The Impact of Black Lives Matter Movement on the Evaluation of Health Providers

Sepehr Khadem
Cornell University

Omid Rafieian
Cornell University

Vrinda Kadiyali
Cornell University

Abstract

This paper examines how the Black Lives Matter (BLM) movement, following the murder of George Floyd in 2020, shaped patient evaluations of Black healthcare providers. Using more than 1.5 million reviews from Healthgrades.com and a matched sample of Black and White providers, we estimate the effect of the BLM movement by treating Floyd's murder as an exogenous shock in time. We find that ratings for Black providers increased by about 0.05 stars relative to White providers, driven by fewer 1-star and more 5-star reviews, alongside a broad decline in negative written feedback. These improvements persisted beyond the immediate aftermath of the protests, indicating durable cultural spillovers rather than short-lived reactions. Heterogeneity analyses show that the effect was concentrated among Black male providers and was larger in counties with higher racial segregation, while protest type and intensity had little explanatory power. Taken together, the results demonstrate that social movements can influence perceptions in evaluation systems, which are widely used in healthcare markets and policy. This raises concerns about the reliability of online ratings as unbiased measures of provider quality and highlights the need for evaluation systems that are less prone to bias-driven volatility.

1 Introduction

1.1 Social Movements and The Case of Black Lives Matter

Social movements play a critical role in shaping public attitudes, policies, and institutions. They can shift cultural norms, influence political agendas, and bring attention to longstanding inequities. The murder of George Floyd in May 2020 brought the Black Lives Matter (BLM) movement to the center of public attention, sparking one of the largest waves of protest in recent U.S. history. Millions of people across all fifty states, as well as in countries around the world, participated in mass demonstrations. These protests were not confined to major cities but extended into suburbs, small towns, and rural communities. Supporters of the movement demanded increased accountability for law enforcement, reforms to address racial disparities in the justice system, and broader structural change to confront entrenched racial inequities in American society.

Scholars have demonstrated that the BLM movement and other large-scale mobilizations can bring about significant and lasting change across multiple domains. For example, Madestam et al. (2013) find that Tea Party protests influenced Republican vote share and public policy, while Levy and Mattsson (2023) and Luo and Zhang (2021) link the #MeToo movement to increased reporting of sexual harassment and shifts in workplace practices. Research on BLM itself highlights a reduction in police homicides in protest-affected areas (Campbell, 2021) and increased demand for antiracist curricula in schools (Agarwal and Sen, 2022). Studies also document how the movement altered online engagement, with Liu et al. (2022) showing short-lived surges in platform activity, and others finding that increased visibility of minority-owned businesses influenced consumer reviewing behavior (Son et al., 2025; Babar et al., 2023; Aneja et al., 2025). From a cultural perspective, Lin et al. (2024) show that negative biases against racial minority representation in films weakened in the wake of BLM, especially where protests were most intense. Taken together, these findings

underscore the broad societal reach of the BLM movement and its capacity to reshape behavior, institutions, and cultural perceptions.

1.2 Research Agendas and Challenges

In this paper, we examine the impact of BLM protests in the healthcare context. A substantial body of work has documented racial disparities in diagnosis, treatment quality, and patient outcomes linked to both patient and provider race (Smedley et al., 2003; Williams and Rucker, 2000). Additionally, studies found that minority physicians frequently experience discrimination from colleagues and patients, with consequences for professional advancement and clinical relationships (Nunez-Smith et al., 2009b,a). They also often report lower levels of patient trust and satisfaction (Saha et al., 1999; Traylor et al., 2010). Further, patients of color are less likely to receive adequate pain management or evidence-based treatments than White patients (Green et al., 2003; Hoffman et al., 2016). At the same time, evidence shows that racial concordance between providers and patients can improve trust, adherence, and outcomes (Alsan et al., 2019). These entrenched inequities make healthcare a critical context for assessing whether social movements such as BLM have the potential to shift public evaluations of minority providers.

Despite the extensive literature on both BLM and healthcare disparities, no prior research has examined whether the BLM movement directly influenced how patients evaluate healthcare providers. This question is especially important because the quality of service for online platforms is unlikely to change in response to the movement, meaning that any observed shifts in ratings may reflect underlying racial biases rather than differences in care. Addressing this gap, our study investigates whether a broad societal movement translated into changes within the healthcare sector. Specifically, we ask:

1. How did the BLM movement following George Floyd's murder change the way patients

rated Black health providers compared to White providers?

2. How long did these changes in patient evaluations last?

3. How did these changes differ across providers' sex, protest variation, and county-level socioeconomic conditions?

Answering these questions presents several challenges. First, we require a dataset with sufficient numbers of reviews for both Black and White health providers to ensure reliable estimates. Second, identifying provider race is difficult because such information is not typically reported in structured forms. Third, we need to construct a sample that allows for apples-to-apples comparisons across providers, ensuring that observed differences are not driven by confounding attributes. Fourth, we must implement a research design that credibly isolates the effect of the BLM movement from other potential confounders.

1.3 Our Approach

To address these challenges, we draw on data from Healthgrades.com, one of the largest online platforms where patients rate and review healthcare providers. This platform provides both a sufficient scale of reviews and the necessary granularity to compare Black and White providers over time.

A key obstacle is the absence of self-reported race information. We address this by employing a face-recognition algorithm to infer provider race from profile photographs, generating consistent and reproducible racial classifications. While imperfect, this approach enables us to systematically assign providers into racial categories and construct a dataset suitable for analysis.

To ensure comparability across providers, we implement a matching strategy. Each Black provider in the sample is paired with a White provider who shares the same specialty, practices in the same or nearby ZIP code, and has a similar review history and observable

characteristics. This design reduces confounding by aligning providers on core factors that could independently influence patient ratings, creating an "apples-to-apples" sample for inference.

Finally, our identification strategy leverages the murder of George Floyd in May 2020 as an exogenous temporal shock. We assume this event—and the subsequent rise of the Black Lives Matter movement—affected patient perceptions of Black providers but not White providers in the same way. By comparing rating changes across matched pairs before and after the event, we separate the differential effect of the BLM movement from broader temporal shocks such as pandemic-related disruptions or seasonal review cycles, as well as from provider-specific reputation trajectories. This difference-in-differences approach allows us to isolate the causal impact of the BLM movement on patient evaluations of Black healthcare providers.

1.4 Findings and Contribution

Our analysis shows that the BLM movement following George Floyd's murder led to measurable changes in how patients evaluated Black healthcare providers. On average, Black providers experienced an increase of about 0.05 stars relative to their White counterparts. This shift was not marginal in character but driven by meaningful changes at the extremes of the distribution: a reduction in 1-star ratings and a corresponding increase in 5-star ratings. Written reviews also reflected this shift, with a broad decline in negative feedback toward Black clinicians and their offices. These patterns suggest that the BLM movement influenced not just the numerical rating but also the narrative through which patients assessed care.

The effect was persistent. Rather than fading after the initial wave of protests, the improvement in evaluations extended well beyond the immediate aftermath. This persistence points to cultural spillovers that reshaped perceptions in more durable ways, suggesting that the movement initiated longer-term shifts in how patients engaged with and rated Black providers. The durability of these changes indicates that social movements can affect social

evaluations in ways that outlast the triggering events.

We also document systematic heterogeneity in the response. The improvement was concentrated among Black male providers, while no significant effect was observed for Black female providers. This sex-based divergence highlights how intersectionality shapes public perceptions and reveals that not all providers benefited equally. At the community level, the impact was larger in counties characterized by higher racial segregation, underscoring the importance of structural conditions in moderating the effects of social movements. In contrast, variation in protest type or intensity did not explain differences in the observed effects, pointing to demographic and structural environments as more decisive drivers than protest dynamics alone.

Substantively, the study makes several contributions. First, it demonstrates that the reach of social movements extends into domains beyond their immediate focus, as seen here in the case of healthcare evaluations. Second, it demonstrates that patient ratings—widely used in healthcare markets, policy design, and provider reputations—are not insulated from societal context. They respond to cultural and political currents, raising concerns about their reliability as unbiased measures of provider quality. Third, by showing that evaluations of Black providers improved even absence of evidence that the quality of care itself changed, the findings underscore the role of racial bias in shaping reputational outcomes and reveal how collective mobilization can disrupt entrenched evaluative patterns.

For policymakers and healthcare organizations, these insights underscore the importance of interpreting online ratings with caution, acknowledging their vulnerability to social and cultural influences. For platform designers, the results point to the importance of safeguards against bias-driven volatility that can distort reputations. More broadly, our findings reveal that social movements can produce enduring cultural spillovers, reshaping perceptions of marginalized groups and altering evaluative practices within key institutions. This raises concerns about the design of evaluation systems themselves: if patient ratings respond not only

to providers' performance but also to external societal shifts, then using them as benchmarks for quality comparison risks conflating care quality with broader cultural dynamics. Designing evaluation systems that are more robust to such shifts—through weighting schemes, safeguards, or complementary measures—is therefore essential to ensure that reputational outcomes and institutional decisions are not unduly influenced by bias-driven volatility.

2 Literature Review

Our study intersects with three existing literatures and fills a clear gap. First, it relates to the growing body of work on the impact of social movements on societal, economic, and political outcomes. Madestam et al. (2013) show that Tea Party protests shifted Republican vote share and public policy. Levy and Mattsson (2023) and Luo and Zhang (2021) connect the #MeToo movement to changes in reporting sexual harassment and hiring practices. Campbell (2021) studies the impact of BLM protests on deaths resulting from police violence in affected neighborhoods, reporting a decrease in police homicides. Agarwal and Sen (2022) examine the BLM movement's effect on classroom curricula about race, documenting a surge in antiracism requests following George Floyd's death. Taken together, these studies establish that social movements can produce measurable shifts across multiple domains. We extend this literature by showing that social movements can also reshape consumer evaluations in the healthcare sector. Specifically, we provide evidence that BLM protests led to persistent improvements in how patients rate Black providers relative to White providers, demonstrating a novel domain—healthcare perceptions—through which social movements influence institutional outcomes.

A narrower strand of work focuses on how the BLM movement affected online review platforms. Liu et al. (2022) analyze whether user engagement on review platforms changes following social causes and report that support typically lasts one to three months. Several

studies extend this line of inquiry by showing that increasing the visibility of minority-owned businesses on online platforms alters reviewer behavior and is influenced by contextual factors (Son et al., 2025; Babar et al., 2023; Aneja et al., 2025). From a cultural perspective, Lin et al. (2024) find that increasing racial minority presence in movie sequels predicted lower ratings and more toxic reviews before BLM, but these negative effects weakened after the advent of BLM, especially when movement intensity was high. Building on this work, our study extends the analysis to healthcare, a domain central to well-being but understudied in the context of online reviews. We show that the BLM movement produced not only short-term engagement but also durable shifts in how patients evaluate Black healthcare providers.

This focus also connects our study to the broader literature on the challenges faced by minority business owners. Research documents disadvantages spanning access to resources, financing, and networks (Fairlie and Robb, 2007; Fairlie et al., 2022), while consumer biases further limit demand for minority entrepreneurs' products and services (Gligor et al., 2021a). Investors, too, react unfavorably to the appointment of a Black CEO (Gligor et al., 2021b). Similar patterns appear in digital markets: minority and female proprietors experience performance gaps compared to majority counterparts (Ayres et al., 2015; Younkin and Kuppuswamy, 2019). Digital platforms often amplify these inequities, with evidence of racial discrimination on Airbnb (Edelman et al., 2017), biased outcomes in freelance markets (Hannák et al., 2017), and lower crowdfunding support for Black founders (Younkin and Kuppuswamy, 2018; Yazdani et al., 2025). By situating our work in this literature, we highlight that such biases are not immutable. Our evidence suggests that the BLM movement mitigated disadvantage in healthcare markets by improving patients' ratings of Black providers, indicating that large-scale mobilization can shift consumer discrimination in professional service contexts.

Finally, our research builds directly on the literature documenting racial discrimination in healthcare. Studies have shown longstanding disparities in treatment, diagnosis, and

outcomes linked to the race of both patients and providers (Smedley et al., 2003; Williams and Rucker, 2000). Minority physicians face discrimination from colleagues and patients, with consequences for their career advancement and relationships (Nunez-Smith et al., 2009b,a). They often encounter lower trust and satisfaction from patients (Saha et al., 1999; Traylor et al., 2010). Patients of color are less likely to receive adequate pain management or guideline-based treatments than White patients (Green et al., 2003; Hoffman et al., 2016), while racial concordance between providers and patients improves trust, adherence, and outcomes (Alsan et al., 2019). By linking this literature with research on social movements, we demonstrate that the BLM movement improved ratings of Black providers. This connection highlights how large-scale societal mobilization can alter patient perceptions and, in turn, the professional standing of minority physicians in healthcare markets.

3 Settings and Data

3.1 Settings

For this study, we focus on reviews collected from the website Healthgrades.com, one of the largest online platforms in the United States, where patients can rate and review healthcare providers. Healthgrades enables users to evaluate physicians and other healthcare professionals across multiple dimensions, including overall satisfaction, bedside manner, wait times, and staff performance. The platform is widely used by patients seeking information about prospective providers, making it a valuable source of unsolicited, organic feedback that reflects real-world patient experiences.

Each provider's page on Healthgrades follows a standard format that includes demographic and professional information, aggregated feedback scores, and patient reviews. For this study, the relevant information is presented in two main sections:

1. **Provider Demographic Info:** At the top of a provider's page, users see basic demographic information such as name, professional prefix/suffix, specialty, location, and a profile photograph. Figure 1a shows an example of this demographic information.

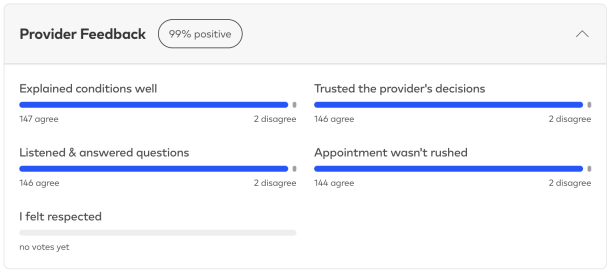
2. **Ratings:** Written patient reviews appear after the aggregated feedback on the provider and office/staff. Each review includes a numeric star rating that indicates the patient's overall satisfaction, along with a written comment. Figures 1d and 1e present sample reviews with ratings. Additionally, patients can provide both positive and negative feedback on a list of what went bad/well in terms of provider feedback and office plus staff feedback. The aggregated feedback is illustrated as the examples in Figures 1b and 1c on the provider's page.

3.2 Data

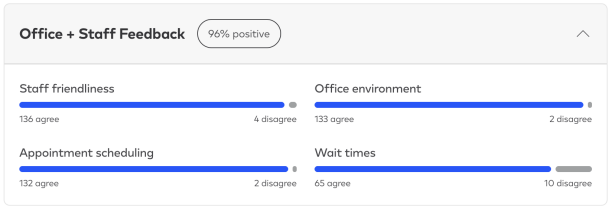
We built our dataset by scraping Healthgrades.com. The scraping program consisted of three processes. In the first process, we collected links to healthcare providers in each specialty along with their overall ratings. In the second process, the output of the previous step was cleaned by removing duplicates and providers without any ratings, as our study focuses only on providers with both ratings and reviews. Finally, in the third process, we scraped all demographic information and reviews for each provider, along with their profile picture.

Since Healthgrades.com launched in late 2015, we restrict our analysis to reviews published from 2016 onward. The dataset consists of textual reviews paired with star ratings, which range from one to five, and includes information about provider characteristics such as name, specialty, and location as shown in Figures 1. For the purposes of this study, we focus specifically on numerical ratings of Black and White healthcare providers to examine whether shifts in patient attitudes following the BLM movement after the murder of George Floyd in 2020 are reflected in differential ratings of these two groups. Only written reviews on

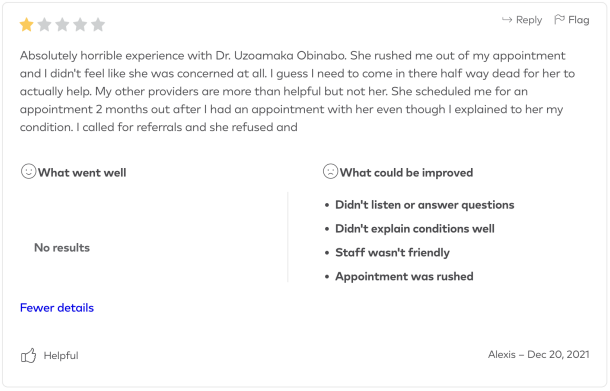
(a) Provider demographic and profile information.



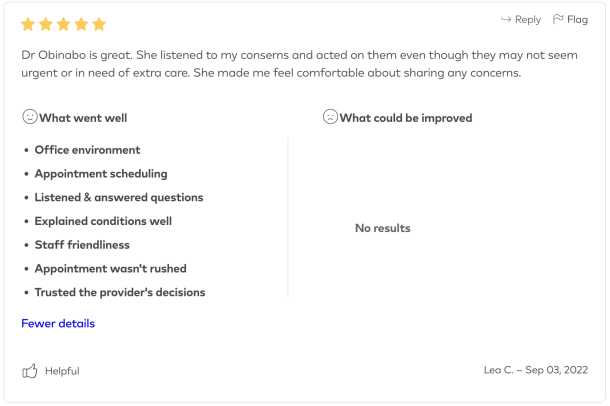
(b) Provider feedback categories.



(c) Office and staff feedback categories.



(d) Low overall rating review.



(e) High overall rating review.

Figure 1: Healthgrades provider page elements.

Healthgrades include the date of the review. Therefore, while we restrict our dataset to providers with written reviews, the present study analyzes only the numerical ratings derived from those reviews.

In total, 2.97 million healthcare provider links were gathered from Healthgrades.com, of which 1.05 million pages contained rating values. Our dataset comprises 6.1 million written reviews from these pages, spanning from 2016 to the end of April 2025. Only 255K providers in the dataset had a profile picture available, who account for approximately 62% of the overall reviews—over 3.8 million reviews in total.

4 Research Design

Measuring the impact of the BLM movement on how patients evaluate Black versus White health providers presents three core challenges. First, publicly available data sources such as Healthgrades.com do not report providers' race, requiring us to construct race information through alternative methods. Second, the sample must be carefully refined to ensure an apples-to-apples comparison, allowing ratings of the treated group to be meaningfully compared to those of the control group. Third, any causal estimate of BLM's impact must account for potential confounding factors—such as shifts in health care delivery, pandemic-related dynamics, or regional variation—that may influence rating changes independently of race. Our research design addresses each of these challenges in turn, and the corresponding strategies are detailed in the subsequent subsections.

4.1 Race Estimation

We fetched the providers' profile pictures while scraping Healthgrades.com. Advances in image processing over the past decade have made it increasingly reliable to estimate demographic attributes, including race, from facial images. Deep learning models, in particular convolutional neural networks, now offer robust accuracy in large-scale classification tasks when trained on diverse datasets. Leveraging these techniques allows us to generate probabilistic estimates of each provider's race, filling the gap left by the absence of self-reported demographic information on the platform.

To perform race estimation, we employ the DeepFace library in Python (Serengil and Ozpinar, 2020). DeepFace is an ensemble method consisting of different state-of-the-art pre-trained models for facial analysis: VGG-Face (Parkhi et al., 2015), Google FaceNet (Schroff et al., 2015), OpenFace[1], Facebook deepface (Taigman et al., 2014), DeepID (Sun

[1]OpenFace, `https://cmusatyalab.github.io/openface/`

et al., 2014), ArcFace (Deng et al., 2019), and Dlib[2]. For this study, we use the VGG-Face, a convolutional neural network–based facial recognition system developed by Parkhi et al. (2015). VGG-Face utilizes a 16-layer deep neural network trained on 2.6 million images of 2,622 individuals to extract features and map them into a high-dimensional vector space, where similarity can be measured with high precision. Although VGG-Face was not trained for demographic inference, Acien et al. (2019) reports a 90.1% accuracy, and Greco et al. (2020) achieves a 94% accuracy. Figure 2 illustrates DeepFace's racial attribute analysis for a sample provider profile.


```
{
 "race": {
  "asian": 1.2331949472427368,
  "indian": 1.0589929819107056,
  "black": 96.674072265625,
  "white": 0.006276246160268784,
  "middle eastern": 0.00217628525570035,
  "latino hispanic": 1.0252983570098877
 },
 "dominant_race": "black"
}
```


Figure 2: Racial attribute analysis with deep learning using the DeepFace library

Several scientific studies have used DeepFace or related frameworks to classify demographic attributes. For instance, Wang and Kosinski (2018) applied deep learning–based facial recognition tools to predict sexual orientation from facial images, while Luca et al. (2024) examined the impact of anti-Asian bias on Airbnb during the pandemic. Moreover, Marx et al. (2025) use DeepFace as a part of their race estimation to measure the impact of BLM on venture funding for Black entrepreneurs.

Applying this method to our dataset, the distribution of estimated dominant racial

[2]Face Recognition with Dlib in Python, `https://sefiks.com/2020/07/11/face-recognition-with-dlib-in-python/`

categories among providers is as follows: White (61.5%), Middle Eastern (10.6%), Asian (10.2%), Latino/Hispanic (9.9%), Black (4.3%), and Indian (3.3%). However, these ratios do not align perfectly with national benchmarks. The Association of American Medical Colleges (AAMC) reports that in 2022, the physician workforce in the United States was composed of 56.5% White, 18.8% Asian, 6.3% Hispanic or Latino (alone or in combination), and 5.2% Black or African American. Fewer than 1.5% of physicians identified as Multiracial (1.3%), Other (1.1%), American Indian or Alaska Native (0.3%), or Native Hawaiian or Other Pacific Islander (0.1%). An additional 10.4% did not report their race or ethnicity (U.S. Physicians in All Specialties, AAMC Report, 2022).

4.2 Matching Procedure

Another challenge arises from the need to construct an apples-to-apples comparison between Black and White providers. Because the number of Black providers in the scraped dataset is much smaller, direct comparisons may reflect differences in sample sizes rather than true differences in patient ratings. To address this, we must carefully design sampling and matching strategies to ensure that comparisons between the two groups are valid and balanced.

Our matching approach proceeds as follows. First, we retain all providers identified as Black. We then perform feature engineering on the dataset to generate comparable attributes across providers. These features include:

- geolocation (latitude and longitude of the provider's ZIP-code centroid)
- gender (dominant gender estimated by DeepFace)
- review history
  - date of first review
  - date of last review

  - lifespan in days between first and last review
  - the average interval in days between consecutive reviews

- professional text profile (a combination of specialty, professional prefix, and professional suffix)

The dataset contains approximately 588 specialties along with diverse prefixes and suffixes. To ensure accurate matching (e.g., comparing nurses to nurses within the same specialty), we construct a text profile for each provider by combining prefix, specialty, and suffix. We then apply TF–IDF vectorization to convert this text corpus into a high-dimensional term-frequency matrix. TF–IDF, widely applied in information retrieval (Salton and Buckley, 1988), computes how important a term is in a document relative to the entire corpus by multiplying its frequency in the document with the inverse of its frequency across all documents. This allows common but uninformative terms to be downweighted, while rare and distinctive terms receive more weight.

Next, we use truncated singular value decomposition (SVD) to reduce the TF–IDF matrix to ten principal components. SVD factorizes the original matrix into orthogonal components that capture the main latent structure of the data. When applied to text, truncated SVD—commonly known as latent semantic analysis (Deerwester et al., 1990)—identifies latent semantic dimensions that group related words and contexts together. This reduction allows us to capture underlying similarities among related specialties and titles while mitigating sparsity and noise. It balances expressiveness with compactness, ensuring that the resulting feature space is suitable for matching providers across groups.

With this feature generation, we obtain a rich set of attributes that allow us to align Black providers with White providers on comparable dimensions such as geography, review history, gender, and professional specialization.

Before introducing the matching algorithm, we first define a loss function that will be

incorporated into it. For a treated provider i and a candidate control j, we consider their feature vectors X_i and X_j. The Euclidean distance $||X_i - X_j||$ measures how far the two providers are from each other in the constructed feature space, representing overall similarity in geography, review history, gender, and professional profile. In addition, we define a penalty term W_j, which reflects the probability that provider j is White. Combining these elements, the loss function is expressed as:

$$L_{ij} = ||X_i - X_j|| - \lambda W_j \tag{1}$$

where λ is a tuning parameter that balances closeness in feature space against racial composition. This formulation ensures that the matching process pairs Black providers with very similar controls while also encouraging selection of those controls most likely to be White.

The matching algorithm proceeds as follows. First, construct a control pool of providers labeled as White. For each treated provider, compute losses against all available controls. Assign the top-10 controls with the smallest losses to the treated provider, and remove selected controls from the pool to avoid reuse. This approach ensures that Black and White providers are matched based on geography, gender, practice profile, and review history, creating balanced groups for subsequent analysis.

We implemented this algorithm for 10,044 Black providers in our dataset. Table 1 compares the matched sample to the complete sample. In the matched sample, 10,044 Black providers were paired with 100,440 White providers, yielding 134,474 and 1,370,769 ratings respectively. The average rating for Black providers in the matched set was 4.54 (SD = 1.22), while the matched White providers had a mean rating of 4.61 (SD = 1.13). By construction, the number of specialties and geographic coverage became more balanced: 300 specialties and 3,820 ZIP codes represented among Black providers compared to 533 specialties and

9,343 ZIP codes for White providers. In contrast, the complete sample contained 136,086 White providers with 1,977,140 ratings across 558 specialties and 10,041 ZIP codes. These comparisons highlight that the matched sample reduces imbalance across provider attributes while maintaining a large number of observations for robust analysis.

Table 1: Comparison of Matched and Complete Samples

	Matched Sample		Complete Sample	
	Black	White	Black	White
Providers (N)	10,044	100,440	10,044	136,086
Ratings Count	134,474	1,370,769	134,474	1,977,140
Mean Rating	4.54	4.61	4.54	4.62
SD Rating	1.22	1.13	1.22	1.11
Specialties	300	533	300	558
ZIP Codes	3,820	9,343	3,820	10,041

4.3 Identification Strategy

The primary goal of this study is to measure the impact of the Black Lives Matter (BLM) movement, following the murder of George Floyd in 2020, on the evaluation of Black healthcare providers relative to White providers. Specifically, we seek to determine whether patient ratings shifted in ways that reflect changes in racial attitudes and perceptions toward Black physicians and other health professionals.

To achieve this, we rely on a balanced, comparable, and apples-to-apples sample of providers. Using the matching framework described earlier, we constructed treatment and control groups consisting of Black and White providers with similar characteristics in terms of specialty, geography, review history, gender, and professional profile. This ensures that observed differences are not driven by structural imbalances across provider types, but instead capture differences that can plausibly be attributed to racial identity and the broader social context.

For causal estimation, we employ a Difference-in-Differences (DiD) research design. Our

identification strategy treats the murder of George Floyd as an exogenous shock in time that plausibly affected ratings of Black providers differently from White providers. This approach leverages temporal variation by comparing the change in average ratings for Black providers before and after 2020 with the corresponding change for matched White providers. By differencing across both time and groups, the DiD estimator removes common shocks affecting all providers, isolating the effect of the BLM movement on perceptions of Black providers. This framework allows us to identify whether the heightened salience of racial justice issues in 2020 produced a measurable shift in patient evaluations of Black healthcare professionals.

Formally, our identification model is specified as:

$$Y_{it} = \beta_0 + \beta(Black_i \times Post_t) + \alpha_i + \tau_t + \theta_i + \epsilon_{it} \tag{2}$$

Here, Y_{it} denotes the number of stars for review i at time t. The coefficient β is the Difference-in-Differences estimator and represents our parameter of interest. $Black_i$ is an indicator variable that equals one if the provider for review i is Black (treated) and zero if White (control). $Post_t$ is an indicator for the post-period (after the 2020 shock). The term α_i captures place fixed effects, controlling for time-invariant regional characteristics. The term τ_t captures time fixed effects, absorbing shocks common to all providers in a given period. Given that the 2020 shock occurred during the COVID-19 pandemic, the combination of time and place fixed effects also helps absorb the broader impact of COVID-19 on provider ratings. Finally, θ_i represents provider fixed effects, which account for unobserved provider-specific heterogeneity. The error term ϵ_{it} captures idiosyncratic variation not explained by the model. To account for within-county correlation in provider ratings, we cluster the standard errors of ϵ_{it} at the county level, since patients typically travel within their county when visiting healthcare providers.

5 Results

We organize the results into four parts. First, we present the main estimates from our identification model, which quantify how the Black Lives Matter movement following the murder of George Floyd affected patients' evaluations of Black providers relative to white providers. Second, we examine the length and persistence of this effect over time to assess whether the observed impact was temporary or sustained. Third, we decompose the main effect across star-rating buckets to identify whether the shift came from reductions in low ratings, increases in high ratings, or both. Finally, we analyze written feedback to distinguish whether the changes were directed toward the provider personally or extended to evaluations of the office and staff.

5.1 Main Results

To estimate whether the 2020 BLM protests shifted patients' evaluations of Black clinicians relative to comparable White clinicians, I implement a difference-in-differences design on a matched sample pairing each Black provider to a similar White provider. The treatment indicator equals one for ratings of Black clinicians; the coefficient on $Post \times Treated$ captures the post–George Floyd change in Black–White rating gaps.

Table 2 reports three specifications. Column (1), with only ZIP and month fixed effects, indicates a positive and statistically significant increase of 0.052 points (s.e. 0.014, $p<.001$) in ratings for Black clinicians relative to their matched White counterparts after the protests. Column (2) adds matched-group fixed effects, comparing each provider with matched providers. The estimate is slightly higher and significant at 0.057 (s.e. 0.011, $p<.001$). Column (3) replaces the matched-group fixed effects with provider fixed effect, providing the most conservative within-provider estimate. The estimate remains positive and significant at 0.030 (s.e. 0.015, $p<.05$).

Table 2: Impact of BLM Protests on Ratings of Black vs. White Providers

	Column (1)	Column (2)	Column (3)
Post × Black	0.052^{***}	0.057^{***}	0.030^{*}
	(0.014)	(0.011)	(0.015)
Num. Obs.	1,505,219	1,505,219	1,505,219
Adj. R^2	0.059	0.115	0.248
FE: Provider			X
FE: Matched-group		X	
FE: ZIP Code	X	X	X
FE: Month	X	X	X

Notes: Each Black provider is matched to a comparable White provider prior to estimation. The outcome is the rating score. Column (1) includes no fixed effects. Column (2) includes matched-group, ZIP-code, and month fixed effects. Column (3) replaces the matched-group FE with the provider; the ZIP and month FE remain. Standard errors in parentheses. Errors are clustered at the county level. Significance levels: $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Taken together, the results indicate a small but statistically significant post-BLM improvement—on the order of three to six hundredths of a rating point—in evaluations of Black providers relative to those of White providers. Given the low number of ratings per provider, provider fixed effects absorb substantial variation and reduce precision; therefore, we present Column (2) as the primary estimate and interpret Column (3) as a more conservative, within-provider, robustness check that yields a slightly lower effect while preserving the provider identification. The plausibility of parallel trends is examined using the Wald Test on the aggregated data for months, quarters, and years.

Building on the main result that Black providers received higher star ratings than comparable White providers on Healthgrades after the 2020 BLM movement, we examine the main impact in more depth across the persistence, changes in good/bad ratings, and separate feedback on the provider from feedback on the office and staff. These tests determine how long the impact persists, whether the post-BLM increase reflects changes in evaluators' sentiment, and which aspects of care reviewers emphasize. We use the same matched sample

and identification strategy as in the main analysis.

5.2 Length and Persistence of the Main Effect

To assess the persistence of the observed effect, we estimate our baseline specification on progressively larger subsets of the matched sample, starting with only one quarter of data following the BLM movement and then sequentially adding subsequent quarters until the full post-treatment sample is included. This approach allows us to examine whether the estimated effect is driven by a short-term spike or whether it is sustained over time.

The estimates indicate that, based on the first quarter of post-BLM data, the coefficient is 0.083. When we extend the window to include two quarters, the estimate decreases to 0.061. As additional quarters are added, the effect continues to decline until six quarters after the movement. After this point, the estimates stabilize and begin to rise slightly, converging to a level consistent with the magnitude observed in the complete matched sample.

As the Figure 3 illustrates, the trajectory of the estimates suggests that the initial response was strongest immediately following the BLM movement, likely reflecting heightened public salience. Over time, the effect attenuates but does not vanish. Instead, it stabilizes at a persistent level, indicating that the improvement in ratings of Black providers relative to white providers is not merely a short-lived reaction but represents a long-term shift in patient evaluation behavior.

5.3 Decomposition of Effect in Star Levels

We use this decomposition to see where the post-BLM shift comes from: fewer punitive 1-stars, more 5-stars, or movements in the middle, and to quantify each star level's contribution to the mean. We estimate changes in the probability that a review is 1, 2, 3, 4, or 5 stars for Black providers relative to matched White providers, comparing the pre-shock period

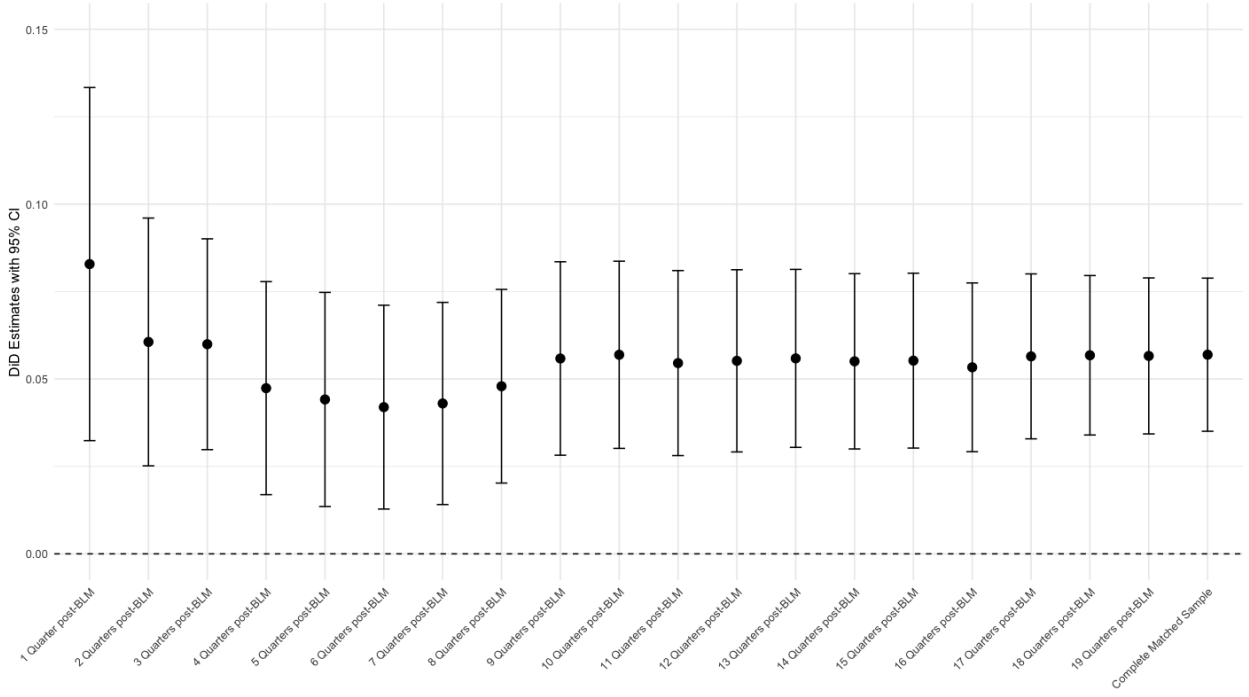


Figure 3: DiD estimates of the effect of BLM across progressively larger post-treatment windows. The effect was strongest right after the BLM movement, reflecting its peak public visibility. It then declined but did not disappear, instead leveling off at a steady point. This shows that higher ratings for Black providers compared to white providers are not just a temporary reaction but signal a lasting change in patient evaluation patterns.

to the post-George Floyd period. Table 3 shows that shifts are concentrated in the tails. The 1-star probability for Black providers declines by 1.4 percentage points (s.e. 0.2 pp, $p<.001$). The 5-star probability rises by 1.3 pp (s.e. 0.3 pp, $p<.001$). Changes at 2-, 3-, and 4-stars are small and imprecise: +0.001 pp (0.07 pp), −0.03 pp (0.06 pp), and +0.10 pp (0.10 pp), respectively. With matched-group, ZIP, and month fixed effects, these shifts imply an approximately 0.054-point increase in the expected rating $\left(\sum_{k=1}^{5} k \cdot \Delta \Pr[r{=}k]\right)$, consistent with the 0.057 main estimate. The pattern indicates fewer punitive reviews and more top marks rather than midpoint compression.

Table 3: Distributional Effects on Star Ratings for Black vs. White Providers

	1-star	2-star	3-star	4-star	5-star
Post × Black	-0.014^{***}	0.00001	-0.0003	0.001	0.013^{***}
	(0.002)	(0.0007)	(0.0006)	(0.001)	(0.003)
Num. Obs.	1,505,219	1,505,219	1,505,219	1,505,219	1,505,219
Adj. R^2	0.107	0.014	0.005	0.009	0.093
FE: Matched-group	X	X	X	X	X
FE: Zip Code	X	X	X	X	X
FE: Month	X	X	X	X	X

Notes: Each column estimates a separate DiD where the outcome is an indicator for receiving a k-star review. Sample matches Black providers to comparable White providers. All specifications include matched-group, ZIP-code, and month fixed effects. Standard errors in parentheses. Errors are clustered at the county level. Significance levels: $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

5.4 Shifts in Patient Feedback on Providers and Office+Staff

To clarify what drives the mean rating gain, we shift from stars to structured feedback and ask whether the post-BLM change reflects evaluations of the provider, the practice environment, or both. As described in the settings section, each review can mark items in two menus—provider feedback and office plus staff feedback—with separate "what went well" (positive) and "what went bad" (negative) selections in the same review. We analyze four predefined provider items and four predefined office plus staff items that are available throughout our sample window; the "feeling respected" item was introduced in 2023 and is excluded. Each item can appear as a positive or negative flag, and we model the probabilities of these flags at the review level to connect dimension-specific shifts to the overall rating effect. We estimate the same identification model separately for each of the 16 feedback indicators to quantify item-specific changes. We now turn to the results.

Panel A in Table 4 shows a uniform decline in negative provider flags. "Appointment was rushed," "Didn't explained condition well," "Didn't listen and answer questions," and "Didn't trust the provider's decision" each fall by about 1.0–1.3 percentage points (s.e. ≈0.2 pp), all $p < 0.001$, with ZIP, month, and matched-group fixed effects. Panel B in Table 4 shows

little and insignificant movement or small declines in positive provider flags. "Appointment wasn't rushed" declines by 2.7 pp (0.8 pp, $p < 0.001$) and "Trusted the provider's decision" declines by 2.4 pp (0.9 pp, $p < 0.01$); changes in "Explained condition well" and "Listened and answered questions" are small and imprecise. These results suggest that the post-BLM improvement primarily stems from fewer negative provider critiques rather than increased positive praise.

Table 4: Shifts in Patient Feedback on Providers

Panel A: Negative feedback on provider

	Appointment was rushed	Didn't explained condition well	Didn't listen and answer questions	Didn't trust the providers' decision
Post × Black	-0.013***	-0.013***	-0.010***	-0.012***
	(0.002)	(0.002)	(0.002)	(0.002)
Num. Obs.	1,505,219	1,505,219	1,505,219	1,505,219
R^2 Adj.	0.081	0.084	0.090	0.092
FE: zipcode	X	X	X	X
FE: month	X	X	X	X
FE: group	X	X	X	X

Panel B: Positive feedback on provider

	Appointment wasn't rushed	Explained condition well	Listened and answered questions	Trusted the provider's decision
Post × Black	-0.027***	-0.004	-0.005	-0.024**
	(0.008)	(0.006)	(0.006)	(0.009)
Num. Obs.	1,505,219	1,505,219	1,505,219	1,505,219
R^2 Adj.	0.326	0.228	0.259	0.363
FE: Matched-group	X	X	X	X
FE: Zip Code	X	X	X	X
FE: Month	X	X	X	X

Notes: Each column estimates a separate DiD where the outcome is an indicator for receiving a patient's feedback on the provider. Panel A includes the negative flags, showing a uniform decline in all negative provider flags. Panel B includes the positive flags, indicating a negative or insignificant effect. All models include ZIP, month, and matched-group fixed effects. Standard errors in parentheses. Errors are clustered at the county level. Significance levels: $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Panel A in Table 5 shows consistent declines in negative office plus staff flags. "Difficult to schedule appointment" falls by 1.6 pp (0.2 pp, $p < 0.001$). "Long wait times" falls by 1.8 pp (0.3 pp, $p < 0.001$). "Bad office environment" falls by 0.8 pp (0.1 pp, $p < 0.001$). "Staff

wasn't friendly" falls by 1.3 pp (0.2 pp, $p < 0.001$). Panel B in Table 5 shows negative or insignificant movements in positive office+staff flags. "Easy appointment scheduling" declines by 1.8 pp (0.9 pp, $p < 0.1$). "Short wait times" declines by 0.9 pp (0.9 pp, n.s.). "Good office environment" changes by 0.1 pp (0.8 pp, n.s.). "Staff was friendly" declines by 2.5 pp (0.9 pp, $p < 0.01$). As with provider feedback, the pattern is a decrease in all negative flags.

Table 5: Shifts in Patient Feedback on Office plus Staff

Panel A: Negative feedback on office+staff

	Difficult to schedule appointment	Long wait times	Bad office environment	Staff wasn't friendly
Post × Black	-0.016***	-0.018***	-0.008***	-0.013***
	(0.002)	(0.003)	(0.001)	(0.002)
Num. Obs.	1,505,219	1,505,219	1,505,219	1,505,219
R^2 Adj.	0.059	0.049	0.031	0.067
FE: zipcode	X	X	X	X
FE: month	X	X	X	X
FE: group	X	X	X	X

Panel B: Positive feedback on office+staff

	Easy to schedule appointment	Short wait times	Good office environment	Staff was friendly
Post × Black	-0.018*	-0.009	0.001	-0.025**
	(0.009)	(0.009)	(0.008)	(0.009)
Num. Obs.	1,505,219	1,505,219	1,505,219	1,505,219
R^2 Adj.	0.295	0.289	0.412	0.359
FE: Matched-group	X	X	X	X
FE: Zip Code	X	X	X	X
FE: Month	X	X	X	X

Notes: Each column estimates a separate DiD where the outcome is an indicator for receiving a patient's feedback on the office plus staff. Panel A includes the negative flags, showing a consistent decline in all negative provider flags. Panel B includes the positive flags, indicating a negative or insignificant effect. All models include ZIP, month, and matched-group fixed effects. Standard errors in parentheses. Errors are clustered at the county level. Significance levels: $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Across provider and office+staff dimensions, negative flags fall for Black providers, while positive flags change little and sometimes tilt toward White providers. Taken together with the star level pattern, fewer 1-star and more 5-star reviews, these results imply less punitive reviewing, not a broad rise in praise.

6 Heterogeneity in Main Effect

This section investigates whether the impact of the Black Lives Matter movement on patient evaluations of Black health providers differs across contexts. We explore three key dimensions of heterogeneity. First, we examine variation by provider sex to assess whether Black male and female providers experienced similar changes in patient ratings. Second, we consider protest-related factors at the county level, including protest type and intensity, to capture how local salience of the BLM movement may have shaped patient responses. Third, we analyze county-level socioeconomic characteristics, such as racial segregation, income, and education, to evaluate whether broader structural conditions moderated the observed effects. Together, these analyses allow us to determine whether the effects of the BLM movement were consistent across settings or contingent on local demographics and social context.

6.1 Heterogeneity by Provider's Sex

We first examine whether the impact of the BLM movement on patient evaluations of Black providers differs by provider sex. Table 6 presents the results. The estimates indicate that the main effect is concentrated among Black male providers, with a positive and statistically significant DiD term across all specifications. For Black female providers, however, the DiD term is small in magnitude and statistically insignificant. This suggests that the observed improvement in ratings following the BLM movement is driven by an increase in evaluations of Black male providers.

6.2 Heterogeneity by Protest Type and Intensity

We next investigate whether the effect of the BLM movement on patient evaluations of Black providers varied depending on the type and intensity of protests. To capture these dynamics, we draw on data from the Armed Conflict Location & Event Data (ACLED) and

Table 6: Impact of BLM Protests on Ratings of Black Providers by Provider's Sex

Panel A: Ratings for Female Providers

	Column (1)	Column (2)	Column (3)
Post × Black	0.032	0.040	0.023
	(0.038)	(0.037)	(0.048)
Num. Obs.	757,285	757,285	425,045
Adj. R^2	0.108	0.164	0.265
FE: Provider			X
FE: Matched-group		X	
FE: ZIP Code	X	X	X
FE: Month	X	X	X

Panel B: Ratings for Male Providers

Post × Black	0.059^{***}	0.063^{***}	$0.030^{\dagger}$
	(0.016)	(0.013)	(0.016)
Num. Obs.	1,080,174	1,080,174	1,080,174
Adj. R^2	0.063	0.118	0.236
FE: Provider			X
FE: Matched-group		X	
FE: ZIP Code	X	X	X
FE: Month	X	X	X

Notes: Panel A shows the estimates for the main effect among ratings for female providers, while Panel B shows it for male providers. The results show that significant effects are observed only for Black male providers, while estimates for Black female providers are not statistically significant. Standard errors in parentheses. Errors are clustered at the county level. Significance levels: $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

the Bridging Divides Initiative (BDI), which provide detailed information on protest events, including their type, timing, and location (Kishi and Jones, 2020). Within this framework, BLM-related events are classified into four categories: peaceful protests, peaceful protests with intervention, violent protests, and peaceful protests met with excessive force. Peaceful protests show no signs of violence. In peaceful protests with intervention, police or other authorities intervened and displayed violence. Violent protests are riots in which protesters or rioters themselves engaged in violent acts. In peaceful protests met with excessive force, external actors such as right-wing or armed groups acted against the protesters and exhibited

violence. We restrict our analysis to events occurring between the murder of George Floyd and the end of July 2020, the peak of protest activity.

We implement two approaches. First, we examine whether the presence of different protest types within a county moderated the treatment effect. Second, we incorporate the intensity of protest activity by interacting the treatment effect with the cumulative number of protest events. These complementary strategies allow us to test whether either the presence or the frequency of protests shaped the observed changes in evaluations of Black providers.

Table 7 shows the results for protest type. Across specifications, the three-way interaction terms are small and generally statistically insignificant, with the exception of a marginal negative estimate for peaceful protests in the most saturated model. This indicates that the main effect is not systematically different across protest categories.

Table 7: Heterogeneity by Protest Type

	Column (1)	Column (2)	Column (3)
Post × Black	0.033	0.046	0.114**
	(0.033)	(0.025)	(0.042)
Post × Black × Peaceful Protest	0.007	0.004	-0.098*
	(0.034)	(0.026)	(0.049)
Post × Black × Peaceful Protest w/ Intervention	-0.013	0.004	-0.032
	(0.025)	(0.021)	(0.038)
Post × Black × Violent Protest	0.040†	0.016	0.039
	(0.024)	(0.020)	(0.040)
Post × Black × Peaceful Protest w/ Excessive Force	-0.024	-0.028	-0.013
	(0.027)	(0.025)	(0.052)
Num. Obs.	1,505,219	1,505,219	1,505,219
Adj. R^2	0.059	0.115	0.248
FE: Provider			X
FE: Matched-group		X	
FE: ZIP Code	X	X	X
FE: Month	X	X	X

Notes: This table shows the heterogeneity of the main effect by protest types. The main effect is not systematically different across protest categories. Standard errors in parentheses. Errors are clustered at the county level. Significance levels: $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table 8 presents the results for protest intensity. Here, the treatment effect interacts with the cumulative number of events in each category within a county. None of the coefficients are statistically significant once additional fixed effects are introduced, and the magnitudes remain close to zero. This indicates that protest frequency did not condition the effect of the BLM movement on ratings of Black providers.

Table 8: Heterogeneity by Protest Intensity

	Column (1)	Column (2)	Column (3)
Post × Black	0.041*	0.053**	0.032†
	(0.016)	(0.019)	(0.019)
Post × Black × Count of Peaceful Protest	0.001*	-0.002	0.000
	(0.001)	(0.001)	(0.001)
Post × Black × Count of Peaceful Protest w/ Intervention	-0.001	0.006	-0.006
	(0.009)	(0.011)	(0.015)
Post × Black × Count of Violent Protest	-0.005	-0.005	0.004
	(0.004)	(0.008)	(0.007)
Post × Black × Count of Peaceful Protest w/ Excessive Force	-0.018	-0.016	-0.042
	(0.025)	(0.030)	(0.041)
Num. Obs.	1,505,219	1,505,219	1,505,219
Adj. R^2	0.059	0.115	0.248
FE: Provider			X
FE: Matched-group		X	
FE: ZIP Code	X	X	X
FE: Month	X	X	X

Notes: This table shows the heterogeneity of the main effect by protest intensity. The main effect is not systematically different across protest intensities. Standard errors in parentheses. Errors are clustered at the county level. Significance levels: $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

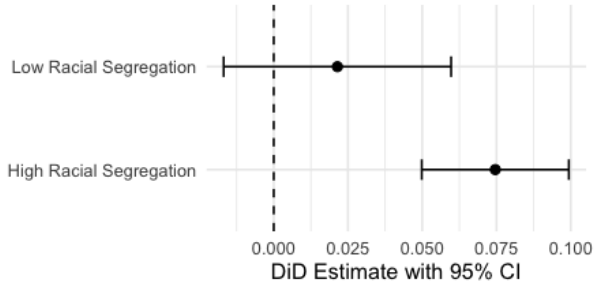
In summary, this subsection analyzed heterogeneity in the treatment effect across both protest type and protest intensity using detailed event data. Across specifications, we do not observe systematic variation in the effect of the BLM movement on ratings of Black providers based on local protest characteristics. A caveat to emphasize is that this does not mean protests themselves had no effect. Rather, it indicates that within our empirical framework, we cannot detect variation in the treatment effect attributable to the type or intensity of protest.

6.3 Heterogeneity by County-Level Socioeconomic Characteristics

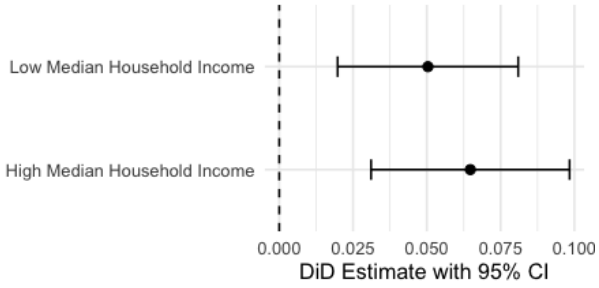
Beyond protest dynamics, the impact of the BLM movement may also vary across counties with different socioeconomic conditions. Structural factors such as racial segregation, racial composition, Urban-rural composition, income distribution, education attainment, and political leaning can influence both the salience of racial justice movements and the ways communities respond. By examining county-level socioeconomic characteristics, we capture these underlying structural conditions and test whether they moderate changes in patient evaluations of Black providers.

Figure 4 displays the estimated treatment effects when the sample is split at the median of each socioeconomic variable. For racial segregation, Panel (a) in Figure 4 shows a pronounced pattern indicating that counties with higher segregation exhibit large and statistically significant effects, while counties with lower segregation show no meaningful change. This suggests that the BLM movement's impact on patient evaluations was concentrated in communities with sharper racial divides. A similar, though weaker, pattern emerges for the share of the Black population in Panel (b) in Figure 4, showing counties with larger Black populations exhibit positive effects.

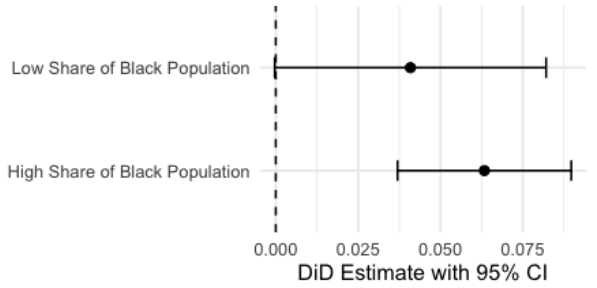
By contrast, for other socioeconomic measures—including urban block ratio, median household income, education attainment, and political leaning—the estimated effects are small and statistically indistinguishable across high and low counties. These null results indicate that, aside from segregation and racial composition, broader economic or political conditions did not systematically moderate the BLM movement's impact on patient evaluations of Black providers. Full regression tables are reported in the Appendix.



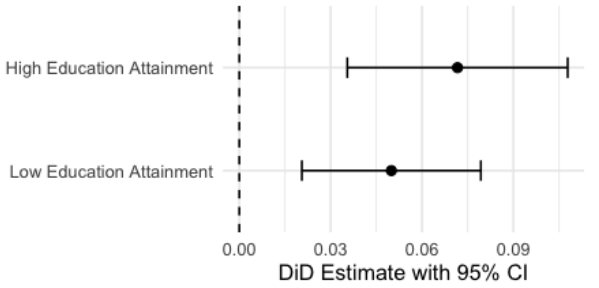
(a) Heterogeneity by Racial Segregation



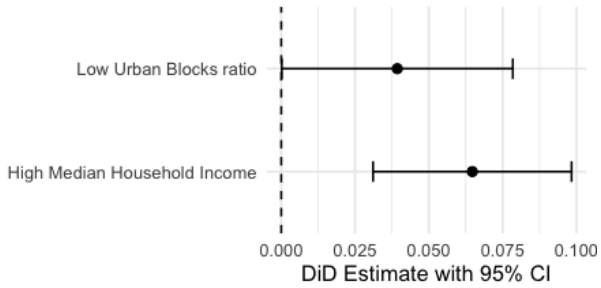
(d) Heterogeneity by Median Household Income



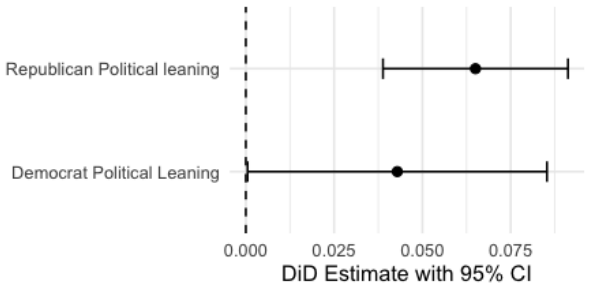
(b) Heterogeneity by Share of Black Population



(e) Heterogeneity by Education Attainment



c) Heterogeneity by Urban Blocks Ratio



(f) Heterogeneity by Political Leaning

Figure 4: DiD estimates with 95% confidence intervals across heterogeneity groups. Panels (a)–(c) show heterogeneity by segregation, Black population share, and urbanization. Panels (d)–(f) show heterogeneity by income, education, and political leaning. Points indicate DiD estimates with 95% confidence intervals, based on median splits of each county-level variable. Regression tables underlying these estimates are reported in the Appendix.

7 Robustness Check

7.1 Generalized Synthetic Control

We re-estimate the effect using the Generalized Synthetic Control (GSC) method, which models outcomes with interactive fixed effects (latent factors) to relax the parallel-trends

assumption and extend synthetic control to settings with multiple/staggered treatments. GSC imputes counterfactuals for treated units by combining observed covariates with a low-rank factor structure estimated from the donor pool, and it selects the number of factors by cross-validation; uncertainty is typically obtained via bootstrap (Xu, 2017).

We apply the gsynth package in R to the provider–month panel, including zip code as an observed covariance, and impose two-way fixed effects. As reported in Table 9, the average treatment effect on the treated (ATT) is 0.066 with a standard error of 0.0205 ($p = 0.0013$), yielding a 95% confidence interval of [0.026, 0.106]. Substantively, ratings for Black providers increased by about 0.07 stars post-BLM relative to their matched White counterparts. This aligns with, and is slightly larger than, the DiD estimates reported in Table 2. Pre-treatment coefficients fluctuate around zero without a coherent trend. While a few isolated months show nominal significance, they do not form a coherent pre-trend (e.g., months $t = -48$ to $t = -45$ exhibit negative values, and $t = -44$ is positive). During the post-period, several months exhibit positive and statistically significant effects (e.g., $t = 2, 3, 4, 10, 13, 15, 17$), reinforcing a sustained average uplift.

Table 9: GSC robustness: Average treatment effect on the treated (ATT)

	ATT	Std. Error	95% CI	p-value
Average ATT	0.066	0.0205	[0.026, 0.106]	0.0013

Notes: Estimates from GSC method on the provider–month panel with Zip code control and two-way fixed effects. Number of latent factors selected by cross-validation over $r \in [0, 5]$. Uncertainty from 2,000 parametric bootstrap replications with `min.T0` = 7.

7.2 Placebo Test

As a robustness check, we conducted a placebo test by redefining the intervention period to May 2019, one year before the actual shock. We restricted the sample to ratings from 2016 up to the true intervention date in May 2020 and re-estimated our difference-in-differences

specification. If the observed effects in our main analysis are indeed driven by the Black Lives Matter movement in 2020 rather than pre-existing dynamics, we should not observe any systematic effect in this placebo setting.

We defined a placebo post indicator equal to one for all reviews after May 2019 and interacted it with the treatment indicator for Black providers. The model was estimated three times: (1) without any fixed effects, (2) including matched group, zipcode, and month fixed effects, and (3) including provider, zipcode, and month fixed effects. Standard errors were clustered at the county level.

Table 10: Placebo Test: Redefining Intervention in May 2019

	Column (1)	Column (2)	Column (3)
Post × Black	-0.020	-0.003	-0.028
	(0.029)	(0.018)	(0.019)
Num. Obs.	581,426	581,426	581,426
Adj. R^2	0.003	0.109	0.256
FE: Provider			X
FE: Matched-group		X	
FE: ZIP Code	X	X	X
FE: Month	X	X	X

Notes: DiD estimates for the placebo test. Placebo test supports that the main effect is not an artifact of spurious pre-trends. Standard errors in parentheses. Errors are clustered at the county level. Significance levels: $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Across all three specifications, the coefficient on the placebo interaction is small and statistically insignificant. For example, in our most saturated specification with provider, location, and time fixed effects, the coefficient is -0.028 with a standard error of 0.019. This reinforces that the improvement in ratings for Black providers documented in our main analysis is not an artifact of spurious pre-trends, but instead reflects the causal impact of the BLM movement in 2020.

7.3 Restricting to Pre-BLM Providers

To further test the robustness of our results, we restrict the sample to providers whose first review was recorded before the BLM shock. This restriction ensures that our estimates are not confounded by providers who entered the review system only after the event, which could introduce selection bias. As shown in Table 11, the estimated treatment effects remain positive and close in magnitude to our baseline results. Specifically, across different fixed effects specifications, the post-treatment effect ranges between 0.030 and 0.058 stars, all statistically significant at conventional levels. This consistency confirms that the observed impact is not driven by changes in the composition of providers entering the market after the BLM movement, but instead reflects a genuine shift in patient evaluations of Black providers relative to white counterparts.

Table 11: DiD Estimates for Pre-BLM Providers

	Column (1)	Column (2)	Column (3)
Post × Black	0.046^{**}	0.058^{***}	0.030^{*}
	(0.015)	(0.012)	(0.015)
Num. Obs.	1,266,633	1,266,633	1,266,633
Adj. R^2	0.059	0.115	0.233
FE: Provider			X
FE: Matched-group		X	
FE: ZIP Code	X	X	X
FE: Month	X	X	X

Notes: DiD estimates are calculated when restricting to only providers with a first review before the BLM shock. This table reflects a genuine shift in patient evaluations of Black providers relative to white counterparts. Standard errors in parentheses. Errors are clustered at the county level. Significance levels: $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

7.4 Restricting to Pre-AI Adoption Era

As an additional robustness check, we restrict the analysis to the pre-AI adoption era, thereby excluding reviews that may have been generated or influenced by large language models (LLMs) or other AI-based systems. Specifically, we trim the data by June 31, 2023, which is recognized as the period when AI-generated reviews began rising substantially across online review platforms. This restriction addresses concerns that the rise of automated review generation could bias the results. The estimates presented in Table 12 demonstrate that the effect remains positive and statistically significant across all specifications. The magnitude of the effect, ranging between 0.047 and 0.055 stars, is highly consistent with our main results. This indicates that the observed treatment effect is not an artifact of AI-generated content, but rather reflects genuine shifts in patient evaluations of Black providers relative to white providers.

Table 12: DiD Estimates for Pre-AI Adoption Era

	Column (1)	Column (2)	Column (3)
Post × Black	0.050**	0.055***	0.047*
	(0.015)	(0.013)	(0.016)
Num. Obs.	1,075,699	1,075,699	1,075,699
Adj. R^2	0.057	0.111	0.247
FE: Provider			X
FE: Matched-group		X	
FE: ZIP Code	X	X	X
FE: Month	X	X	X

Notes: DiD estimates are calculated when restricting to the pre-AI adoption era. This table shows that the main effect exists even before the AI-generated reviews started to rise. Standard errors in parentheses. Errors are clustered at the county level. Significance levels: $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

7.5 Continuous Specification Using Probability of Being Black

In our final robustness check, we replace the binary treatment indicator (Black provider) with a continuous measure: the predicted probability of being Black from the DeepFace model. This probability score ranges from 0 to 1 and takes values above 0.5 for providers identified as Black. By using this continuous variable, the treatment effect is allowed to scale with the degree of confidence in racial classification rather than relying solely on a binary cutoff. Conceptually, this specification captures how patient ratings shift as the perceived likelihood of a provider being Black increases.

Table 13 reports the estimates across three specifications. The results are consistent with our main findings, with effect sizes between 0.050 and 0.059 stars. The significance remains strong when including matched-group and location-time fixed effects, while the specification with provider fixed effects shows a weaker level of confidence, likely due to limited within-provider variation in the DeepFace probability score. Overall, this robustness exercise confirms that our main results are not sensitive to the binary coding of race and hold when we allow for a continuous representation of racial classification.

Table 13: DiD Estimates for the Probability of Being Black

	Column (1)	Column (2)	Column (3)
Post × Probability(Black)	0.054^{***}	0.059^{***}	$0.050^{\dagger}$
	(0.015)	(0.012)	(0.018)
Num. Obs.	1,505,219	1,505,219	1,505,219
Adj. R^2	0.059	0.115	0.248
FE: Provider			X
FE: Matched-group		X	
FE: ZIP Code	X	X	X
FE: Month	X	X	X

Notes: DiD estimates are calculated with a continuous specification using the probability of being Black. This table suggests that our main effect is not sensitive to the binary coding of race. Standard errors in parentheses. Errors are clustered at the county level. Significance levels: $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

8 Conclusion

This paper examined how the Black Lives Matter (BLM) movement, following the murder of George Floyd in 2020, shaped patient evaluations of Black healthcare providers. Using reviews from Healthgrades.com and a difference-in-differences design with matched Black–White provider groups, we found that Black providers experienced an average increase of 0.05 stars compared to their White counterparts. This improvement was driven by a decrease in 1-star reviews, an increase in 5-star reviews, and a significant reduction in negative written feedback. The effect was concentrated among Black male providers, absent for Black female providers, stronger in racially segregated counties, and persistent well beyond the immediate aftermath of the movement. Measures of protest intensity did not explain variation in the impact.

These findings demonstrate that large-scale social movements can have a lasting impact on professional service markets, influencing consumer perceptions of minority providers in a durable manner. For policymakers and healthcare leaders, the results underscore both the persistence of racial discrimination in evaluations and the potential for societal mobilization to mitigate, though not eliminate, such biases. For online review platforms, they reveal that ratings are sensitive to broader cultural currents, raising concerns about their stability as indicators of provider quality. More broadly, the evidence highlights how collective action can reshape evaluations in domains with entrenched inequities, offering new insight into the social and institutional consequences of protest movements.

Our study contributes to the growing body of literature on the impact of social movements on economic and social outcomes, with a particular focus on digital platforms. We extend prior work by showing that BLM altered consumer behavior in the healthcare sector, a domain where racial disparities are both deeply entrenched and socially consequential. By linking a major social movement to changes in patient evaluations of clinicians, we highlight how shifts in societal attitudes can influence perceptions of minority professionals in critical

service markets.

References

Acien, A., Morales, A., Vera-Rodriguez, R., Bartolome, I., and Fierrez, J. (2019). Measuring the gender and ethnicity bias in deep models for face recognition. In Vera-Rodriguez, R., Fierrez, J., and Morales, A., editors, *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pages 584–593, Cham. Springer International Publishing.

Agarwal, S. and Sen, A. (2022). Antiracist curriculum and digital platforms: Evidence from black lives matter. *Management Science*, 68(4):2932–2948.

Alsan, M., Garrick, O., and Graziani, G. (2019). Does diversity matter for health? experimental evidence from oakland. *American Economic Review*, 109(12):4071–4111.

Aneja, A., Luca, M., and Reshef, O. (2025). The benefits of revealing race: Evidence from minority-owned local businesses. *American Economic Review*, 115(2):660–89.

Ayres, I., Banaji, M., and Jolls, C. (2015). Race effects on ebay. *RAND Journal of Economics*, 46(4):891–917.

Babar, Y., Mahdavi Adeli, A., and Burtch, G. (2023). The effects of online social identity signals on retailer demand. *Management Science*, 69(12):7335–7346.

Campbell, M. (2021). Protests and police violence: Evidence from black lives matter. *American Political Science Review*, 115(4):1117–1135.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Deng, J., Guo, J., Xue, N., and Zafeiriou, S. (2019). Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision*

and Pattern Recognition (CVPR).

Edelman, B., Luca, M., and Svirsky, D. (2017). Racial discrimination in the sharing economy: Evidence from a field experiment. *American Economic Journal: Applied Economics*, 9(2):1–22.

Fairlie, R. W. and Robb, A. M. (2007). Why are black-owned businesses less successful than white-owned businesses? the role of families, inheritances, and business human capital. *Journal of Labor Economics*, 25(2):289–323.

Fairlie, R. W., Robb, A. M., and Robinson, D. T. (2022). Black and white: Access to capital among minority-owned start-ups. *Management Science*, 68(4):2377–2400.

Gligor, D., Newman, C., and Kashmiri, S. (2021a). Does your skin color matter in buyer–seller negotiations? the implications of being a black salesperson. *Journal of the Academy of Marketing Science*, 49:969–993.

Gligor, D. M., Novicevic, M., Feizabadi, J., and Stapleton, A. (2021b). Examining investor reactions to appointments of black top management executives and ceos. *Strategic Management Journal*, 42(10):1939–1959.

Greco, A., Percannella, G., Vento, M., et al. (2020). Benchmarking deep network architectures for ethnicity recognition using a new large face dataset. *Machine Vision and Applications*, 31:67.

Green, C. R., Anderson, K. O., Baker, T. A., Campbell, L. C., Decker, S., Fillingim, R. B., Kaloukalani, D. A., Lasch, K. E., Myers, C., Tait, R. C., Todd, K. H., and Vallerand, A. H. (2003). The unequal burden of pain: Confronting racial and ethnic disparities in pain. *Pain Medicine*, 4(3):277–294.

Hannák, A., Wagner, C., Garcia, D., Mislove, A., Strohmaier, M., and Wilson, C. (2017). Bias in online freelance marketplaces: Evidence from taskrabbit and fiverr.

Hoffman, K. M., Trawalter, S., Axt, J. R., and Oliver, M. N. (2016). Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences

between blacks and whites. *Proceedings of the National Academy of Sciences*, 113(16):4296–4301.

Kishi, R. and Jones, S. (2020). Demonstrations and political violence in america: New data for summer 2020. Accessed insert date you accessed the page .

Levy, R. and Mattsson, M. (2023). The effects of social movements: Evidence from #metoo. SSRN Electronic Journal.

Lin, Y.-W., Yang, S., Han, W., and Lu, J. G. (2024). The black lives matter movement mitigates bias against racial minority actors. *Proceedings of the National Academy of Sciences*, 121(29):e2307726121.

Liu, X., Han, M., Nicolau, J. L., and Li, C. (2022). Online engagement and persistent reactions to social causes: The black-owned business attribute. *Tourism Management*, 88:104407.

Luca, M., Pronkina, E., and Rossi, M. (2024). The evolution of discrimination in online markets: How the rise in anti-asian bias affected airbnb during the pandemic. *Marketing Science*, 0(0).

Luo, H. and Zhang, L. (2021). Scandal, social movement, and change: Evidence from #metoo in hollywood. *Management Science*, 68(2):1278–1296.

Madestam, A., Shoag, D., Veuger, S., and Yanagizawa-Drott, D. (2013). Do political protests matter? evidence from the tea party movement. *The Quarterly Journal of Economics*, 128(4):1633–1685.

Marx, M., Wang, Q., and Yimfor, E. (2025). Minimum viable signal: Venture funding, social movements, and race. *Management Science*.

Nunez-Smith, M., Pilgrim, N., Wynia, M., Desai, M. M., Bright, C., Krumholz, H. M., and Bradley, E. H. (2009a). Health care workplace discrimination and physician turnover. *Journal of the National Medical Association*, 101(12):1274–1282.

Nunez-Smith, M., Pilgrim, N., Wynia, M., Desai, M. M., Jones, B. A., Bright, C., Krumholz,

H. M., and Bradley, E. H. (2009b). Race/ethnicity and workplace discrimination: results of a national survey of physicians. *Journal of General Internal Medicine*, 24(11):1198–1204.

Parkhi, O., Vedaldi, A., and Zisserman, A. (2015). Deep face recognition. In *BMVC 2015 - Proceedings of the British Machine Vision Conference 2015*, pages 1–12. British Machine Vision Association.

Saha, S., Komaromy, M., Koepsell, T. D., and Bindman, A. B. (1999). Patient-physician racial concordance and the perceived quality and use of health care. *Archives of Internal Medicine*, 159(9):997–1004.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing Management*, 24(5):513–523.

Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Serengil, S. I. and Ozpinar, A. (2020). Lightface: A hybrid deep face recognition framework. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 23–27. IEEE.

Smedley, B., Stith, A., and Nelson, A. (2003). *Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care*. National Academies Press.

Son, Y., Wowak, K. D., and Angst, C. M. (2025). Does greater visibility benefit minority businesses? evidence from an online review platform. *Production and Operations Management*, 34(4):711–724.

Sun, Y., Chen, Y., Wang, X., and Tang, X. (2014). Deep learning face representation by joint identification-verification. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Taigman, Y., Yang, M., Ranzato, M., and Wolf, L. (2014). Deepface: Closing the gap to

human-level performance in face verification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708.

Traylor, A. H., Schmittdiel, J. A., Uratsu, C. S., Mangione, C. M., and Subramanian, U. (2010). Adherence to cardiovascular disease medications: does patient-provider race/ethnicity and language concordance matter? *Journal of General Internal Medicine*, 25(11):1172–1177.

Wang, Y. and Kosinski, M. (2018). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology*, 114(2):246–257.

Williams, D. R. and Rucker, T. D. (2000). Understanding and addressing racial disparities in health care. *Health Care Financing Review*, 21(4):75–90.

Xu, Y. (2017). Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis*, 25(1):57–76.

Yazdani, E., Chakravarty, A., and Inman, J. (2025). Racial inequity in donation-based crowdfunding platforms: The role of facial emotional expressiveness. *Journal of Marketing*, 89(4):140–160.

Younkin, P. and Kuppuswamy, V. (2018). The colorblind crowd? founder race and performance in crowdfunding. *Management Science*, 64(7):3269–3287.

Younkin, P. and Kuppuswamy, V. (2019). Discounted: The effect of founder race on the price of new products. *Journal of Business Venturing*, 34(2):389–412.

A Length and Persistence of the Main Effect

To assess whether the BLM movement's impact was temporary or enduring, we conduct additional persistence tests. Specifically, we re-estimate our main difference-in-differences specification on trimmed subsets of the matched sample that exclude the first 3, 6, 12, and 24 months following the May 2020 shock. All specifications retain matched-group, ZIP code,

and month fixed effects, with clustering as in the main analysis. This design directly asks whether the effect remains once the immediate post-BLM surge is removed.

Table 14 presents the results. The baseline ATT on the full matched sample is 0.057 (s.e. 0.011). Excluding the first 3 months yields 0.056 (0.012), excluding 6 months yields 0.057 (0.012), excluding 12 months yields 0.059 (0.012), and excluding 24 months yields 0.062 (0.012). All estimates are statistically significant at $p < 0.001$. The effect is stable across specifications and even slightly increases as more early months are trimmed.

Table 14: Persistence: Trimming Early Post-BLM Months

	Complete Matched Sample	Trimmed by 3 months post-BLM	Trimmed by first 6 months post-BLM	Trimmed by first 12 months post-BLM	Trimmed by first 24 months post-BLM
Post × Treated	0.057***	0.056***	0.057***	0.059***	0.062***
	(0.011)	(0.012)	(0.012)	(0.012)	(0.012)
Num. Obs.	1,505,219	1,469,832	1,435,890	1,356,647	1,208,697
R^2 Adj.	0.115	0.116	0.116	0.118	0.118
FE: group	X	X	X	X	X
FE: zipcode	X	X	X	X	X
FE: month	X	X	X	X	X

Notes: Coefficients are DiD estimates for Black vs. matched White providers. Each column trims the early post-BLM window as labeled. Standard errors in parentheses. All models include matched-group, ZIP, and month fixed effects. Significance levels: $^{\cdot}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

These results indicate that the observed shift in ratings of Black providers is durable rather than a short-lived enthusiasm effect. The slight upward trend with longer trimming is consistent with reinforcement or diffusion of the effect over time, not decay. This pattern aligns with the persistence analysis reported in the main text and visualized in Figure 3, confirming that the BLM movement produced lasting changes in patient evaluations.

In addition to trimming, we estimate quarterly dynamic effects to examine how the impact evolves over time. Figure 5 plots the DiD coefficients quarter by quarter after May 2020, each with a 95% confidence interval. The estimates fluctuate somewhat across quarters, but they consistently remain positive, with several quarters showing statistically significant effects.

Importantly, there is no evidence of a rapid fade-out: even several years after the initial protests, point estimates remain above zero and comparable in magnitude to the baseline ATT.

This dynamic analysis complements the trimming results by showing that the effect is not confined to the immediate aftermath but persists across multiple quarters. Together, Table 14 and Figure 5 provide consistent evidence that the BLM movement induced a lasting upward shift in patient evaluations of Black providers.

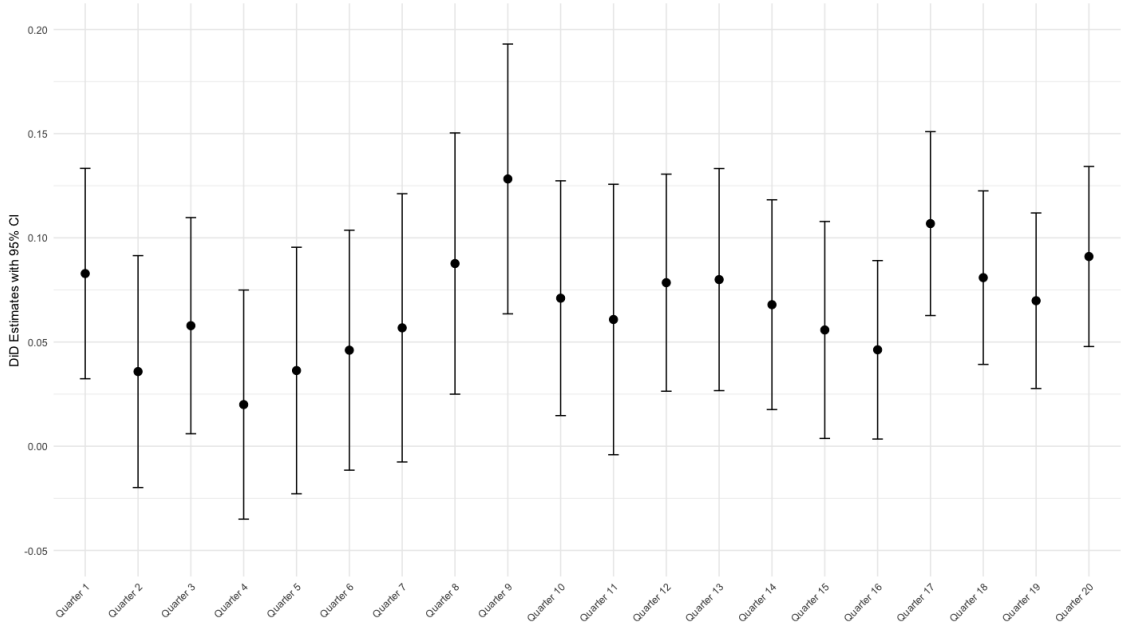


Figure 5: Dynamic Quarterly Effects of BLM on Ratings of Black vs. White Providers. Each point represents the DiD estimate for the corresponding quarter after May 2020, with 95% confidence intervals.

B Regression Tables for Heterogeneity Analysis by Socioeconomic Moderators

Table 15: Heterogeneity by Racial Segregation

	Column (1)	Column (2)	Column (3)
Panel A: Ratings with Racial Segregation Higher than Median			
Post × Black	0.071***	0.075***	0.065***
	(0.017)	(0.013)	(0.017)
Num. Obs.	756,289	756,289	756,289
Adj. R^2	0.058	0.127	0.2443
FE: Provider			X
FE: Matched-group		X	
FE: ZIP Code	X	X	X
FE: Month	X	X	X
Panel B: Ratings with Racial Segregation Lower than Median			
Post × Black	0.029	0.021	-0.011
	(0.025)	(0.020)	(0.024)
Num. Obs.	748,247	748,247	748,247
Adj. R^2	0.060	0.132	0.253
FE: Provider			X
FE: Matched-group		X	
FE: ZIP Code	X	X	X
FE: Month	X	X	X

Notes: Standard errors in parentheses. Errors are clustered at the county level. Significance levels: $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table 16: Heterogeneity by Racial Segregation

Panel A: Ratings with Share of Black Higher Higher than Median

	Column (1)	Column (2)	Column (3)
Post × Black	0.059***	0.063***	0.018
	(0.016)	(0.013)	(0.019)
Num. Obs.	757,285	757,285	757,285
Adj. R^2	0.057	0.125	0.247
FE: Provider			X
FE: Matched-group		X	
FE: ZIP Code	X	X	X
FE: Month	X	X	X

Panel B: Ratings with Share of Black Higher Lower than Median

Post × Black	0.023	0.041†	0.037
	(0.026)	(0.021)	(0.025)
Num. Obs.	747,929	747,929	747,929
Adj. R^2	0.061	0.132	0.249
FE: Provider			X
FE: Matched-group		X	
FE: ZIP Code	X	X	X
FE: Month	X	X	X

Notes: Standard errors in parentheses. Errors are clustered at the county level. Significance levels: $^{\dagger}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.