★ OpsStar

WHITE PAPER · VERSION 1.0

# Forecasting Federal Deficiency Citations in U.S. Nursing Homes: Validation of the Stockdale Survey Risk Index™ (SSRI)

Dan Stockdale, MA, LNHA

OpsStar LLC

November 29, 2026 · Data: CMS Provider Data Catalog, release of October 1, 2026 · Correspondence: support@opsstar.com

## ABSTRACT

Every certified U.S. nursing home receives an unannounced standard health inspection about once a year, and the federal requirements (F-tags) that inspectors cite drive star ratings, penalties and public reputation. Facilities have limited time to prepare, yet few public tools tell them which requirements they are most likely to be cited for. This paper reports the out-of-sample accuracy of the Stockdale Survey Risk Index (SSRI), a proprietary statistical model built only from public Centers for Medicare & Medicaid Services (CMS) data that ranks 126 health F-tags by their chance of being cited on a facility's next standard survey. In a holdout test on 2,931 facilities never seen in training, the SSRI achieved an area under the ROC curve (AUC) of 0.849, and its top 10 tags contained 41.3% of all tags cited. In a national five-fold cross-fitted backtest covering 14,647 facilities and 105,302 citations, each facility's top 10 contained 40.6% of the tags cited, 5.1 times the 7.9% expected by chance and about half of the 81.8% that any 10-tag list could reach. Predicted chances were well calibrated: tags given about a 25% chance were cited 24.8% of the time. The results suggest that public data can meaningfully focus survey preparation. Forecasts are now frozen and scored prospectively each month.

**Keywords:** nursing homes; long-term care; CMS health inspections; F-tags; deficiency citations; survey readiness; predictive modeling; regulatory compliance; quality of care; calibration

## Executive summary

**The problem.** Standard surveys are unannounced and can cite any of more than a hundred federal requirements. Administrators and directors of nursing have to decide where to spend scarce preparation time, usually from instinct, the last survey, or a generic checklist.

**What the SSRI does.** For every certified nursing home with survey history, the SSRI ranks the F-tags surveyors are most likely to cite on the next standard survey and gives each a chance. Every forecast is dated and frozen so it can be scored honestly when CMS posts the result.

**40.6%**
of cited tags were in the SSRI top 10 (42,771 of 105,302)

**5.1×**
better than picking 10 tags at random (7.9%)

**0.849**
ranking accuracy (AUC) on 2,931 facilities the model never saw

**93%**
of cited facilities had at least one tag correctly called

**How we know.** Every number in this paper is out of sample: each facility was scored by a version of the model trained without it. The national backtest covers 14,647 facilities, essentially every certified nursing home with enough survey history.

**Can you trust the percentages?** Yes, within the range where nearly all forecasts fall. When the SSRI says 25%, the tag was cited about 25% of the time; at 45%, about 45%.

**What it cannot do.** It cannot see complaints, facility-reported incidents or what surveyors observe on site, so some citations will always be a surprise. It is a planning tool, not a guarantee and not an official CMS report.

**How to use it.** Start mock surveys, audits and in-services with your top-ranked tags. OpsStar's free Staff Survey Prep Plan pairs each one with a concrete fix, and facility and chain licenses are available for survey-readiness training.

# 1 Introduction

Nursing homes that take Medicare or Medicaid are inspected by state survey agencies on behalf of CMS. A standard health survey occurs at least every 15 months, with a statewide average of about 12, and is unannounced. Inspectors evaluate compliance with the federal requirements for long-term care facilities, each identified by an F-tag number in Appendix PP of the CMS State Operations Manual [1]. The results feed the health inspection domain of the CMS Five-Star Quality Rating System [2], can trigger civil money penalties, and are published on Medicare Care Compare.

Because the survey can cite any requirement, facilities face an allocation problem: which areas deserve attention first? Published deficiency data show strong regularities. A small number of tags, such as infection prevention and control (F880) and food safety (F812), are cited at a large share of facilities, and citation habits differ by state and region. Facility history, staffing and quality data are also public. The SSRI was built to turn these signals into a facility-specific, ranked and calibrated forecast.

This paper describes the data, the general modeling approach and, in full, the validation design, and reports the results. The model's specific inputs, transformations, weights and code are proprietary to OpsStar LLC and are not disclosed.

# 2 Data

All inputs come from the CMS Provider Data Catalog, nursing home datasets, release of October 1, 2026 [3]:

- **Health deficiencies** [4]: tag-level citations from standard health surveys, with survey dates, over the three most recent standard survey cycles.
- **Provider information** [5]: location, ownership, certification, Special Focus Facility status, survey cycle dates, and payroll-based staffing and turnover measures.
- **Quality measures** [6]: MDS-based quality measures for short- and long-stay residents.

The analysis covers 126 health F-tags, those cited at 0.5% or more of facilities nationally. Complaint and facility-reported incident surveys, life safety (fire) citations, and any resident- or staff-level information are excluded. No protected health information is used.

# 3 Methods

The SSRI is a supervised probabilistic classifier. Each training example is a facility and tag pair: the information available from CMS data before a standard survey, and whether that tag was cited on the survey. Inputs fall into four families:

**Survey history**
The facility's own citations on earlier standard surveys.

**Local survey patterns**
How often inspectors in the facility's area cite each tag.

**Workforce**
Staffing levels and staff turnover.

**Quality and status**
Quality measures, citation load and enforcement status.

The model outputs a probability for each of the 126 tags. Tags are ranked by probability, and the top 10 form the facility's forecast. The training set contains 11,716 facilities and 1,476,216 facility and tag examples. The model is rebuilt monthly as CMS releases new data.

## 4 Validation design

Three tests are used. Tests 1 and 2 are reported here; test 3 is ongoing.

**1. Holdout test**

Facilities are split 80/20 by a fixed hash of their CMS certification number. The model is trained on 11,716 facilities and evaluated on the other 2,931, which it never sees.

**2. National backtest**

Each facility's most recent standard survey is predicted from the two before it, by a model trained on the other four fifths of facilities (5-fold cross-fitting). No facility is scored by a model that saw it.

**3. Prospective scoring**

Forecasts are dated and frozen. When CMS posts a newer standard survey, the forecast is scored and published on that facility's OpsStar Snapshot as “SSRI Accuracy.”

**Metrics.** Discrimination is measured by the area under the ROC curve (AUC) [7] over all facility and tag pairs. Probability accuracy is measured by the Brier score [8] and log loss, and calibration by comparing predicted and observed citation rates within probability bands. Practical usefulness is measured by the share of all cited tags that appear in the facility's top 10. Two reference points frame this share: a random choice of 10 of the 126 tags (7.9%), and the ceiling for any 10-tag list, since facilities cited for more than 10 tags cannot have all of them caught (81.8%).

## 5 Results

### 5.1 Holdout test

**Table 1. Holdout performance on 2,931 unseen facilities**

| Measure | Result | Interpretation |
|---|---|---|
| AUC | **0.849** | Chance a cited tag is ranked above an uncited one (0.5 = coin flip, 1.0 = perfect). |
| Top-10 tags cited | **2.91** | Average number of the top 10 that were cited, out of 7.1 citations per facility. |
| Share of cited tags in top 10 | **41.3%** | Share of all cited tags that were in the facility's top 10. |
| Brier score | 0.04369 | Mean squared error of predicted probabilities (lower is better). |
| Log loss | 0.1622 | Penalizes confident errors (lower is better). |

Across the five cross-fitting folds used in the backtest, holdout AUC ranged from 0.849 to 0.855, so the result does not depend on one favorable split.

## 5.2 National backtest

Across 14,647 facilities, inspectors cited 105,302 health tags on the most recent standard survey. 42,771 of them (40.6%) were in that facility's SSRI top 10. This is 5.1 times the random reference of 7.9% and about 50% of the 81.8% ceiling (Figure 1). 93% of the 13,754 facilities with at least one citation had at least one tag correctly called.

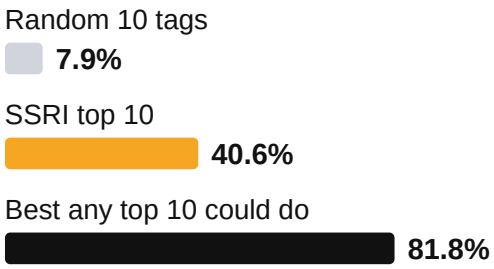


Figure 1. Share of the 105,302 cited tags that appeared in a facility's top 10, national backtest, against random and ceiling references.

**Table 2. Backtest results by number of tags cited on the scored survey**

| Tags cited | Facilities | Avg. tags correctly called | Share caught |
|---|---|---|---|
| 1–3 | 3,510 | 1.1 | 55% |
| 4–6 | 3,597 | 2.4 | 48% |
| 7–10 | 3,352 | 3.7 | 44% |
| 11–15 | 2,068 | 4.9 | 38% |
| 16+ | 1,227 | 6.3 | 30% |

Facilities cited for more tags had more tags called correctly, while the share caught fell because a 10-tag list covers less of a long citation list. 893 facilities had no health citations on the scored survey and are excluded from Table 2.

### 5.3 Calibration

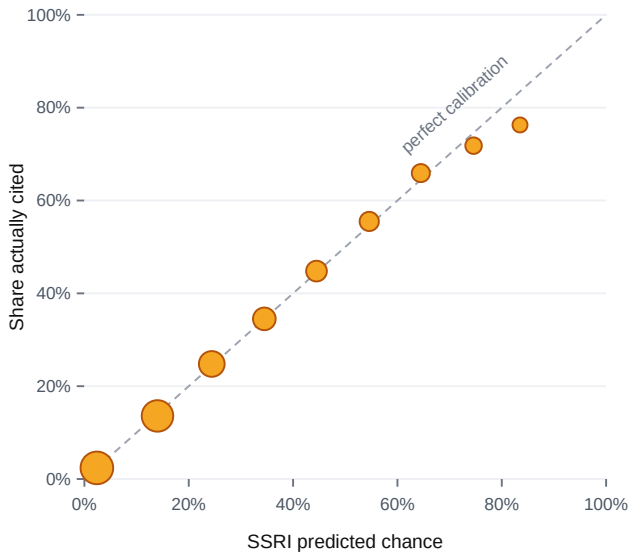


Figure 2. Predicted probability vs. observed citation rate, holdout test. Dot area reflects the number of predictions in each band.

**Table 3. Calibration by band**

| Predicted | Observed | Predictions |
|---|---|---|
| 2% | 2.4% | 311,227 |
| 14% | 13.6% | 33,263 |
| 24% | 24.8% | 11,780 |
| 34% | 34.5% | 6,014 |
| 44% | 44.8% | 3,422 |
| 55% | 55.5% | 1,863 |
| 64% | 65.9% | 1,046 |
| 75% | 71.8% | 493 |
| 84% | 76.3% | 177 |

Predicted and observed rates agree closely in every band below 70%, which contains over 99.8% of predictions. In the two highest bands the model is somewhat over-confident.

### 5.4 Illustration

Three anonymized backtests show typical results: a facility in Alabama (September 2025 survey) was cited for 6 tags, 2 of them in its top 10; one in Massachusetts (December 2025) for 5, with 2 caught; and one in Washington, D.C. (April 2023) for 15, with 7 caught. For the Alabama facility the current forecast ranks F812 food safety (66%), F880 infection prevention and control (48%) and F689 accidents and supervision (34%) highest.

## 6 Limitations

- **Public data only.** Surveys also respond to complaints, facility-reported incidents and on-site observation, which the SSRI cannot see.
- **Retrospective evaluation.** Backtests use historical surveys. Inspection practice, CMS guidance and the F-tag list change. Prospective scoring (test 3) is the definitive test.
- **Survey timing.** Scored surveys range from 2019 to 2026, with 90% in 2025 and 2026.
- **Undisclosed method.** Because the model is proprietary, independent replication of the method is not possible; the validation design is disclosed so results can be judged, and prospective results are published per facility.
- **Probabilities, not verdicts.** A high chance does not mean a citation will occur, and a low chance does not indicate compliance. The SSRI is not the CMS CASPER or iQIES report and says nothing about individual staff.

# 7 Conclusion

Using only public CMS data, the SSRI ranks the F-tags a nursing home is most likely to be cited for with strong discrimination (AUC 0.849) and close calibration, and its 10-tag forecast captures 40.6% of real citations nationally, about half of what any 10-tag list could. For administrators, this means preparation time can be pointed at the areas surveyors are most likely to examine and cite. OpsStar will report prospective results in future versions of this paper.

# References


1. Centers for Medicare & Medicaid Services. *State Operations Manual, Appendix PP: Guidance to Surveyors for Long Term Care Facilities.* Baltimore, MD: CMS.
2. Centers for Medicare & Medicaid Services. *Design for Care Compare Nursing Home Five-Star Quality Rating System: Technical Users' Guide.* Baltimore, MD: CMS.
3. Centers for Medicare & Medicaid Services. Provider Data Catalog: Nursing homes including rehab services. https://data.cms.gov/provider-data/topics/nursing-homes (release of October 1, 2026).
4. Centers for Medicare & Medicaid Services. Health Deficiencies dataset. Provider Data Catalog (release of October 1, 2026).
5. Centers for Medicare & Medicaid Services. Provider Information dataset. Provider Data Catalog (release of October 1, 2026).
6. Centers for Medicare & Medicaid Services. MDS Quality Measures dataset. Provider Data Catalog (release of October 1, 2026).
7. Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*(1), 29–36.
8. Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78*(1), 1–3.



**Notice and disclosure.** The author is affiliated with OpsStar LLC, which develops and licenses the SSRI. “Stockdale Survey Risk Index” and “SSRI” are trademarks of OpsStar LLC. The SSRI methodology, model and software are proprietary trade secrets of OpsStar LLC. This paper may be shared and quoted with attribution; use of SSRI outputs or the name in a product or service requires written permission (opsstar.com/terms). OpsStar is not affiliated with CMS or Medicare. This paper is not legal, medical or regulatory advice.



**Suggested citation.** Stockdale, D. (2026). *Forecasting Federal Deficiency Citations in U.S. Nursing Homes: Validation of the Stockdale Survey Risk Index (SSRI)*. White paper, version 1.0. OpsStar LLC. https://opsstar.com/ssri