group.

This is the approach Embold uses for scoring. We compare each provider measure against a market target to derive measure-level risk ratios. Risk ratios are always scaled such that high values equate to higher risk of undesirable practice choices or outcomes.

We average those risk ratios across measures to derive the provider's overall risk profile. That overall risk profile has a distribution like the upper right panel that can be summarized by the "best guess" and confidence bands like the bottom right panel. A risk profile of 1 indicates average performance in the market. Scores less than 1 are desirable, as they indicate reduced probability of undesirable events. These scores can also be considered as percentages: A score of 1.1 means the provider is 10% higher risk than peers; a score of 0.95 means the provider has 5% less risk than peers.

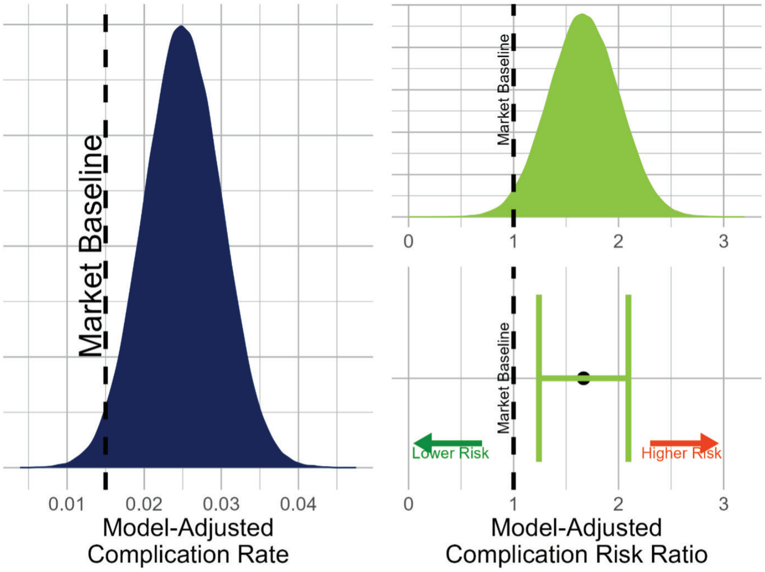


Figure 1. An estimate of provider score (blue) is compared against peers to determine provider risk (green). That profile can be summarized by a point estimate and confidence limits. For Risk scores (green plots), scores above 1 indicate higher probability of undesired practices.

3.1. DEFAULT CURATION BENCHMARKING

Embold creates scores on providers that are benchmarked nationally. This ensures that scores across state borders, and across the country, are comparable. A risk of this approach is that scores are no longer "graded on a local curve," so there may be pockets and regions of systemic over- or under-performance compared to national targets.

There are two ways to address this issue. First, the benefits administrator may evaluate and address these access issues through program design and ancillary approaches with Embold's support. The Provider Guide tool will still use provider scores when ordering providers in search results, so the best-available providers in the user's search radius will be returned first (see section 8). Implementations that use other navigation tools and access Embold scores via the API may consider a similar rule for ordering searches to promote the best providers within the user's search radius.

Second, Embold also supports the concept of "evaluate nationally; curate locally." If customers prefer, our scores can be updated and put back on a local curve. This would mean that providers that are "good for their local area" have high scores, whether or not those scores would be considered high-performing on a national scale. We recommend this rescaling only for implementations that include a single geographic region (for example, a single state), or include regions that are non-adjacent (for example, multiple large metros spaced out across the country). Members who might conceivably search within 2 differently-benchmarked areas (for example, those living on the border of 2 states that are locally benchmarked) may see similarly-performing providers have their scores inflated or deflated because they are compared against different peer groups.

All discussions of scores and risk profiles below are conducted based on Embold's statistically risk-adjusted performance scores for providers.

4. CONFIGURATION DECISIONS

This document describes a variety of options a customer can take when designing a tiered benefits program. They can be summarized as follows:

- How to define the peer group to determine "acceptable" performance?
    - As described below, Embold by default compares providers nationally. This makes interpretation easier when deploying in markets that cross state boundaries. However, customers may choose to grade "on a local curve" if they are geographically concentrated.
- How should Cost Performance be incorporated in provider ratings?
    - As described below, , Embold supports a variety of approaches to integrate Cost Performance in provider evaluation. Our default approach is to incorporate it when identifying both over- and under-performing providers. Customers have chosen to use it only for one or neither, though, when their goals for the program are less sensitive to provider cost.

- How many tiers should the benefits program have?
    - Embold recommends a 3-tier system of high-performing, low-performing, and average-performing providers. However, we can also support implementations that divide into two tiers. These are typically “high performers vs. everyone else” and “low performers vs. everyone else.”
    - The number of tiers used for financial incentives (copay differentials, network status) does not necessarily have to be the same as used in Provider Guide display. For example, there may be a higher copay associated with seeing low-performing providers but no financial differentiation between seeing average- and high-performing providers. This would not stop us from labeling high-performers as such in our tools and promoting them to the top of member searches.

All of the above are configurable; we recognize that new customers may need consultation from the Embold team to make decisions that align with program success. This consultation is part of our implementation process.

5. IDENTIFYING LOW PERFORMERS

Embold supports a number of options to support customer preference in identifying under-performers. These can be summarized as:

- **Curating on a blended Clinical Performance / Cost score**
- **Curating on Clinical Performance only**
- **Curating on individual domains (Appropriateness, Effectiveness, Cost)**

All methods follow the same general approach. Along the metric chosen, establish a of minimally-acceptable performance compared to the market baseline. If providers score worse than this baseline, and we have sufficient confidence that their score is worse than average, they are designated as low-performing.

Figure 2 shows two hypothetical providers’ risk profiles. The tails represent the 90% confidence band of our model-adjusted performance evaluation. The circles represent the “most likely” point estimate of the provider’s performance.

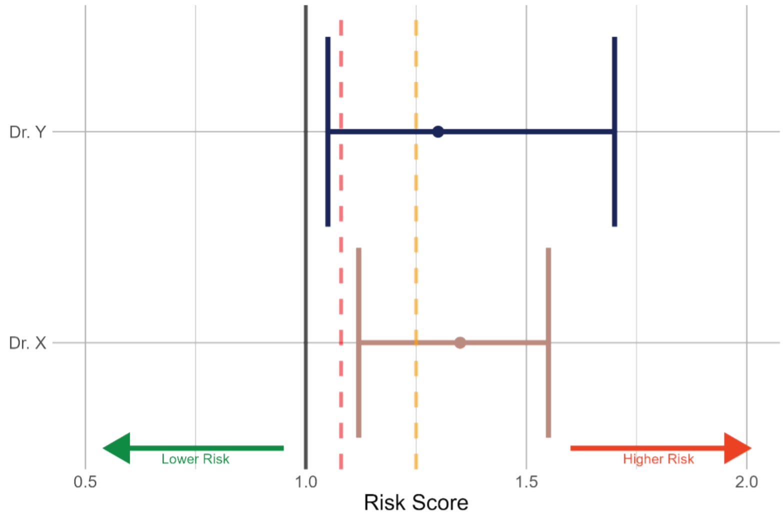


Figure 2. Risk score profiles for two hypothetical providers, showing their point estimate (dot) and 90% confidence band (error bars). A risk score greater than 1 indicates that the provider is worse than average with respect to their peer group, and vice versa. The gold and red dashed lines reflect the minimum performance threshold and performance buffer, respectively. To be considered low performing, a provider must have their point estimate exceed the minimum performance threshold, and their lower confidence limit exceed the buffer threshold. In this example, Dr. X would be low performing while Dr. Y would not.

The golden line indicates the risk cutoff in this scenario. Any provider that exceeds this risk threshold is considered outside the range of acceptable performance. Both providers in the figure have model-adjusted scores that exceed the risk threshold.

The dashed red line indicates a "performance buffer" that is worse than average but still in a "grey zone" (note that this is distinct from the solid black line at 1.0, which indicates the average performance of all providers). If there is still a chance that the provider performs within this band, they will be deemed acceptable, even though their point estimate fails the first test. In this case, our model does not have 90% confidence that Dr. Y performs outside of that allowable performance band. However, our models are 90% confident that Dr. X's performance falls outside of it.

This band is enforces a strong hurdle for a provider to be identified as under-performing. For providers that are on the border, Embold biases towards being conservative and erring on the side of defaulting providers to an "acceptable" status. When providers fall outside this band, we have a high degree of confidence that their score is not due to regular annual fluctuation or statistical noise.

In this example, Dr. X would be identified as low-performing for benefit design purposes, while Dr. Y would be allowed in as "acceptable performance", because failing both tests is required. However, as described below in Section 5.4, Dr. Y would be ineligible for promotion as a "high performing / high value" provider due to the borderline scores seen here.

Below, we describe several options to identify under-performers. These approaches incorporate Cost Performance to different degrees. During implementation, we suggest that customers choose only one of the options described below. If there are multiple pathways for a provider to be deemed an under-performer this may lead to confusion for providers on how to improve their clinical practice. All three approaches follow the pattern in Figure 2 and require both a poor score and high confidence in that score.

5.1. EVALUATING ON OVERALL (COST + CLINICAL) PERFORMANCE

Embold produces a 50/50 weighted composite of Clinical and Cost scores. This overall score is based on the provider's Clinical and Cost performance holistically, weighting all clinical measures equally against each other and weighting Cost performance and Clinical performance equally. In general, the 50/50 score approximates a bell-shaped distribution and gives excellent program-level results on both cost savings and clinical improvements.

Our approach to curating out providers on this scale is to identify those who score at least 12% worse than average. On our 0-100 score scale, this translates to a score of ≤ 35. This indicates a higher propensity to have adverse outcomes, render unnecessary care, and/or cost more. Looking at Figure 2, this would mean drawing the golden dashed line at 1.12. In addition, we require 90% confidence that the provider is at least 8% worse than average as well, which equates to having the red dashed line at 1.08.

5.2. EVALUATING ON CLINICAL PERFORMANCE ONLY

Embold also provides a 100% Clinical Performance overall score. This score excludes Cost from our evaluation and tends to be bell-shaped. Implementations that focus on Clinical Performance only can expect greater quality improvements at the tradeoff of reduced cost savings. Curation works similarly to the above – our recommendation is to determine those that are at least 12% worse than average with a high degree of confidence that the provider is at least 8% worse than average.

5.3. EVALUATING ON DOMAINS SEPARATELY

Some customers have chosen to break down clinical measure concepts into Appropriateness and Effectiveness domains separately instead of looking at them in totality, alongside the Cost domain. This may occur when specific domains align with existing clinical reporting capabilities the customer has that make program monitoring easier (for example, off-the-shelf HEDIS measures).

In such a case, Embold supports sequential domain-specific curation that works similarly to above: if a provider fails the “low score + high confidence” tests on any domain, they are designated as an under-performer.

Note that this approach does have tradeoffs. First, Embold requires 4 clinical measures to receive an overall score or clinical-only score (all of which are used equally in the blended approaches above), with at least one coming from each clinical sub-domain. For domain-specific curation on Appropriateness and Effectiveness, we add an additional requirement that a provider have 2+ measures in each domain we test. The only exceptions to this are two of our newly released specialties: Lung Cancer Surgery and Bariatric Surgery. The nature of those specialties is such that the measure panel is heavily biased towards Effectiveness measures, so those specialists are curated “out” based only on their Effectiveness scores.

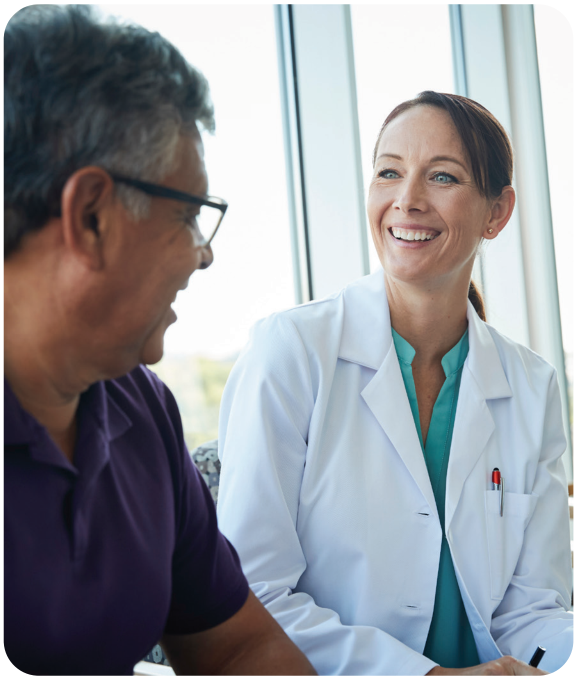
This domain-specific measure requirement results in fewer providers being evaluated overall, as they eliminate doctors who have 3 in one domain but 1 in another. This is around a 7% reduction in the number of providers that have scores.

Second, for reporting purposes and member usability, Embold aims to report scoring in as simple a metric as possible – usually in a single number “overall score”. Layering on domain-specific curation will lead to instances where a provider has a seemingly acceptable overall score, but underperformance in a specific domain causes them to be deemed low-performing. Peers with similar overall scores will have different curation statuses depending on how they perform on specific sets of measures. Because of this potential confusion, Embold recommends not displaying numeric scores in any member-facing tools when curating on domain scores and, instead, including clear badging in the user experience.

Finally, because domains are constructed of fewer measures than overall scores, the distribution of performance is less bell-shaped and varies by specialty. Because of this, Embold has developed Specialty-Domain specific cutpoints of minimally acceptable performance based on expected clinical variation by specialty.

Users may curate on any combination Appropriateness, Effectiveness, and Cost. The approach for setting domain-specific performance cutpoints is described below.

5.3.1. DEFINING ACCEPTABLE DOMAIN PERFORMANCE



Upon clinical review and feedback from our Scientific Advisory Board, we have set the maximum allowable Effectiveness, Appropriateness, and Cost risk profile by specialty. Given the mix of measure concepts within specialties, Embold recognizes that expected degrees of performance variation will differ by specialty, and as a result, the range of what is not meaningfully different from "average" varies. For some specialty-domains, providers may be curated out if their risk profile is 10% worse than target rates. For others, variation is such that a provider might need to be 25% worse to be considered outside the realm of standard acceptable clinical practice variation. Embold has set thresholds based on our most recent production runs and monitors any required adjustment to these cutpoints when releasing new data.

We maintain the additional requirement that we are also 90% confident that the provider is at least 8% riskier than target rates on any curated domain.

5.4 NOTE ON "BORDERLINE" PROVIDERS

As noted above, a provider must fail two checks to be identified as an under-performer: the metric of interest must demonstrate an unacceptably high risk, and we must have high confidence that that score is indicative of worse-than-average performance.

There are frequently situations where a provider fails only one of the two checks – e.g., they have a poor score on the Clinical composite but not enough data to pass the confidence check. If an implementation were to use a different metric to identify high-performers (for example, using only Clinical scores for low-performance but the Overall (Cost-inclusive) score for high-performance), such a provider might leap-frog to be high-performing by virtue of focusing on a different set and weighting of performance measures. To avoid this scenario of "Provider X was almost curated out based on the low-performing metric, but when we look at the high-performing metric they are in fact a top-tier provider," Embold has implemented a rule to limit those "borderline underperformers" to the "average" group at best.

In other words:

- *If a provider fails both the score check and confidence check for low performance: low performing.*
- *If a provider fails one of the checks: average performing.*
- *If a provider passes both checks: use the "high performing" logic below to determine whether they are high or average performing.*

In Figure 2, Dr. Y is an example of this. They are not deemed low-performing by virtue of having not quite enough confidence in their current score. However, because they failed one of the checks, they are restricted to being part of the "average" cohort of providers, even if the metric used to identify high-performers treats them quite favorably.

6. IDENTIFYING HIGH PERFORMERS

Among the providers remaining after identifying low-performers and those that are average-at-best, we have developed an approach that promotes high performers. These providers will be returned at the top of searches and can be identified with a label or badge. Customers may also deploy a financial incentive to encourage members to choose these providers, though that is implementation-specific.

To reiterate – if a provider has been identified as a low performer, their status is set. There is no way to overwrite that decision and move them into the high-performing group based on performance on other factors until such time as that provider's sufficiently improves on the metric used to identify low performers.

Customers may choose from any (or any combination) of the below to identify high performers. We encourage customers to review the size of the resulting group, though, as part of the implementation process. Financially incentivizing too broad a pool of providers may cause unexpected program cost. As above, we work closely with customers to ensure that the resultant curation meets program goals.

Note that all of these approaches follow a similar pattern as before: If a provider has a strong performance score, and we have confidence in that score, they are included in the "high performing" group. Our degree of confidence required is lower, though. Figure 3 shows two hypothetical providers' overall performance risk profiles. The tails represent the 75% confidence band of our model-adjusted performance evaluation. The circles represent the "most likely" point estimate of the provider's performance.

The green line indicates the risk cutoff of a 12% improvement over target rates. In both cases, the providers' model-adjusted scores are better (lower) than the risk threshold. High performance also requires 75% confidence that the provider's score is lower than average (i.e., risk score of 1.0).

In this example, Dr. X would be flagged as high-performing, as both tests are met. Their model-adjusted risk profile is low enough, and there is sufficient confidence that their score is truly better than average. Dr. Y, while having good performance, does not have sufficient confidence to earn the "high performing" designation.

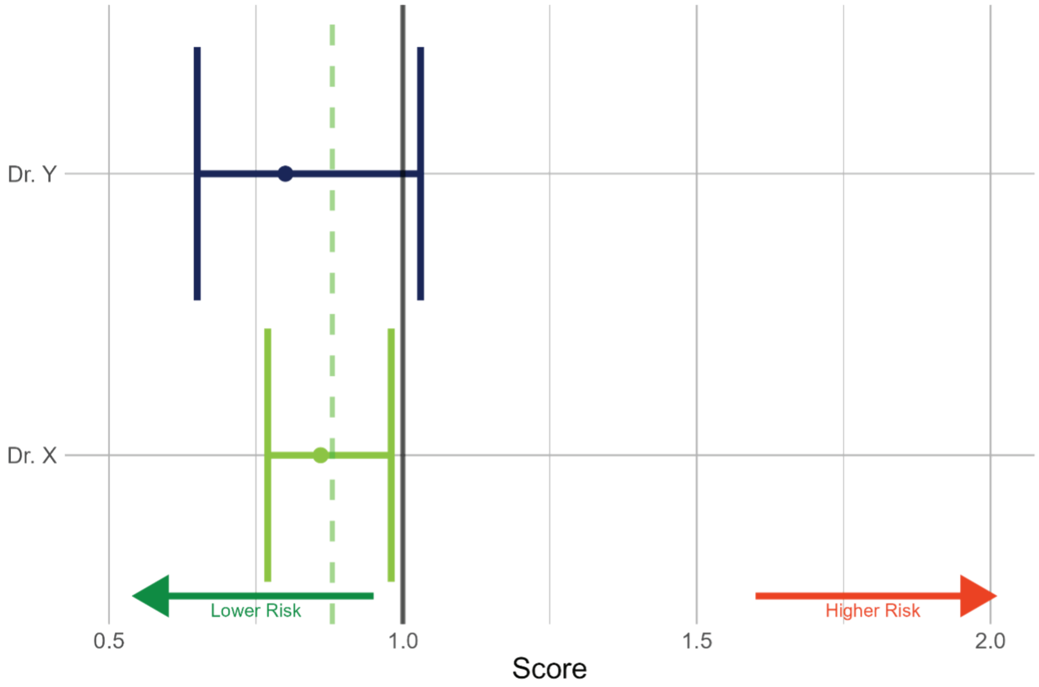


Figure 3. Risk score profiles for two hypothetical providers, showing their point estimate (dot) and 75% confidence band (error bars). A risk score less than 1 indicates that the provider is better than average with respect to their peer group, and vice versa. The green dashed lines reflect the maximum performance threshold. To be considered high performing, a provider must have their point estimate below the maximum performance threshold, and their confidence band must lie entirely below the average value of 1.0. In this example, Dr. X would be high performing while Dr. Y would not.

6.1 HIGH PERFORMACE GROUPS

Group 1: strong overall (Cost + Clinical) performance

If a provider's overall performance profile is 12% better than market targets (e.g. risk propensity compared to target is 0.88 or lower) they pass the first hurdle. On our 0-100 scale, this translates to a score of 65+. We again layer on a requirement of confidence in that score, requiring that we are also at least 75% confident that the provider is truly better than average.

Note the differences from our approach to identifying low performers. In this case, the required confidence threshold is 75%, and the confidence zone must exclude 1.0 instead of a "buffer target." This asymmetry in required confidence as compared to low performers was developed with feedback from our Scientific Advisory Board; it reflects the fact that it is more acceptable to steer towards a provider who has shown promising (but perhaps not definitive) results than it is to steer away from a provider without very strong evidence. In other words, for providers at the margin between low and acceptable, we are conservative and biased towards calling them acceptable; for providers that are at the margin between acceptable and high, we are more willing to label them as high.

Group 2: very strong clinical performance

Group 1 consists of a variety of provider phenotypes: high-Clinical, moderate-Cost; average-Clinical, low-Cost; and high-Clinical, low-Cost. However, we also know that there are some high-cost providers whose clinical performance justifies the expense for cost-minded customers. As such, a provider can also be flagged as "high performance" if their Clinical risk profile alone is 20% better than market (with a 75% confidence in being truly better than average), regardless of their cost performance.

We do not recommend this group be the sole recipients of the high-status designation. Instead, they can be considered as an addition to the strong overall cohort in Group 1.

Group 3: strong clinical performance

For customers who are cost-agnostic in their high-performance designation, Embold suggests an alternative approach similar to Group 1: a Clinical Performance score that is 12% better than market targets, with a 75% confidence behind that score being better than average.

Not recommended: Using domains to identify high performers

Embold recommends against using performance on specific domains (Appropriateness, Effectiveness, Cost) to identify high-performers. The reason for this that there are very few “unicorn” providers that perform better than average on everything. Most are a mix of high, average, and low performing aspects of care. By looking at their full profile, we can make an evaluation on whether they are generally more good than bad

This leads to a larger pool of doctors identified as high-performing with minimal impact on expected program impact on member outcomes.

7. FUTURE REFRESHES

Allowing all providers to become acceptable

Embold’s recommended curation methodologies are not percentile-based, and there is no fixed “minimum number of providers curated out” or “minimum number of highly-rated providers” baked into our process. They generally align with target curation sizes we have seen lead to successful programs in the past:

Percentage ratings:

- *10-20% of providers steered away from as low performing.*
- *20-35% of providers promoted as high performing.*
- *40-60% of providers designated as average.*

Our method makes it possible for all low-performing providers to improve over time and be included as “acceptable” in our program. These providers will have to shift their performance patterns towards acceptable thresholds; they must either get their risk profiles below our targets or do well enough that we lose confidence they are truly under-performing. The converse is also true – to the extent that providers respond to other perverse incentives, the number of providers that fail the clinical tests may increase year-to-year. Our approach is not designed to exclude a fixed number or proportion of providers from the acceptable pool. As we continue to refine and add more measures of clinical practice, we may need re-evaluate the specialty-specific performance targets in the future that reflect these new inputs.

7.1. SMOOTHING REFRESHES AFTER YEAR 1

After an initial deployment, we are mindful that changing provider designations can have disruption on patient relationships, especially when financial incentives are in place. There is a balance between sharing scores that reflect meaningful changes in provider behavior and avoiding disruptions that have their own consequences. To address this, Embold has worked with several customers to allow for implementation-specific approaches to smooth updates from year to year.

These include:

- *For providers that were "high performing" in year 1, make it harder to be deemed an "underperformer" in year 2 by increasing the required risk profile or confidence to receive such a designation during refreshes. In other words, further raise the burden of proof to be designated low-status when prior years led us to the opposite conclusion.*
- *For providers that were "high performing" in year 1, make it easier to remain a high-performer in year 2, by requiring less confidence in high-performance in year 2.*
    - *If total number of providers deemed "high performing" is a concern, also limit the number of "average to high" switchers by raising the bar for those providers in year 2.*
- *For providers that were "low performing" in year 1, require more evidence than the default to designate them "high performing" in year 2.*
    - *We may also set the limit so high that providers are only allowed to move one performance tier (either up or down) during a data refresh.*

Embold is aware that provider performance is a continuum, and even our best efforts at statistical modeling leave room for a degree of unavoidable uncertainty for every provider. Adding to the fact that providers are humans capable of change and treat patient cohorts that change year to year, there is some inherent noise when determining a status for any given provider. Drawing lines to break up the population into 3 distinct groups will inevitably leave room for misclassification at the boundaries. The approaches above have been developed with years of experience and extensive feedback from our customer base to minimize these sorts of errors, but they cannot be reduced to zero.

8. SEARCH ORDERING IN PROVIDER GUIDE

When members conduct searches in Provider Guide that return a list of providers (e.g. searching for specialties, sub-specialties, or conditions), the search order will return high performers first, acceptable performers second, and low performers last. These tiers are determined by the rules outlined above. Within those three tiers, providers will be ordered by the score that is most relevant to the search: Embold Overall (Cost-inclusive or Clinical-only, depending on implementation) Composite score (if searching for the overall specialty or provider by name) or by the sub-specialty score (if searching for scored sub-specialists). This ensures that, even in regions where there are no high-performers available, the user is seeing the best of the "acceptable" providers at the top of their search results.

embold HEALTH™

Embold Health is a doctor-led healthcare analytics company that helps employers identify and guide their members to high-performing doctors, which improves patient outcomes and lowers costs.

Ratings are based on objective clinical performance data of the individual physician against regional peers and latest medical care standards.

Embold's vision is to raise the quality of health care by providing every healthcare consumer in America with actionable, objective doctor quality metrics, empowering them to make smarter health care decisions.

TO LEARN MORE, VISIT EMBOLDHEALTH.COM