# Evaluation of the Primary Care First Model

# Study Protocol and Statistical Analysis Plan

Draft (September 24, 2024)

# Background and Objectives of the Primary Care First Evaluation

## Background

Primary care is critical to improving health outcomes and reducing overall health care costs. To improve the way primary care providers deliver care and to reimburse them for value over volume, the Centers for Medicare & Medicaid Services (CMS) have increasingly relied on alternative payment models (APMs). As part of this effort, the Center for Medicare & Medicaid Innovation (Innovation Center) within CMS launched the Primary Care First (PCF) Model in 2021. The PCF Model builds on principles and experiences from past Innovation Center initiatives, including the Comprehensive Primary Care Initiative, Multi-Payer Advanced Primary Care Practice demonstration, and Comprehensive Primary Care Plus (CPC+). The PCF Model tests whether financial risk taken on by primary care practices and performance-based payments for outcomes can (1) reduce total Medicare fee-for-service (FFS) expenditures; (2) reduce use of health care services, such as acute hospitalizations; and (3) preserve or enhance quality of care.

CMS enrolled practices in the PCF Model in one of two cohorts. Cohort 1 practices participate from January 1, 2021, through December 31, 2025. Cohort 2 practices participate from January 1, 2022, through December 31, 2026. Cohort 2 includes many practices that participated in CPC+. CMS defines a primary care practice under the PCF Model as one or more primary care providers (physician, nurse practitioner, physician's assistant, or clinical nurse specialist) working within the same physical office location or practice site.

The PCF Model provides a combination of capitation, volume-based, and performance-based payments to participating practices. Specifically, practices receive (1) a per-beneficiary-per-month prospective payment that depends on the average health of their attributed Medicare beneficiaries; (2) a Flat Visit Fee for primary care visits, subject to a geographic adjustment factor, and (3) a Performance-based Adjustment (PBA). The PBAs depend on practices' performance on several quality measures in addition to their performance in reducing beneficiaries' use of inpatient care or total cost of care, relative to national and regional benchmarks. Practices must meet a limited set of care delivery requirements and can use the PCF Model's flexible use of payments to invest in strategies that best suit their practices' unique patient population and resources. In return, practices take on limited financial risk in exchange for performance-based payments that reward participants that meet certain performance and quality benchmarks for selected outcomes.

This evaluation provides an opportunity to understand the effectiveness of the PCF Model across different practice organizations and beneficiary populations. PCF is intended for practices with advanced primary care capabilities, but practices are likely at different stages in their ability to implement care delivery changes, which could result in variation of the model's impacts on outcomes.

### Objectives of the evaluation

The primary goal of this evaluation is to determine whether the PCF Model leads to better care for Medicare FFS beneficiaries attributed to participating practices and lowers costs for CMS. In addition to evaluating how the model affects total Medicare FFS expenditures and the acute hospitalization rate (the two primary outcomes of the evaluation), we will also evaluate the impact of the PCF Model on secondary spending and service utilization outcomes.

## Data Sources

The evaluation will rely on three types of data sources: (1) Medicare FFS claims and enrollment data, (2) payment data for the PCF Model and other CMS programs, and (3) area-level data sets with information on beneficiary and practice characteristics.

### Medicare data

We will use Medicare Parts A and B claims and enrollment data from the Enrollment Database and Master Beneficiary Summary File as the principal secondary data sources for Medicare beneficiaries. Claims data provide information on service utilization, expenditures, and diagnostic history. We will use claims data with at least 90 days of runout, the standard for research purposes, and all data processes and programs will be subject to our rigorous quality assurance practices. The Enrollment Database provides information, by month, for beneficiaries enrolled in Medicare during the study period, including whether they were enrolled in Parts A and/or B, whether Medicare was their primary payer of medical bills, whether they were dually enrolled in Medicare and Medicaid, basic demographic data, and date of death if applicable. We will acquire and process all Medicare data within the Chronic Conditions Warehouse Virtual Research Data Center to generate cost and utilization outcomes and other variables.

### CMS payment data

To measure the payments that practices receive through participation in the PCF Model and other CMS initiatives, we will use Medicare claims, Non-Claims Based Payment (NCBP) files, other files on CMS advanced APM bonuses paid, and information from the PCF Payment, Operations, Management, and Quality contractor. For PCF-related payments, we will use claims to calculate amounts of Flat Visit Fees for primary care services and obtain PBAs from CMS. These data will inform the magnitude of financial supports and penalties practices receive through participation in the PCF Model. In addition, we will include revenue that PCF and comparison practices may receive from their participation in other CMS-sponsored initiatives. Specifically, the paid amounts on Medicare FFS claims reflect adjustments for the Merit-Based Incentive Payment System. For payments through other initiatives not represented in the claims, such as Quality Payment Program payments for practices in advanced APMs and the Medicare Shared Savings Program (SSP) adjustments for practices in accountable care organizations, we will use NCBP data from CMS, the SSP Public Use Files, and other CMS-provided data.

### Other data sources

We will obtain practice characteristics from several data sets.

**PCF participant list.** We will obtain a list of PCF participating practices and their addresses from CMS. The CMS list includes the National Provider Identifiers (NPIs) and Tax Identification Numbers the practices use for billing.

**IQVIA OneKey.** We will obtain a list of *all* primary care practices across the United States and their primary care clinicians each year from IQVIA OneKey, a proprietary database. We will link the CMS practice list to IQVIA OneKey and use OneKey's NPI lists as the basis for the evaluation's analytic files (for both the PCF and potential comparison groups) to ensure consistent methodology for defining practices across the two groups. The OneKey data also include practice characteristics (such as size and whether the practice is part of a hospital-based system) used to select the comparison group and for analysis.

**Data sources for other characteristics, including characteristics of the practices' locations.** We will construct and maintain a database of practice information to match PCF practices to comparison practices and to track practices' organizational characteristics and participation in initiatives over the course of the evaluation. This file will draw on the following data sources: (1) medical home recognition from the National Committee for Quality Assurance; (2) the Area Health Resource File and American Community Survey, which provide characteristics of the practices' locations (such as urbanicity, median household income, and percentage of households below the federal poverty level); (3) the Social Vulnerability Index and Pandemic Vulnerability Index; (4) CMS's Master Data Management, which identifies participants in CMS initiatives both at baseline and after the start of the model; and (5) CMS's NCBP data, which identify participation in models that are not in the Master Data Management, such as episode-based payment models.

## Study Design

To estimate the impact of the PCF Model on outcomes of interest, we will compare outcomes of PCF practices, calculated from beneficiaries assigned to PCF practices, with outcomes calculated for beneficiaries assigned to matched comparison practices. The core strategy for estimating impacts is a quasi-experimental difference-in-differences design, where we estimate impacts based on differences in outcomes before and after model start between beneficiaries served by PCF practices and beneficiaries in a matched comparison group. The model testing period for the difference-in-differences regression analysis is the five years of PCF, and the baseline period is the two years before practices joined PCF. These periods reflect slightly different calendar years for Cohort 1 (2019–2020 baseline and 2021–2025 model testing period) versus Cohort 2 (2020–2021 baseline and 2022–2026 model testing period). For both cohorts, we remove the first two calendar quarters of data from the annual observation for 2020. This is due to concerns that large swings in outcomes during the early months of the

COVID-19 pandemic could violate the parallel trends assumption underpinning a difference-in-differences regression analysis.

We will conduct an intent-to-treat evaluation, following outcomes over the full five-year model (for each cohort) among all practices that join the PCF Model, including those that drop out before the end of the model period. Similarly, we will follow all Medicare FFS beneficiaries ever assigned during the model if they remain in Medicare FFS, even if some stop visiting a PCF (or comparison) practice. Following the logic CMS uses to calculate PCF payments, we will first attribute beneficiaries to practices based on their most recent Medicare annual wellness visit or welcome-to-Medicare visit, or, for beneficiaries with neither visit type in the previous 24 months, based on the plurality of their primary care visits in the previous 24 months. We will then assign beneficiaries to practices where they were first attributed in a period (baseline and model-testing years). The intent-to-treat approach stabilizes the sample size and limits the possibility that differential attrition between the PCF and comparison groups bias impact estimates.

The comparison group will comprise a matched set of primary care practices located within the PCF regions but that are not participating in the PCF Model. The goal of matching is to identify a set of comparison practices with comparable observable characteristics to those of PCF participants. Our matching approach selects the comparison group at the practice level (defined as the physical site where the practice is located), and we will roll up data from beneficiaries to the practice-year level as the unit of observation in the regression analysis.

The evaluations primary research questions are as follows:

1. Does the PCF Model reduce total Medicare FFS expenditures relative to the comparison group, as measured among all Medicare FFS beneficiaries assigned to PCF practices, in one or more of the five years of the PCF Model testing period?
2. Does the PCF Model reduce the acute hospitalization rate, as measured among all Medicare FFS beneficiaries assigned to PCF practices, in one or more of the five years of the PCF Model testing period?

We will use both frequentist and Bayesian methods to estimate impacts of the PCF Model (described in more detail below). When publishing findings, we plan to present both frequentist impact estimates, including traditional *p*-values and confidence intervals, and probabilities based on Bayesian estimates, as described in subsequent sections of this protocol.

## Outcomes

We plan to estimate impacts of the PCF Model for each model performance year on two primary outcomes and additional secondary outcomes (Table 1). We will construct the outcomes for all PCF beneficiaries and their matched comparison beneficiaries. The primary outcomes enable us to answer the primary research question. An important use of secondary outcomes is to identify intermediate factors potentially driving the impacts on primary outcomes. For example, if we observe reductions in the acute hospitalization rate (a primary outcome), we can then examine whether these reductions are associated with reductions in medical admissions. Even if we do not observe impacts on primary outcomes, the secondary outcomes can provide evidence about the impact of practices' changes on patients' care.

Our list of secondary outcomes includes three measures defined at the discharge level (that is, with the measure's denominator defined as hospital discharges rather than beneficiary-year) and we might change the measure specification for these three if we find impacts on acute hospitalizations that could affect the measures' denominator. Specifically, if the PCF Model is effective at reducing acute hospitalizations, then this could affect the relative severity of hospitalizations among beneficiaries with a qualifying discharge. For example, if PCF reduced hospitalizations by 10 percent by avoiding the lowest-acuity hospitalizations, the remaining discharges would no longer be similar between the PCF and comparison groups. As a result, we may observe higher unplanned readmission rates, higher unplanned acute care rates, and higher post-acute care expenditures in the PCF group. To address this issue, we may assess impacts on versions of the relevant measures that are calculated over all assigned beneficiaries, not just those with a qualifying discharge, in the event we observe meaningful impacts on acute hospitalizations (a primary outcome). The three relevant measures are as follows: (1) proportion of discharges with a 30-day unplanned readmission; (2) proportion of discharges with any unplanned acute care within 30 days; and (3) post-acute care expenditures per post-acute care episode.

Table 1. Outcome measures for PCF impact evaluation, with rationale for inclusion

| Measure | Rationale for inclusion |
|---|---|
| **Primary outcomes** | |
| Total Medicare Parts A and B expenditures<br>($ per beneficiaries per month) | Impacts on expenditures are central to CMS's decisions to expand an Innovation Center model and used to determine model PBA payments for practices. |
| Acute hospitalization rate (per 1,000 beneficiaries per year) | Given the model payment structure rewards practices for decreasing hospitalizations, we hypothesize acute hospitalizations are the primary mechanism for reduced expenditures. |
| **Secondary outcomes** | |
| Medical admissions (per 1,000 beneficiaries per year) | Evaluations of similar primary care models (such as CPC+) found participating practices reduced medical admissions. We may expect to see impacts on this before impacts are evident on the broader acute hospitalization measure, which includes surgical admissions. |
| Outpatient ED visits | Care delivery activities that seek to reduce acute hospitalizations may reduce ED utilization. |

| Measure | Rationale for inclusion |
|---|---|
| (per 1,000 beneficiaries per year) | |
| Primary-care-substitutable ED visits<br>(per 1,000 beneficiaries per year) | Care delivery activities that seek to reduce acute hospitalizations may reduce ED utilization; effects may be concentrated among primary-care-substitutable ED visits. |
| Proportion of inpatient discharges at the practice that had a 30-day all-cause unplanned readmission | We may expect to see reductions in hospital readmissions through practices' focus on episodic care management. |
| Proportion of inpatient discharges at the practice with unplanned 30-day acute care | We may expect to see reductions in unplanned acute care (including ED visits, observation stays, and unplanned readmissions) through practices' focus on episodic care management. |
| PAC expenditures per PAC episode | Fewer acute hospitalizations may result in lower total PAC expenditures if savings in higher-cost institutional care (such as skilled nursing facility stays) offset potential cost increases in lower-cost home health expenditures. |
| Inpatient expenditures<br>($ per beneficiaries per month) | Fewer acute hospitalizations may result in lower inpatient expenditures, contributing to lower total Medicare FFS expenditures. |

CMS = Centers for Medicare & Medicaid Services; CPC+ = Comprehensive Primary Care Plus; ED = emergency department; FFS = fee-for-service; PAC = post-acute care; PBA = Performance-based Adjustment; PCF = Primary Care First.

In addition to examining whether the model impacts primary and secondary outcomes, we will seek to understand the causal mechanisms driving any impacts we observe using a set of leading indicators. Leading indicators measure process changes practices make that are expected to lead to changes in primary and secondary outcomes. For example, the literature shows continuity of care is associated with reduced hospitalizations and lower costs (Nyweide et al. 2013; Bazemore et al. 2018). As such, if we observe reductions in hospitalizations or Medicare expenditures, we may examine effects for a leading indicator that measures continuity of care beneficiaries are receiving as part of their health care management. We will present the results for leading indicators as exploratory analyses. In addition, we might update the leading indicators we ultimately examine during the PCF performance period to reflect hypothesized mechanisms of impacts for the outcomes in Table 1, based on impact estimates and qualitative data we collect during the evaluation.

# Regression Modeling

## Frequentist approach

Our main frequentist specification is a linear regression model that follows the approach introduced by Sun and Abraham (2021). Equation (1) shows the main regression specification:

$$\bar{y}_{jt} = \rho_j + \alpha_t + \sum_{c} \sum_{\tau \neq -1} \delta_{c,\tau} PCF_j * 1\{C_j = c\} * 1\{t - C_j = \tau\} + \beta_t X_j + \beta_t \bar{X}_{jt} + \varepsilon_{jt} \qquad (1)$$

where $\bar{y}_{jt}$ denotes the mean outcome for practice *j* in year *t*; $\rho_j$ and $a_t$ denote practice and calendar-year fixed effects, respectively; $PCF_j$ is an indicator that denotes whether a practice is in the PCF group; $C_j$ is the year when practice *j* is first treated; and $\tau$ represents the relative year before and after the model testing period starts (for example, $\tau$ = -2 denotes two years prior to PCF). We define Year 1 of the model testing period as the first year of a practice's participation in the model—so that Year 1 corresponds to calendar year 2021 for Cohort 1 and 2022 for Cohort 2. We also control for $X_j$ and $\bar{X}_{jt}$, which represent practice characteristics (such as health system affiliation) and practice averages of beneficiaries' characteristics (such as age and other demographics). These characteristics are measured at the start of the baseline and model testing periods and interacted with each year. We weight practices by matching and beneficiary observability (total days enrolled in Medicare FFS) weights each year, thus preserving the interpretation of estimating the causal impact of PCF on the average beneficiary. Finally, we cluster standard errors at the practice level.

The Sun and Abraham approach produces estimates that can be interpreted as average treatment effects that are robust to contamination from treatment-effect heterogeneity in settings with staggered model start (as is in the case for PCF). The Sun and Abraham approach works as follows:

1. Estimate *cohort-specific* average treatment effects for each year relative to the model start date.
2. Calculate cohort shares in each relative year, which are equivalent to the (weighted) shares of assigned beneficiaries in Cohort 1 PCF practices and Cohort 2 PCF practices, relative to the total number of assigned beneficiaries to PCF practices in the same relative year.
3. Estimate the overall (combined) treatment effect in each relative year by combining cohort-specific estimates from Step 1 within each relative year, using cohort shares in Step 2 as weights.

In Equation (1), $\hat{\delta}_{c,\tau}$ represents a consistent estimate for the *cohort-specific* average treatment effect in each relative year, conditional on covariates. Aggregating the coefficients $\hat{\delta}_{c,\tau}$ using the cohort shares yields a consistent estimator of the average treatment effect of PCF overall for each relative year. Thus, applying the Sun and Abraham approach to Equation (1), we estimate model impacts (across both cohorts) for each of the five model performance years. We will also generate a cumulative (average) estimate of PCF's impacts over the full five-year period, based on the annual estimates.

To address the evaluation's primary research questions—assessing whether impacts occurred in *any* of the five model performance years—we will then, for each primary outcome, conduct an F test assessing whether any of the five annual impact estimates is statistically significantly different from 0. Consistent with other CMS evaluations, we will use a two-sided test with alpha = 0.1 as our threshold of statistical significance.

### Hybrid frequentist-Bayesian approach

We plan to use a hybrid frequentist-Bayesian methodology designed to draw on the complementary strengths of frequentist and Bayesian approaches. Following Lipman et al. (2022), we will use a two-stage modeling strategy that pairs a frequentist difference-in-differences regression with a Bayesian meta-regression. In the first stage, we will fit a cohort-specific frequentist difference-in-differences regression, as shown in Equation (1). In the second stage, we will fit a Bayesian meta-regression to the impact estimates and their estimated variance-covariance matrix. We will estimate hybrid Bayesian models for all the outcomes listed in Table 1 and for the set of subgroups described in the next section (Table 2). This approach improves the precision and plausibility of the impact estimates by borrowing strength across subgroups and simultaneously adjusting for multiple comparisons (Gelman et al. 2012). The Bayesian approach also enables us to draw probabilistic conclusions through statements such as, "There is a 60 percent chance that PCF reduced acute hospitalizations by 2 percent or more."

To follow the best practice that prior distributions used in program evaluation represent prior information, rather than researchers' subjective beliefs, we will review the literature to assemble evidence about the magnitudes of impact estimates observed in previous similar evaluations (Kaplan 2019). We will use this literature review to define, for each outcome, the prior distribution for a central parameter in our model: the overall impact of PCF. For other parameters in the model, we will incorporate information about the likely magnitudes of impacts for different outcomes and subgroups into the prior distributions.

## Subgroups

The impacts of PCF could differ across subgroups. Therefore, we will estimate the effects of PCF on the list of practice-level subgroups shown in Table 2. We will use both frequentist and hybrid frequentist-Bayesian methods to examine impacts by subgroups, as described previously. To reduce the risk of spurious results when making multiple comparisons, we will limit the subgroup analysis to the primary outcomes listed in Table 1. As with the leading indicators, we may choose to examine effects on additional subgroups beyond those listed in Table 2 (such as whether a practice is affiliated with a hospital-based health system) in exploratory analyses to help shed light on causal mechanisms.

Table 2. List of subgroups for PCF evaluation

| Subgroup |
|---|
| 1. Participants in CPC+ versus nonparticipants in CPC+ |
| 2. SSP participants versus SSP nonparticipants measured when PCF began |

CPC+ = Comprehensive Primary Care Plus; SSP = Medicare Shared Savings Program; PCF = Primary Care First.

## Testing for Parallel Trends and Reporting Results

Parallel trends in outcome trajectories between PCF and comparison practices (in the absence of the PCF Model) is a necessary assumption for interpretating results from difference-in-differences designs as unbiased estimates of causal impacts. We do not intend to publish results for outcomes or subgroup–outcome combinations that fail our test of baseline trends, except in supplemental appendices noting study limitations.

When producing impact estimates for each outcome and subgroup, we will test for the possibility of PCF–comparison differences in baseline outcome trends. Our parallel trends test will proceed in two steps:

**Step 1:** We will use our difference-in-differences regression model (Equation [1]) to assess whether PCF–comparison trends differ between the first and second years of the two-year baseline period. We will consider our findings to have passed the parallel trends test for each outcome for which the estimated PCF–comparison difference in baseline trends (Year 1 to Year 2) is not statistically significantly different from 0 at the 90 percent confidence level ($p > 0.1$). That is, we will assume baseline trends are parallel if we do not find statistically significant evidence ($p \leq 0.1$) of nonparallel trends. Given that statistical power to detect divergent trends among subgroups is less than the full population, we will use a narrower confidence interval (80 percent) for subgroup–outcome combinations. Using a narrower confidence level for subgroups helps identify true trend differences that would otherwise not be statistically detectable but that could meaningfully alter conclusions about impacts of the model during later performance years. Finally, we will assume—for any given outcome—that we fail the parallel trends test for all subgroups if we have failed the test for the full population. For all outcomes and subgroup–outcome combinations that fail the parallel trends test, we will proceed to the next step.

**Step 2:** For any outcomes and subgroup–outcome combinations that fail the parallel trends criteria described in Step 1, we will assess whether accounting for the estimated trend difference in Step 1 meaningfully alters conclusions we would draw from the impact estimates. Specifically, we will check whether conclusions change in terms of direction or statistical significance of the impact estimates (for example, a change from statistically significant to insignificant) after accounting for baseline trend differences projected into the model testing period. In most cases, we anticipate being able to determine pass/fail in Step 2 by comparing the magnitude of the baseline trend difference to the impact estimates for a given outcome. For example, suppose we find evidence that PCF reduces a service use outcome by one visit per 1,000 beneficiaries per year in Model Year 3 but also find a significant baseline trend difference between PCF and comparison practices of two visits per 1,000 beneficiaries per year. In this scenario, we would fail the parallel trends test because we cannot distinguish whether the statistically significant impact estimate in Model Year 3 is due to the PCF Model or is a result of pre-existing trend differences (which would equal six visits per 1,000 beneficiaries per year by Model Year 3, assuming the trend persists during the intervention). For cases in which comparing the magnitude of the baseline trend difference to the impact estimates does not yield a clear pass/fail decision, we will follow Bilinski and Hatfield's (2019) recommendations and calculate how the impact estimates compare to estimates generated from a difference-in-differences model that adjusts for differences in linear baseline trends.

We expect the final reported results to cover the following:

- For the full population of PCF practices, all outcomes in Table 1 that pass the parallel trends test
- For the subgroups shown in Table 2, primary outcomes for which the subgroup–outcome combination passes the parallel trends test

## References

Bazemore, A.W., S. Petterson, L.E. Peterson, R. Bruno, Y. Chung, and R.L. Phillips. "Higher Primary Care Physician Continuity Is Associated with Lower Costs and Hospitalizations." *Annals of Family Medicine*, vol. 16, no. 6, 2018, pp. 492–497.

Bilinski, Alyssa, and Laura A. Hatfield. "Nothing To See Here? Non-Inferiority Approaches to Parallel Trends and Other Model Assumptions." *arXiv*:1805.03273, 2019. https://doi.org/10.48550/arXiv.1805.03273.

Gelman, A., J. Hill, and M. Yajima. "Why We (Usually) Don't Have to Worry About Multiple Comparisons." *Journal of Research on Educational Effectiveness*, vol. 5, no. 2, 2012, pp. 189–211.

Kaplan, David. "Bayesian Inference for Social Policy Research." OPRE Report #2019-36. U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research, and Evaluation, 2019.

Lipman, E.R., J. Deke, and M.M. Finucane. "Bayesian Interpretation of Cluster-Robust Subgroup Impact Estimates: The Best of Both Worlds." *Journal of Policy Analysis and Management*, vol. 41, no. 4, 2022, pp. 1204–1224.

Nyweide, D.J., D.L. Anthony, J.P.W. Bynum, R.L. Strawderman, W.B. Weeks, L.P. Casalino, and E.S. Fisher. "Continuity of Care and the Risk of Preventable Hospitalization in Older Adults." *JAMA Internal Medicine*, vol. 173, no. 20, 2013, pp. 1879–1885.

Sun, L., and S. Abraham. "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects." *Journal of Econometrics*, vol. 225, no. 2, 2021, pp. 175–199.