Monthly filled posts tracker - methodology

Statistics in development

May 2026

Contact us: analysis@skillsforcare.org.uk

Contents

Overview

Introduction

This publication describes the methods used by Skills for Care (SfC) to produce the estimate of the number of filled posts in CQC-regulated locations in the adult social care independent sector in England over time.

These estimates are published

- Monthly in our filled post tracker
- Annually as part of our 'Size and Structure' and 'State of' publications

The statistics are currently labelled as **statistics in development**, reflecting ongoing improvements to methods and data sources.

Users should continue to use our annual 'Size and Structure' and 'State of' publications for official purposes.

Background

Prior to November 2024, estimates of filled posts were produced annually using a largely manual process.

Between annual publications, trends were monitored using ASC-WDS updates. This approach relied on a small subset of providers and did not adjust for the representativeness of the sample.

A data engineering pipeline was developed by Skills for Care to automate production, improve timeliness and consistency, make fuller use of historical ASC-WDS data, and provide monthly estimates that are more responsive to changes in the sector. The methodology used is described in this document.

Data sources and coverage

Data sources

The estimates are derived from four main sources:

- CQC Care Directory (locations and regulatory data)
- ASC-WDS (filled posts)
- CQC PIR (people directly employed)
- ONS Postcode Directory (geography)

Coverage

The analysis focuses on CQC-regulated locations in the independent sector.

Coverage of the ASC-WDS can be calculated in multiple ways depending on the context:

- Around 55% of CQC-regulated locations have engaged with ASC-WDS within the last two years.
- However, as ASC-WDS is collected as a snapshot, the proportion of locations with an updated submission in any given month is lower (typically around 10–15%).
- The methodology addresses this by using imputation to extend observed data across time, meaning that the majority of estimates (around 70–80%) are informed by real submissions, either directly or indirectly.

Coverage is monitored monthly as a key performance indicator.

1 Methodology overview

The pipeline for estimating the number of filled posts consists of the following stages:

- **Ingestion**
- **Merging**
- **Cleaning**
- **Imputation** (stage one)
- **Modelling**
- **Estimation**
- **Diagnostics**

The approach combines multiple data sources to produce a complete time series of filled post estimates at location level. This ensures that estimates are always based on the most reliable available data.

1.1 Ingestion

Data are collected from multiple sources on a regular schedule and prepared for use in the pipeline.

1.1.1 Adult Social Care Workforce Data Set (ASC-WDS)

- Method: Automatic.
- How collected: CSV file deposited in cloud storage by in-house developers.
- Frequency: 1st, 8th, 15th and 23rd of the month.
- Data collected since: 2006.
- Notes: prior to 2019 it was known as the National Minimum Data Set for Social Care (NMDS-SC) service.
- Contains: Workplace and worker level data, including how many filled posts each workplace employed and which job roles they are in. Workplaces add their Care Quality Commission (CQC) location ID which is used to link the datasets together.

1.1.2 Office for National Statistics Postcode Directory (ONSPD)

- Method: Manual.
- How collected: CSV file downloaded from ONS webpage. The file is converted to a CSV file and uploaded to cloud storage.
- Frequency: Annual (May publication).
- Data collected since: May 2012.
- Contains: All postcodes in the United Kingdom.

1.1.3 Care Quality Commission (CQC) Care Directory

- Method: Automatic.
- How collected: API call.
- Frequency: 1st, 8th, 15th and 23rd of the month.
- Data collected since: March 2013.
- Notes: From March 2013 to February 2022 data was collected manually by downloading an Excel file from CQC's website each month. Since March 2022 the data has been collected via the API call.

The Care Directory is filtered to only contain locations who are:

- Currently registered.
- Identified as adult social care.
- Not specifically 'specialist colleges'.

The ONS postcode directory is joined into the Care Directory via a postcode matching process.

1.1.4 CQC Provider Information Return (PIR)

- Method: Manual.
- How collected: Excel file downloaded from a secure webpage. The file is converted to a CSV file and uploaded to cloud storage.
- Frequency: Monthly.
- Data collected since: October 2019.
- Contains: CQC location ID which is used to link the datasets together.

1.2 Merging

The CQC Care Directory, containing all registered adult social care locations at each point in time, is used as the primary dataset. ASC-WDS and CQC PIR data are joined in to create a unified location-level dataset.

The local authority sector workplaces are removed from this pipeline, retaining only independent sector workplaces. Skills for Care conducts a separate annual data return of local authority adult social services departments.

1.3 Cleaning

Further detail and worked examples of these cleaning steps are provided in chapter 4.1.

1.3.1 Data preparation

A structured data preparation stage is applied to the merged location-level dataset before imputation and modelling. This stage standardises the monthly source data, resolves known structural issues and derives the variables required for later estimation steps.

Key processes include:

- reducing multiple files within the same month to a single monthly snapshot.
- deriving operational variables such as time registered and time since dormant.
- removing dual-registered care home records to avoid double counting.
- imputing missing care home bed count figures.
- deriving ASC-WDS filled posts from submitted workplace and worker-level information.
- removing repeated consecutive workforce values so that only meaningful changes are retained.
- creating derived ratios and bed-size bands used in later quality checks and modelling.

This stage ensures that the core dataset is internally consistent and ready for downstream estimation.

1.3.2 Outlier detection and anomaly treatment

Additional cleaning rules are applied to identify ASC-WDS workforce values that are unlikely to represent a credible location-level staffing figure.

These checks include:

- invalid placeholder values.
- provider-level submissions recorded against a single location.
- unusually high or low staffing-to-bed ratios for care homes.

Where values appear to represent invalid or misallocated data, they are removed from the estimation process.

For care homes, extreme staffing-to-bed ratios are Winsorized (capped) to a more plausible value rather than removed entirely. Validation work showed that unusually large values were often linked to genuinely larger services, while unusually small values were more common

among smaller services. Winsorizing therefore preserves useful information about relative service size while reducing the risk that extreme values distort later imputation and modelling.

1.4 Imputation (stage one)

Imputation is used to estimate filled posts for periods where no workforce data is available. This is necessary because:

- ASC-WDS submissions are voluntary and irregular.
- CQC PIR data is typically collected annually.

As a result, there are gaps in the time series for most locations. The imputation method is based on estimating how workforce levels change over time and applying this to known data points.

See Appendix - chapter 4.2 for worked examples of imputation steps.

1.4.1 Estimating change over time

A monthly rate of change trend in workforce change is calculated using observed ASC-WDS data from locations with multiple submissions. This trend is then used to interpolate between known values and extrapolate forwards and backwards.

This approach ensures that imputed values follow observed patterns in the data and remain consistent with known values.

To ensure robustness:

- Only locations with sufficient data are included.
- Extreme or atypical changes are excluded.
- Rates are smoothed over time to reduce volatility.

The resulting trend represents the typical change in workforce levels across the sector.

1.4.2 Application of the trend

The estimated trend is applied to known values to generate a continuous time series.

- For periods between known values - values are interpolated in a way that ensures alignment with both endpoints.
- For periods before or after known values - values are extrapolated based on the estimated trend.

For those locations who have submitted ASC-WDS data, this produces a complete set of workforce estimates over time.

1.4.3 Use of multiple data sources

For non-residential locations, where ASC-WDS data is unavailable or outdated, CQC PIR data is used to supplement the time series. ASC-WDS data is prioritised where both sources are available.

1.5 Modelling

Regression models are used to estimate filled posts for locations where no workforce data is available.

Separate models are used for care homes and non-residential services. For care homes, the model estimates a filled posts per bed ratio, which is then converted to filled posts using the number of beds. For non-residential services, the model directly estimates filled posts.

Models are applied only where observed or imputed values are not available.

1.5.1 Model inputs

The models use a range of characteristics to describe each location. These include workplace characteristics from the CQC Care Directory, geographical areas from the ONSPD and trend analysis from the ASC-WDS. A full list of features is provided in Appendix - chapter 4.3.3.

1.5.2 Model training and testing

The models are trained using historical data where filled posts are known.

- 80% of locations are used to train the model.
- 20% are used to test how well the model performs.

This allows us to check how accurate the model is on data it has not seen before.

Model performance is measured using the following metrics:

- R-squared (how well the model explains variation in the data).
- Proportion of predictions within 10 and 25 filled posts of the know values.

1.5.3 Model updates

The models are retrained each time the pipeline is run. This ensures that the latest available data is used and the models reflect current patterns in the sector.

Each model run is versioned, with metadata recorded on model performance, model settings and the features used for training.

1.6 Imputation (stage two)

Imputation is applied at two stages of the pipeline.

The previously described imputation (see chapter 1.4) was used to extend observed ASC-WDS data using a national trend-based approach. This increased the volume of usable data and provided a consistent time series for model training.

Following modelling, a second imputation is applied using model-based predictions. This produces more location-specific estimates by reflecting how workforce levels are expected to change over time for each individual location based on their features.

This two-stage approach ensures that models are trained on a sufficiently large dataset, while final estimates reflect location specific characteristics rather than just national averages.

1.7 Estimation

Final estimates are produced using a hierarchical approach:

- Observed values are used where available - approximately 10-15% per month.
- Imputed values are used where sufficient historical data exists - approximately 65-70% per month.
- Modelled values are used where no submitted data is available - approximately 20-25% per month.

This ensures that estimates are always based on the most reliable available information, with modelled values used only where necessary.

1.8 Diagnostics / performance metrics

Uncertainty varies by estimation method. Observed values are considered the most reliable, while imputed and modelled values are subject to greater uncertainty.

To quantify this, validation is carried out by comparing estimated values to observed submissions, allowing typical error ranges to be calculated separately for imputed and modelled estimates.

As estimates are produced using a hierarchical approach, overall uncertainty reflects the mix of methods used rather than a single confidence interval. Analysis to quantify uncertainty ranges for each estimation method will be incorporated into a future update of this methodology document.

2 Strengths and limitations

2.1 Strengths

Use of high-quality workforce data

Estimates are primarily based on ASC-WDS submitted workforce data, which provides detailed, location-level information on filled posts. This ensures that a substantial proportion of estimates are grounded in observed data. For ASC-WDS data to be used, the workplace-level staff figure must be similar to the number of individual worker-level records submitted which provides a higher level of confidence in these values.

Integration of multiple data sources

The methodology combines ASC-WDS with CQC and PIR data to improve coverage and consistency. This allows gaps in one source to be supplemented by others resulting in a more complete, and robust, dataset.

Comprehensive coverage through a hierarchical approach

A structured hierarchy is used to produce estimates for all locations and time periods, prioritising observed data, followed by imputed values, and using modelling only where necessary. This ensures full coverage while maintaining data quality.

Timely and regular production

Estimates are produced monthly, allowing workforce trends to be monitored in a timely way while maintaining consistency with underlying data sources.

Strong performance for care homes

The methodology performs particularly well for care homes, where workforce size is closely related to observable characteristics such as bed capacity. This results in high-quality estimates at both location and aggregate levels.

Alignment with external benchmarks

Aggregate estimates show good alignment with external data sources, providing assurance that the methodology produces realistic and credible outputs.

2.2 Limitations

Incomplete coverage in ASC-WDS

ASC-WDS is not a complete census of all CQC-regulated locations. Around 55% of locations have engaged with the service within the last two years, while the proportion submitting data in any given month is lower (typically around 10–15%).

This is addressed by using imputation methods to extend observed data across time, meaning that the majority of estimates are informed by real submissions either directly or indirectly.

Reliance on modelling and assumptions

Where no workforce data is available, estimates are derived using regression models and trend-based imputation. These approaches introduce uncertainty as they rely on statistical relationships rather than direct observation.

To minimise this, modelling is used only where necessary, following a hierarchical approach that prioritises observed and imputed values. Model performance is regularly evaluated using standard metrics, and models are retrained using the latest available data.

Lower accuracy for non-residential services at location level

Model performance is weaker for non-residential services compared to care homes, particularly at individual location level, due to greater variation in service types and workforce structures. This is mitigated by including a wide range of location characteristics in the models.

The final estimates are only used at aggregate level, where estimates are shown are more stable and show good alignment against external benchmarks.

Sensitivity to sudden changes in workforce levels

Estimates may be less accurate where workforce levels change sharply between reporting periods, for example due to service closures, mergers or rapid expansion.

The methodology reduces the impact of such cases by excluding extreme changes when estimating trends and applying smoothing techniques. These events are retained in the underlying data but are prevented from disproportionately influencing wider estimates.

3 Quality, governance and development

3.1 Compliance with the Code of Practice for Statistics

These statistics are produced in line with the principles of the Code of Practice for Statistics, which promotes trustworthiness, quality and value in official statistics.

Skills for Care uses workforce intelligence from the Adult Social Care Workforce Data Set (ASC-WDS), together with additional administrative and regulatory data sources, to produce evidence-based insights on the adult social care workforce.

The estimates are currently published as statistics in development, reflecting ongoing improvements to methods, data sources and quality assurance processes.

Statistical outputs are subject to internal quality assurance and sign-off processes prior to publication. Methodological changes, revisions and known limitations are documented transparently to support appropriate interpretation and use of the statistics.

Further information on Skills for Care's policies and procedures relating to the Code of Practice for Statistics is available on the Skills for Care website.

3.2 Revisions

The methodologies outlined in this paper are periodically reviewed and updated. For information about the revisions made between publications, see the Filled posts monthly tracking revisions notes report.

The monthly pipeline re-estimates the full historical time series each time it is run, allowing estimates to incorporate the latest available ASC-WDS, CQC PIR and CQC registration data. As a result, previously published monthly estimates may change slightly over time as additional information becomes available. Revision notes are published where methodological improvements or data corrections materially affect the estimates.

Annual publications use a fixed snapshot of the data and are not revised within the publication year.

3.3 Data quality and assurance

Data quality is assessed and managed throughout the production process to ensure that estimates are robust and reliable.

Validation checks are applied at multiple stages to identify issues in source data and derived variables. These include checks for missing or inconsistent values, duplicate records, unexpected ranges and invalid categorical values.

In addition to standard validation, bespoke business rules are applied to reflect known characteristics of the data, such as plausible workforce ranges and relationships between variables.

Where critical issues are identified, processing is halted and investigated before estimates are produced.

Quality assurance also includes review of intermediate outputs and final estimates to ensure they are consistent with expected patterns and external benchmarks.

3.4 Reproducibility and governance

The pipeline follows the principles of Reproducible Analytical Pipelines (RAP). Processing is automated to minimise manual intervention, code is version controlled and subject to peer review, and validation checks are applied throughout. Infrastructure and processing are also designed to be reproducible. Separate production and development environments are used to ensure stability.

3.5 Future development

As development continues, we plan to improve the granularity of outputs and expand the range of statistics that can be published from the pipeline.

Please see our website for our publication calendar and future plans.

4 Appendix

4.1 Data cleaning methods

This appendix provides further detail on the data cleaning steps applied to workforce data prior to imputation and modelling.

4.1.1 Dual registration

The CQC Care Directory includes instances of dual registration, where two providers are recorded as managing services at the same location. These appear as separate records with different location IDs but the same service details.

To avoid double counting the filled posts at these locations, one record from each dual-registered pair is removed.

The retained record is selected using the same ordering principles as CQC's guidance:

- earliest registered location.
- if both locations are registered on the same date, the lowest Location ID number.

4.1.2 Calculating ASC-WDS filled posts

ASC-WDS collects two separate measures of workforce size for each location:

- a location-level total staff figure.
- worker-level records, where one record represents one worker.

The ASC-WDS filled posts measure is derived by comparing these two submitted figures where the two measures provide sufficiently consistent evidence of the likely number of filled posts.

Where the two measures match exactly and the shared value is at least 3, that value is used directly.

Where an exact match is not available, a value is derived where both measures are at least 3 and sufficiently similar. Similarity is defined as either:

- an absolute difference of fewer than 5 posts, or
- a percentage difference below 10%.

In these cases, the average of the two measures is used.

A corresponding source field records whether the final value was based on exact agreement or the average of two similar submitted values. This approach prioritises strong agreement between independently submitted workforce measures while allowing for small differences in reporting.

Some ASC-WDS returns contain the value '999', which is used elsewhere in ASC-WDS as a placeholder for unknown values. Although this field is not formally coded in that way, investigation showed these records were not credible as workforce counts. These values are therefore set to missing before further processing.

Table 1. Calculating ASC-WDS filled posts from the total staff figure and worker record count

Location ID	Total staff figure	Worker record count	Rule applied	ASC-WDS filled posts
1-001	45	45	Exact match	45
1-002	50	53	Similar (average)	51.5
1-003	20	52	Values not similar	null
1-004	999	0	Invalid value	null

4.1.3 Deduplication of ASC-WDS filled posts

ASC-WDS monthly snapshots retain the most recently submitted value for each location. As a result, the same reported filled post value may appear repeatedly across consecutive periods even where no new submission has been made.

For estimating change over time, consecutive duplicate values are removed so that rates of change are based only on genuine updates rather than carried-forward snapshot values.

This ensures that repeated historical values do not distort estimates of workforce change.

Table 2. Calculating ASC-WDS filled posts from the total staff figure and worker record count

Location ID	Time period	ASC-WDS filled posts	ASCWDS filled posts deduplicated
1-001	Month 1	100	100
1-001	Month 2	90	90
1-001	Month 3	90	
1-001	Month 4	95	95
1-001	Month 5	100	100
1-001	Month 6	100	

4.1.4 Outlier detection – grouped providers

In some cases, providers with multiple CQC locations incorrectly submit the workforce for the whole organisation into a single ASC-WDS location account. These records can create implausibly large staffing counts at one location while leaving related locations with no data.

Potential grouped-provider submissions are identified where:

- a provider has multiple CQC locations.
- only one location has ASC-WDS data.
- the reported workforce is disproportionately large relative to:
  - beds at the location and/or provider (for care homes).
  - PIR employment counts (for non-residential locations).

Where identified, the location-level ASC-WDS value is set to missing.

Table 3. Example of a grouped care home provider

Provider ID	Location ID	Number of beds (location)	Number of beds (provider)	ASC-WDS filled posts	Cleaned ASC-WDS
1-101	1-001	5	50	-	-
1-101	1-002	10	50	75	null
1-101	1-003	15	50	-	-
1-101	1-004	20	50	-	-

4.1.5 Outlier detection – filled posts per bed Winsorization

For care homes, staffing values are compared against the expected number of posts based on the number of beds. Locations are grouped into comparable bed-size bands, and an expected posts-per-bed ratio is calculated for each group.

Values falling in the most extreme 5% of standardised residuals are treated as outliers. Rather than removing these observations, the values are Winsorized by capping them at the highest or lowest permitted posts-per-bed ratio within their bed-size group.

Minimum permitted ratio bounds are also applied to preserve realistic staffing levels in smaller care homes.

Table 4. Example of a grouped care home provider

Location ID	Number of beds	ASC-WDS filled posts	Filled posts per bed ratio	Winsorized ratio	Adjusted ASC-WDS filled posts
1-001	5	40	8.0	5.0	25
1-002	100	50	0.5	0.75	75

4.1.6 Forward filling of last submitted values

After trend-based cleaning, the most recent known value is carried forward for a limited period where no newer workforce data is available.

This allows some locations to remain at the same value between months. This reflects real world patterns more accurately than full de-duplication where no locations would remain on the same value.

The length of time a value is carried forward depends on the size of the location, with smaller locations assumed to remain stable for longer periods than larger ones.

This step helps reduce artificial volatility caused by irregular reporting and provides a realistic starting point for later imputation.

Table 5. Forward filling of the last submitted value

Location ID	Time period	ASC-WDS filled posts	ASCWDS posts forward filled
1-001	Month 1	75	75
1-001	Month 2		
1-001	Month 3	80	80
1-001	Month 4		
1-001	Month 5	100	100
1-001	Month 6		100
1-001	Month 7		100
1-001	Month 8		

4.1.7 Derived variables

Additional variables are created to support later stages of the pipeline, such as the imputation process and the regression models. These include measures such as:

- time registered – the number of months the location has been registered with CQC for.
- time since dormant – the number of months since the location was last identified as dormant.
- banding the number of beds scale into categories.

4.2 Imputation methods

This section provides detailed information on how missing workforce data for locations who have submitted ASC-WDS or CQC PIR data at some point in time are estimated.

4.2.1 Overview

Imputation is used to create a complete time series of filled posts for each location. This is necessary because workforce data is not available for all locations in all months.
The approach is based on estimating how workforce levels typically change over time and applying this to known data points.

4.2.2 Estimating the rate of change trendline

To estimate how workforce levels change, only locations with sufficient and consistent data are included. Locations are required to have at least two reported values and remain within the same broad service type (care home or non-residential). This ensures that estimated changes are based on stable and comparable data.

Constructing monthly observations

Observed data is not available for every month so to allow for more consistent comparisons over time, short term gaps in reporting are filled using straight-line interpolation. This step increases the number of usable data points when estimating change over time.

Table 6. Short term interpolation between known submissions

Location ID	Time period	ASC-WDS filled posts	ASCWDS posts interpolated
1-001	Month 1	20	20
1-001	Month 2		22
1-001	Month 3	24	24
1-001	Month 4		25
1-001	Month 5		26
1-001	Month 6		27
1-001	Month 7	28	28
1-001	Month 8		

Outlier handling in rate of change estimates

Extreme changes in monthly filled posts can distort estimates of typical workforce trends. To reduce this effect, unusually large increases or decreases are identified using a percentile-based approach and excluded when calculating the overall monthly rate of change.

These changes are not necessarily treated as errors as they may reflect genuine events affecting an individual location, such as a service expanding rapidly, absorbing activity from another location or splitting into multiple locations.

The submitted filled posts for these locations are retained in the source data and continue to be used for that location's own time series. However, the associated month-to-month change is

excluded from the calculation of the broader trendline, on the basis that it is unlikely to represent the typical pattern of workforce change across the sector. This ensures that the estimated trend reflects common underlying behaviour rather than exceptional location-specific events.

Worked example of a monthly rate of change

Table 7. Location level data for monthly rate of change calculations

Location ID	Month 1	Month 2	Included in monthly change?	Reason why excluded
1-001	-	20	No	Missing in Month 1
1-002	10	-	No	Missing in Month 2
1-003	15	17	Yes	
1-004	20	18	Yes	
1-005	25	28	Yes	
1-006	30	80	No	Extreme increase for this period
1-007	40	8	No	Extreme decrease for this period

Using only the included locations from the table above:

- Sum of Month 1 = 15 + 20 + 25 = 60
- Sum of Month 2 = 17 + 18 + 28 = 63
- Monthly rate of change = 63 / 60 = 1.05

The excluded values for location 1-006 and 1-007 are retained in the location's own time series but are not used to estimate the broader monthly sector trend.

Smoothing the rate of change

A monthly rate of change is calculated using observed data across all eligible locations.

To reduce short-term volatility, the rate is based on a rolling three-month comparison of totals rather than a single month alone. For each period, the total filled posts across the current three-month window is divided by the total across the equivalent preceding three-month window.

This smoothing approach reduces the influence of short-term reporting variation and produces a more stable estimate of how workforce levels typically increase or decrease over time.

$$\text{Rate of change } t = \frac{(\text{Month } t + \text{Month } t\text{-}1 + \text{Month } t\text{-}2)}{(\text{Month } t\text{-}1 + \text{Month } t\text{-}2 + \text{Month } t\text{-}3)}$$

Constructing the trendline

The sequence of monthly changes is combined to form a cumulative trend over time.
This trendline represents how workforce levels evolve relative to a starting point and is used as the basis for both interpolation and extrapolation.

Table 8. Calculating the cumulative rate of change trendline

Time period	Rate of change	Cumulative rate of change
Month 1	1.00	
Month 2	1.01	1.01
Month 3	1.02	1.03
Month 4	1.00	1.03
Month 5	0.99	1.02

4.2.3 Applying the trend to estimate missing values

The estimated trend is applied to known data points to generate a continuous time series using extrapolation and interpolation to produce a complete set of estimated values for each location.

Extrapolation

For periods before the first known value or after the last known value, missing values are estimated by applying the cumulative trendline forwards or backwards to estimate values in adjacent periods.

This allows the series to extend beyond known submissions while remaining consistent with the estimated pattern of workforce change over time.

In the example below, the known value in Month 5 is used as the anchor, with earlier and later periods estimated by applying the cumulative change relative to that point.

Table 9. Applying the trend – extrapolation

Time period	Cumulative rate of change	ASC-WDS filled posts	Extrapolated filled posts
Month 1	1.00		25.00
Month 2	1.01		25.25
Month 3	1.03		25.75
Month 4	1.02		25.50
Month 5	1.00	25	25.00
Month 6	0.98		24.50

Interpolation

Where two known submissions exist for the same location, missing values between them are estimated using a trend-based interpolation approach.

The process starts from the earlier known value and applies the estimated trendline forward to create an initial extrapolated path between the two dates.

Because the value at the later date is also known, the difference between the extrapolated end value and the observed end value is then calculated. This difference is treated as a residual adjustment.

The residual is distributed proportionally across the missing periods according to how far each month lies between the two known submissions. This ensures that:

- the interpolated series follows the overall trendline shape.
- the final interpolated value aligns exactly with the known end point.

This produces a smooth series that reflects both the estimated sector trend and the observed values at each end of the gap.

Interpolated posts = Extrapolated posts + (proportion through gap × End point residual).

Table 10. Applying the trend - interpolation

Time period	Cumulative rate of change	ASC-WDS filled posts	Extrapolated filled posts	Proportion through gap	End point residual	Interpolated filled posts
Month 1	1.00	25	25.00	0.00	-2.00	25.00
Month 2	1.01		25.25	0.25	-2.00	24.75
Month 3	1.03		25.75	0.50	-2.00	24.75
Month 4	1.02		25.50	0.75	-2.00	24.0
Month 5	1.00	23	25.00	1.00	-2.00	23.00

4.2.4 Combining data sources

ASC-WDS is used as the primary source of workforce information. As both ASC-WDS and PIR data are incomplete, a combined filled posts series is created which carries forward ASC-WDS values over time and is selectively supplemented with PIR data where appropriate.

ASC-WDS figures are prioritised as they are derived from internally consistent submissions (total staff and worker-level records), providing a high level of confidence in the reported values. In contrast, PIR data is based on a single reported figure and is therefore more susceptible to reporting variation.

PIR collects ‘people directly employed’ figures so these values are first converted to estimated filled posts. This is calculated by using a global ratio derived from locations where both data sources are available and broadly consistent.

PIR data is only introduced into the combined series where it provides meaningful additional information beyond the existing ASC-WDS trend. The rules for creating the combined values are:

- ASC-WDS values are always carried forward.
- PIR values are used to supplement the series where:
  - no ASC-WDS data has ever been submitted, or
  - the most recent ASC-WDS submission is more than two years older than the PIR submission and the PIR value differs substantially from the carried-forward ASC-WDS value.

PIR figures are not introduced where PIR and ASC-WDS values are similar as it does not provide additional information and may reflect differences in data collection rather than genuine change.

This approach ensures that the series remains anchored to the more reliable ASC-WDS data, while incorporating PIR data where it provides a better reflection of current workforce levels and increases the volume of usable data for estimation and modelling.

4.3 Modelling methodology

This appendix provides additional detail on the models used to estimate filled posts.

4.3.1 Rolling average model

A rolling average is calculated to provide a national benchmark of workforce levels over time, which smooths short-term fluctuations while remaining responsive to gradual changes in workforce levels.

For care homes, the rolling average is calculated as the national mean of the ASC-WDS filled posts per bed ratio, using a rolling window over recent months and is split based on location size. Workforce size is closely related to the number of beds, with smaller homes typically having higher posts-per-bed ratios than larger homes. Banding by bed size ensures these differences are preserved in the benchmark.

For non-residential services, the rolling average is calculated as the national mean of ASC-WDS filled posts using a rolling window over recent months.

The rolling average is used in two ways:

- as an input feature within regression models, supporting the estimation of underlying national trends.
- as a fallback estimate for individual locations where insufficient information is available from other sources.

4.3.2 Model type

Filled posts are estimated using linear regression models with regularisation (Lasso). This approach allows relationships between workforce size and location characteristics to be estimated while reducing the risk of overfitting.

Separate models are used for care homes and non-residential services to reflect differences in how workforce size is determined.

Lasso regression was selected because it performs well where many related input features are available and helps reduce overfitting by shrinking less informative features towards zero.

4.3.3 Model features

The models use a range of characteristics to describe each location. These include information on services provided, location type, geography, and size.

From March 2022 onwards, a dormancy indicator is available in the CQC data which contains a Yes/No flag. This identifies locations that are registered but not actively providing services at the time. For non-residential services, the time since a location was last dormant is included as a model input where available.

A list of feature categories used in each model is shown in the table below.

Table 11. Model features

Feature category	Care home model	Non-residential (no dormancy, prior to March 2022)	Non-residential (with dormancy, March 2022 onwards)	Description
Activities / services offered	✓	✓	✓	Activity count, service count
Time trend	✓	✓	✓ (including cubic term)	Indexed time variable capturing sector trends
Workforce trend	✓	✓		Rolling average (posts or posts per bed ratio)
Capacity	✓			Number of beds
Registration duration		✓	✓	Time since registration
Dormancy			✓	Time since last dormant
Geography	✓	✓	✓	Region indicators
Rural/urban	✓	✓	✓	Rural/urban classification
Services	✓	✓	✓	Service type indicators
Specialisms	✓	✓	✓	Client group indicators
Related locations		✓	✓	Previously registered as another

4.3.4 Model training

Input features are derived and transformed prior to modelling. This includes:

- encoding categorical variables into binary indicators.
- capping extreme values to reduce the influence of outliers.
- deriving additional variables such as rolling averages and time-based measures.

These steps ensure that features are in a suitable format for modelling and that extreme values do not disproportionately influence model estimates.

4.3.5 Model training

The models are trained using historical data where filled posts are known.

To assess performance, the data is split into training and test sets:

- 80% of locations are used to train the model.
- 20% are used to evaluate model performance.

This approach provides an indication of how well the model generalises to unseen data.

4.3.6 Model performance

Model performance is monitored separately for each model and includes the following metrics:

- R-squared (how well the model explains variation in the data)
- Proportion of estimations within 10 and 25 of known values

Modelled estimates are only used for locations where no known ASC-WDS or CQC PIR submission is available for the relevant period, representing approximately 20% to 25% of locations in a typical month. The estimates are not intended for analysis at individual location level. When aggregated across grouped geographical areas or service groups, over and under estimation at individual locations tend to offset each other, resulting in stronger performance at aggregate level than is reflected by location-level performance metrics alone.

Table 12. Performance metrics

Metric	Care home model	Non-residential (no dormancy, prior to March 2022)	Non-residential (with dormancy, March 2022 onwards)
R-squared	0.39	0.11	0.15
Proportion of estimates within 10	66.9%	23.3%	28.0%
Proportion of estimates within 25	91.8%	56.4%	61.5%

4.3.7 Model updates and versioning

The models are retrained each time the pipeline is run. This ensures that the latest available data is used and the models reflect current patterns in the sector.

Each model run is versioned, and key information such as model performance and configuration is recorded for reproducibility.

4.3.8 Model-based imputation

Following model estimation, a second imputation stage is applied to further improve the completeness and quality of the time series.

While the initial imputation applies a national trend in percentage terms, this second stage uses location-specific model predictions to estimate how workforce levels change over time. This produces a more personalised set of estimates that reflect each location's characteristics.

Applying model-based change

For each location, the change in modelled filled posts between consecutive time periods is calculated in absolute terms. These changes are then applied to the observed or previously imputed values to estimate missing periods.

This differs from the earlier approach:

- Initial imputation applies a multiplicative (percentage) change.
- Model-based imputation applies an additive (absolute) change.

This reflects that model predictions capture expected differences in workforce size based on location-specific features, rather than assuming uniform proportional growth across all locations.

Model-predicted changes are applied both backwards and forwards in time:

- Earlier periods are estimated by reversing the modelled change/
- Later periods are estimated by applying the modelled change forward/

This produces a continuous time series that reflects the expected trajectory for that specific location.

Table 13. Extrapolation based on changes in model predictions

Time period	Model prediction	Change from previous	ASC-WDS filled posts	Extrapolated filled posts
Month 1	25.8			31.6
Month 2	25.9	+0.1		31.7
Month 3	26.2	+0.3	32.0	32.0
Month 4	26.0	-0.2		31.8

Where two observed values exist, interpolation is applied using the same approach described in chapter 4.2.2. The only difference is that model-based absolute changes are used in place of percentage trend changes.

Skills for Care
West Village
Wellington Street
Leeds
LS1 4LT

T: **0113 245 1716**
E: **info@skillsforcare.org.uk**

skillsforcare.org.uk

facebook.com/skillsforcare

linkedin.com/company/skills-for-care