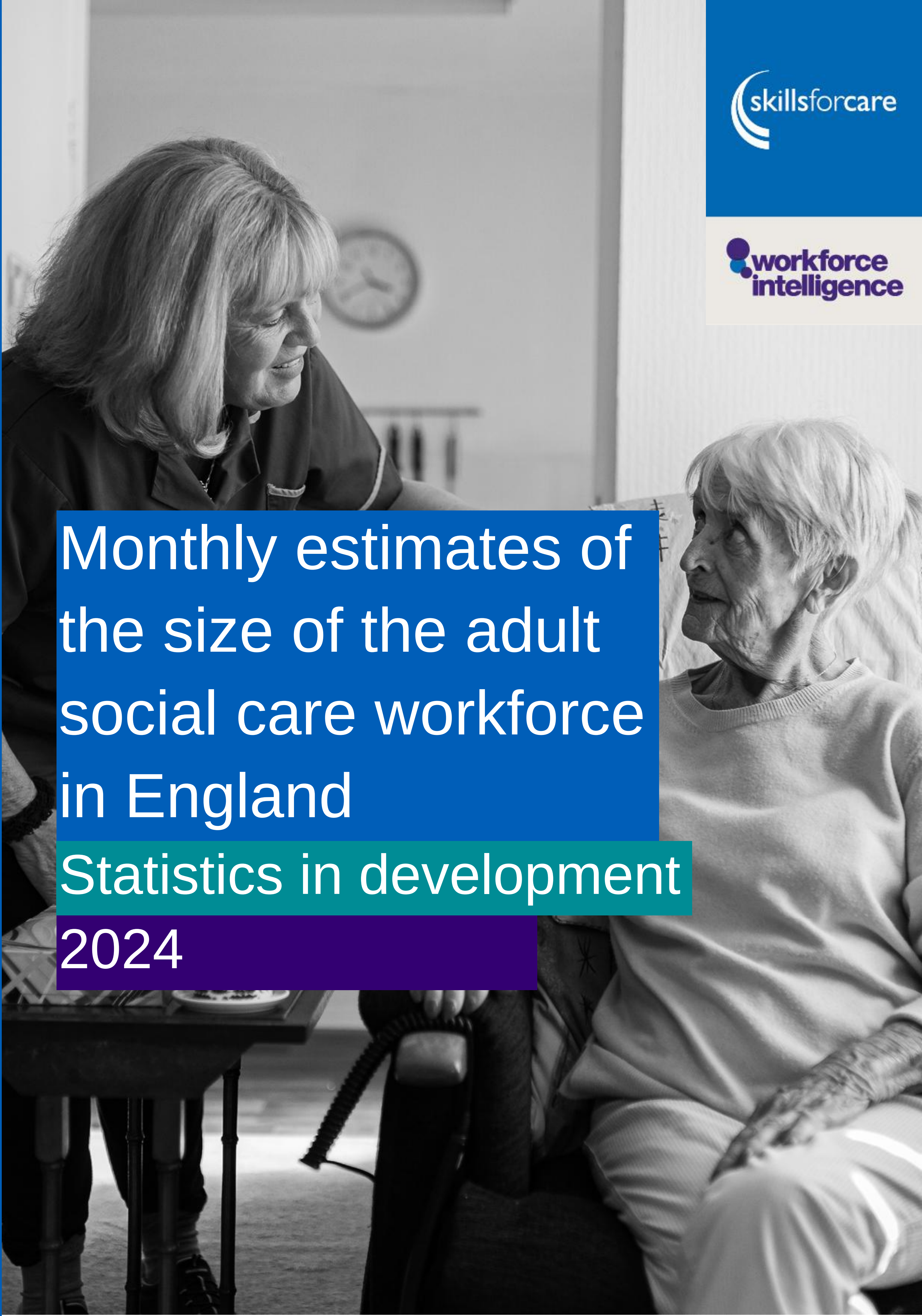


skillsforcare

workforce intelligence

Monthly estimates of the size of the adult social care workforce in England

Statistics in development

2024

Revisions history

The methodologies outlined in this paper are periodically reviewed and updated. For information about the revisions made between publications, see the table below.

Version	Date	Item	Revision
V1	November 2024	All	First publication
V2	December 2024	Non-residential methods	Non-residential estimates are now included in the tracker for the first time. The methods for these are now included in this paper throughout.
		Rolling average	The rolling average is now calculated based on rate of change.
		Accuracy metrics	We have added a column to our accuracy metrics to show the percentage of locations that pass either metric. This is because the individual metrics are useful for different sizes of location

Contents

Summary

We have been working on using data engineering to automate our processes for creating estimates of the size of the adult social care workforce in England. We are now able to share the results of this work, for independent sector care homes, in our improved monthly tracking dashboard.

The key benefits of using data engineering include:

- We can now provide estimates of the number of filled posts in care homes and non-residential locations[1] **every month rather than once per year**).
- We now make use of all ASC-WDS data collected over time and have implemented new analytical methods which have **improved the accuracy and consistency of our estimates** (see the technical detail section below for how we have achieved this).
- The new methods are also more responsive to changes in the sector. **Changes happening in the sector will be reflected sooner in our estimates**.
- Once the data engineering work is complete, we will be able to share estimates at a **more granular level than before**.

As these new statistics are still **under development**, please use the annual figures in our 'Size and Structure' and 'State of' publications from 2024 for official purposes.

It is likely that these new statistics will change as development continues.

We plan to use our new estimates for filled posts in the CQC-regulated independent sector officially in the 'Size and Structure' and 'State of' reports in 2025.

It is important to note that the annual figures in the monthly tracking dashboard do not match our previously published annual estimates, due to the improved methods we are using. The following pages describe these improvements in more detail.

[1] Note that our non-residential estimates are currently only showing the percentage change each month. Absolute estimates we be added in future releases.

Technical detail

1. Our previous methods

Prior to automating our estimates of the number of filled posts in the CQC-regulated independent sector, we were only able to provide annual estimates of the size of the sector and monthly updates on the subset of the sector who were updating ASC-WDS during the year.

The monthly tracking was not able to account for new locations opening or locations closing during the year. Additionally, we were only able to report on percentage change since the previous annual estimates; we were unable to produce a whole sector estimate each month.

2. Our new methods

Our new automated method uses all available ASC-WDS data over time and estimates the number of filled posts for each location at the end of each month, using one of three methods:

A. **Using up to date ASC-WDS data**: If we have ASC-WDS data completed on the number of filled posts for a particular location in a particular month then that figure is used (after some data quality filtering).
B. **Using ASC-WDS data completed/updated during a different time period**: If we have ASC-WDS data for a particular location but it is not completed/updated in the month we are estimating, then we create an estimate based on data we have in the ASC-WDS for that location at different points in time (imputation).
   - We make estimates from the most recent known value at a location forwards in time and the earliest known value for that location backwards in time (extrapolation).
   - We make estimates for months between two known ASC-WDS values (interpolation).
   - Both extrapolated and interpolated estimates follow the trend seen in ASC-WDS data completed each month. A simplified example would be that if locations completing ASC-WDS in February 2024 were 1% larger, relative to their number of beds, than locations completing ASC-WDS in January 2024, then a location completing ASC-WDS data in January 2024 would have an extrapolated value for February 2024 that was 1% higher than the value they submitted in January 2024.
C. **Using a regression model:** For **non-residential locations only**, if we do not have ASC-WDS data for a particular location at all, but we do have a figure of people directly employed from the CQC Provider Information Return (PIR) dataset, then we use a linear regression model to create an estimate of filled posts for that location based on the CQC data.
D. **Using a machine learning model**: If we do not have ASC-WDS data for a particular location at all (or in the case of non-residential locations, no ASC-WDS or PIR data, see Accuracy), then we use a machine learning model to create an estimate of filled posts for that location based on data we have in the ASC-WDS for similar locations.
   The care home machine learning model uses the following features about a location to estimate the number of filled posts:
   - the number of beds
   - the number of services provided
   - the rolling rate of change of the filled posts to beds ratio at the time
   - the types of services provided
   - the geographical region
   - the ONS classification of rural-urban indicators.

   The non-residential machine learning models use the following features about a location to estimate the number of filled posts:

   - the number of activities provided
   - the number of services provided
   - the number of specialisms provided
   - the length of time the location has been registered
   - the rolling rate of change of the filled posts for non-residential locations at the time

- the types of services provided
- the geographical region
- the ONS classification of rural-urban indicators.
  whether the location was dormant at the time (this only exists for more recent data)

Combining the estimates for each location at each point in time allows us to provide an estimate for the number of filled posts each month for every location. From those total figures we can calculate the percentage change over time.

3. Coverage

As Table 1 shows, we have data relating to about two thirds of all care homes, all of which can now be used to generate our estimates. Our old methods were within-year estimates, meaning we could only use data held in ASC-WDS at that point in time. Being able to create estimates across the full history of our dataset allows us access to more data points for many more locations and allows us to anchor our estimates around what was known earlier or later in time.

For our non-residential estimates, Table 1 shows that we have data relating to over 80% of all non-residential locations (based on data from ASC-WDS and the CQC PIR dataset). This translates to a large improvement in accuracy non-residential estimates because we have some data, rather than none for around four fifths of establishments. Previously less than half of our estimates were based on data specific to that establishment.

This advantage also applies to the machine learning estimates. Our machine learning models have all the data collected over time to learn from rather than a single point in time.

Table 1. Proportion of locations estimated by each method by service type, March 2024
Source: Skills for Care estimates

Estimate source	Description	Care homes with nursing	Care homes without nursing	Non-residential
Current ASC-WDS data	The number of filled posts based on data we have in the ASC-WDS for that location at that point in time	11%	11%	9%
Imputed ASC-WDS data	An estimate of filled posts based on data we have in the ASC-WDS for that location at different points in time	54%	59%	33%
Regression model from CQC PIR data	An estimate of filled posts based on CQC PIR data (non-residential only)	n/a	n/a	40%
Machine learning model(s)	An estimate of filled posts based on data we have in the ASC-WDS for similar locations over time	36%	30%	18%

4. Accuracy

We run diagnostic checks on our regression model estimates and machine learning model estimates against known ASC-WDS values to assess the quality of the estimates produced. We have also assessed the quality of the imputation process by comparing known ASC-WDS data, to what we would have estimated for these locations using imputation if they hadn't submitted.

The following metrics have been assessed for imputation, the regression model, and each machine learning model:

- The percentage of estimates that are within 10 filled posts of the ASC-WDS value
- The percentage of estimates that are within 25% of the ASC-WDS value
- The percentage of estimates that are within 10 filled posts or 25% of the ASC-WDS value

Checking the proportion of estimates within 10 filled posts is useful for checking estimates of smaller establishments. For larger establishments, this metric isn't as insightful as predicting 789 when the actual value is 800 would fail on this measure but is still a good estimate. Checking the proportion of estimates within 25% is useful for checking estimates of larger establishments. For smaller establishments, this metric isn't as insightful, as predicting 5.01 when the actual value is 4, would fail on this measure, but is still a good estimate. Because of these limitations, we also provide the proportion of estimates that pass one or both of these estimates.

The table shows how the imputed estimates are very accurate (within 10 filled posts or 25% of the actual number of filled posts over 97% of the time) and therefore will result in an improvement on estimates produced with the previous method.

The care home machine learning model also predicts with high accuracy (to a lesser extent) with over 88% of estimates within 10 filled posts or 25% of the actual figure.

For non-residential locations, the regression model for PIR data predicts with high accuracy, with 82% of estimates within 10 filled posts or 25% of the actual figure. The non-residential model with dormancy as a feature has a moderately high level of accuracy, with 75% of estimates within 10 filled posts or 25% of the actual figure. Where dormancy data is not available, the accuracy reduces slightly with around two thirds of estimates within 10 filled posts or 25% of the actual figure.

Because of these differences in accuracy, we use the most accurate model wherever possible in our estimates. For example, for non-residential estimates, we use the ASC-WDS data where this exists and is of good enough quality. Then establishments where we have some ASC-WDS data, we use imputation to fill in any gaps. For non-residential establishments without any ASC-WDS data, but who have PIR data, we use the PIR model. Finally, we use the care home machine learning model or the non-residential machine learning models, prioritising the with-dormancy model over the without-dormancy model wherever possible.

Additionally, our estimates are never used at location level, they are always aggregated into geographical areas. So, an estimate falling outside of these metrics does not necessary cause an issue. The size and direction of the differences for each model has been analysed and the models are 'wrong' equally as often on the high and low side which means that, once aggregated, the statistics presented are unlikely to be skewed (i.e. the models are equally as likely to predict 70 filled posts for a location with 50 filled posts as they are to predict 30 filled posts).

Table 2. Proportion of location level estimates within predefined distance of actual ASC-WDS value, 2013 to 2024

Source: Skills for Care estimates

Estimation type	Metric	Care homes with nursing	Care homes without nursing	Non-residential
Imputation (forward extrapolation)	% within 10 filled posts of known value	90%	97%	85%
	% within 25% of known value	96%	92%	96%
	% within 10 filled posts or 25% of known value	97%	98%	97%
Care home machine learning model	% within 10 filled posts of known value	53%*	78%	n/a
	% within 25% of known value	87%	83%	n/a
	% within 10 filled posts or 25% of known value	88%	95%	n/a
Regression model (non-residential only)	% within 10 filled posts of known value	n/a	n/a	54%
	% within 25% of known value	n/a	n/a	72%
	% within 10 filled posts or 25% of known value	n/a	n/a	82%
Non-residential machine learning model (with dormancy)	% within 10 filled posts of known value	n/a	n/a	38%
	% within 25% of known value	n/a	n/a	71%
	% within 10 filled posts or 25% of known value	n/a	n/a	75%
Non-residential machine learning model (without dormancy)	% within 10 filled posts of known value	n/a	n/a	28%
	% within 25% of known value	n/a	n/a	62%
	% within 10 filled posts or 25% of known value	n/a	n/a	65%

* The percentage within 10 for care homes with nursing is expected to be smaller as care homes with nursing are larger on average than care homes without nursing.

5. Case Studies

The following case studies use example data and are intended to show how each step of the method would be applied to a location in practice. The first case study describes a location that has submitted data to ASC-WDS, whereas the second case study describes a location that has never submitted data.

5.1. Case study 1: A care home with ASC-WDS data

Location A is a medium-sized independent CQC-regulated care home in England that has been open for over six years. However, they only started submitting data to ASC-WDS in 2020 and provided updates to their data in 2021 and 2023. As ASC-WDS generally retains data for up to two years, they also appear to have data in 2022 and 2024, as it has been carried forward from their previous updates (see Table 3). The CQC also hold data on the number of beds at Location A, which is combined with the ASC-WDS data to assist with estimation.

Table 3. Location A's ASC-WDS submissions over the past 6 years

Source: Example data

Year	ASC-WDS raw	CQC number of beds
2019		20
2020	90	20
2021	900	20
2022	900	20
2023	90	20
2024	90	20

Previous estimation method

Our previous method would have estimated each year of data separately, and without knowledge of later years' data. In 2019, we would have estimated the number of filled posts using a regression model and the number of beds, see Table 4. Care homes are known to have a strong relationship between the number of filled posts and beds, however some places may have higher ratios (for example if they care for people with higher support needs). In our example, Location A is one such care home, which doesn't fit the typical pattern. However, we have no other data available and so we fill the gap with a regression model that estimated 25 filled posts based on the number of beds.

In 2020, Location A signs up to the ASC-WDS and provides data saying they have 90 filled posts. This would be used as their estimate for that year.

In 2021, they resubmit their data and say they have 900 filled posts. However, the CQC data says that they don't have any more beds (they accidentally added an extra zero when inputting their data). Due to the CQC data, we can see that their reported filled posts look too high, but we can only see the 2021 data, and not their 2020 submission. Therefore, we filter out their data and use the value from the regression model to replace the very high data in ASC-WDS.

Location A doesn't update their data in 2022 and so the 900 value from 2021 is carried forward into our 2022 data. It is again replaced by a modelled value because it is so much higher than would be expected for a 20-bed care home.

In 2023, Location A corrects their data back to 90 filled posts and as in 2020, we use this value for our estimates.

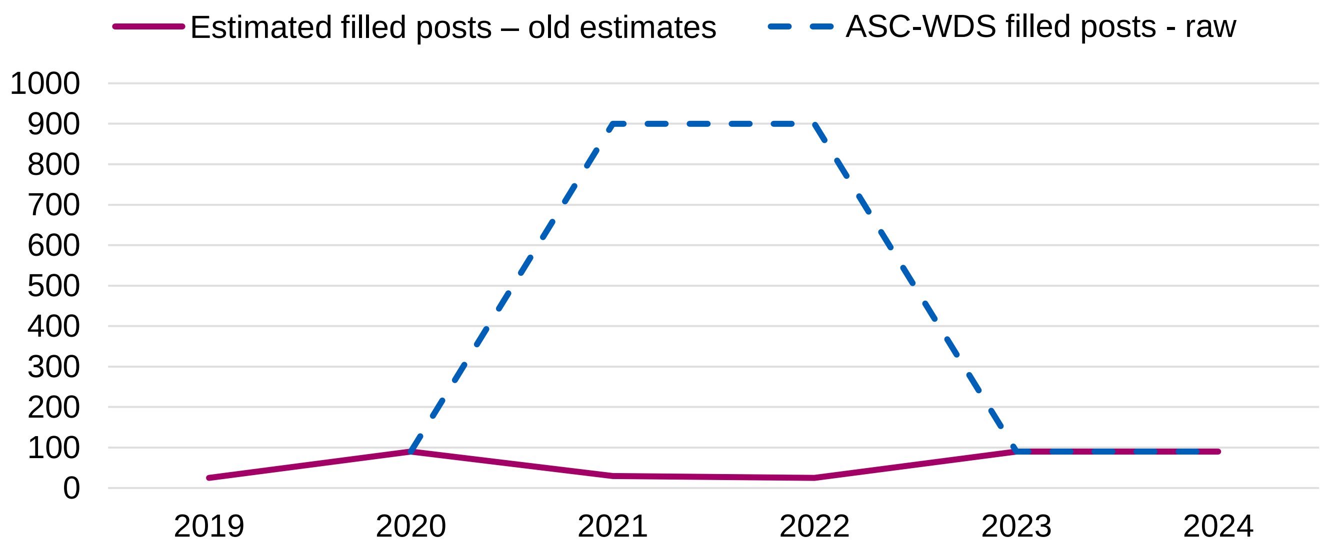
Finally in 2024, Location A doesn't update their data again, meaning that the value of 90 is carried forwards from the previous year. This is again used for our estimates, see Table 4.

Table 4. Location A's estimates using the old methods over the past six years
Source: Example data

Year	ASC-WDS filled posts - raw	CQC number of beds	Estimated filled posts – old estimates
2019		20	25
2020	90	20	90
2021	900	20	30
2022	900	20	25
2023	90	20	90
2024	90	20	90

Chart 1 shows these estimates plotted alongside the data in ASC-WDS for illustration.

Chart 1. Location A's ASC-WDS submissions and estimates using the old methods over the past six years
Source: Example data



New estimation method

Our new method takes several steps to smooth out and improve these estimates. The first step in our new methods is to remove repeated values across different snapshots over time (deduplication), see Table 5. The data in 2022 and 2024 were not updated submissions, so these values are removed. The data for 2020, 2021 and 2023 were all updated submissions, so they are retained in the deduplication stage. There was no data available in 2019, so this row remains blank.

Table 5. Location A's ASC-WDS submissions deduplicated over the past six years
Source: Example data

Year	ASC-WDS raw	CQC number of beds	Deduplication
2019		20	
2020	90	20	90
2021	900	20	900
2022	900	20	
2023	90	20	90
2024	90	20	

The next step in our process is to remove outliers in our data. We use several methods to clean and filter the data; this example only shows how we handle outliers in our data.

As mentioned above, the number of filled posts has a strong relationship to the number of beds in a care home. For this reason, we calculate the ratio between the number of beds and the ASC-WDS data and look for outliers in the ratio. Particularly high or low ratios are capped, see Table 6.

Table 6 shows the three years which have deduplicated data (2020, 2021 and 2023) all have a ratio calculated. The years without deduplicated data (2019, 2022, and 2024) are ignored for this step. The ratio is calculated by dividing the deduplicated filled posts by the number of beds. The ratio of 4.5 in 2020 and 2023 is within our accepted tolerance of 0.75 to 5.0. The ratio of 45.0 is therefore considered an outlier and recalculated using our highest accepted ratio (5.0). The result of this is that the value of 900 filled posts for 2021 is replaced by an estimate of 100.

Table 6. Location A's ASC-WDS submissions filtered over the past six years
Source: Example data

Year	ASC-WDS raw	CQC number of beds	Deduplication	Ratio	Capped ratio	Filtering
2019		20				
2020	90	20	90	4.5	4.5	90
2021	900	20	900	45.0	5.0	100
2022	900	20				
2023	90	20	90	4.5	4.5	90
2024	90	20				

Finally, after the data has been deduplicated and filtered, we impute the missing values based on the data from other points in time for that location. We use a rolling average of the rate of change of the filled posts to bed ratio (for locations that have updated ASC-WDS data) to extrapolate the most recent known value forwards in time and the earliest known values backwards in time.

Assuming the rolling average is as shown in Table 7, this gives an estimate of 80 filled posts for 2019 and 85 filled posts for 2024. We then interpolate between two known values at a location using the rolling average of the rate of change of the filled posts to bed ratio, so that not all interpolated values are a straight line. For Location A, let's say this trend gives us value of 95 in 2022.

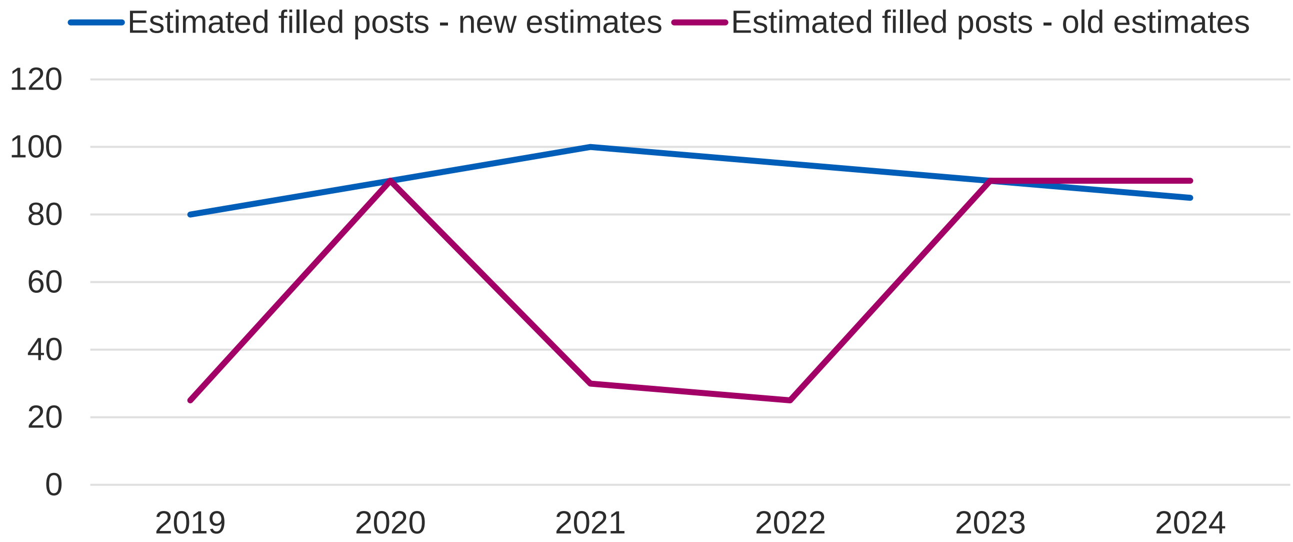
Table 7. Location A's ASC-WDS submissions imputed over the past six years
Source: Example data

Year	ASC-WDS filled posts	CQC number of beds	Deduplicated number of filled posts	Ratio of filled posts to beds	Capped Ratio	Filtering values	Imputed ratio of filled posts to beds	Estimated filled posts – new estimates
2019		20					4	80
2020	90	20	90	4.5	4.5	90	4.5	90
2021	900	20	900	45	5	100	4.5	100
2022	900	20					4.5	95
2023	90	20	90	4.5	4.5	90	4.5	90
2024	90	20					4.25	85

These imputed figures are used as our estimates for Location A. As shown in Chart 2, using imputation rather than regression results in a much smoother trend for Location A. We are only able to do this because our new methods allow us to use all the data for this location at once, rather than having isolated yearly snapshots of data.

Another improvement is that the estimate for 2024, for example, changes from 2023 based on the pattern of changes observed in locations that do have updated ASC-WDS in 2024. In the old method their estimated value stayed the same. As such, the new method will be able to respond quicker to changes in the sector.

Chart 2. Location A's ASC-WDS submissions and estimates using the old and new methods over the past six years
Source: Example data



5.2. Case study 2: A care home without ASC-WDS data

Location B is also a medium-sized independent CQC-regulated care home and domiciliary care provider in England that has been open for over six years. However, they have never submitted data to ASC-WDS. The CQC hold data on the number of beds at Location B, which is used as a starting point for our estimates, see Table 8.

Table 8. Location B's ASC-WDS submissions over the past 5 years
Source: Example data

Year	ASC-WDS raw	CQC number of beds
2019		20
2020		20
2021		20
2022		20
2023		20
2024		20

Using our old method, all these years would be estimated individually using a regression model that takes into account the number of beds at the location, see Table 9.

Table 9. Location B's estimates using the old methods over the past six years
Source: Example data

Year	ASC-WDS raw	CQC number of beds	Old estimates
2019		20	55
2020		20	51
2021		20	55
2022		20	57
2023		20	60
2024		20	55

With our new method, as we have no data in ASC-WDS for Location B, we are not able to impute any data. Instead, the data from all the known locations over time is used to estimate the filled posts at Location B, using a machine learning model (see Our new methods). This model takes into account more features than the regression model we used previously and has access to data from every available year, instead of just the current year.

This leads to a slight improvement in estimates in general. However, there can be substantial improvements for care homes which offer additional services. In this example, Location B offers domiciliary care as well as being a care home and receives a higher estimate as a result, see Table 10.

Table 10. Location B's estimates using the new and old methods over the past six years
Source: Example data

Year	ASC-WDS raw	CQC number of beds	Old estimates	Care home model – new estimates
2019		20	55	82
2020		20	51	83
2021		20	55	80
2022		20	57	85
2023		20	60	87
2024		20	55	85

5.3. Case study 3: A non-residential location without ASC-WDS data that has CQC PIR data

Location C is a large-sized independent CQC-regulated non-residential location in England that has been open for over six years. They have never submitted data to the ASC-WDS. Using our old estimates, these locations would receive the average value for non-residential locations. However, the CQC hold data on the number of people directly employed at Location C since 2020, which is used as a starting point for our estimates, see Table 11'

.

Table 11. Location C's estimates using the old methods over the past six years

Source: Example data

Year	ASC-WDS raw	Old estimates	PIR raw
2019		55	
2020		51	100
2021		55	100
2022		57	90
2023		60	95
2024		55	95

PIR data was first collected in 2020, so we begin by deduplicating and imputing the PIR data, just as we would for ASC-WDS data (see Case study 1: A care home with ASC-WDS data). See Table 12.

Table 12. Location C's imputation of PIR data over the past six years

Source: Example data

Year	ASC-WDS raw	Old estimates	PIR raw	PIR deduplicated	PIR imputed
2019			55		102
2020		100	51	100	100
2021		100	55		95
2022		90	57	90	90
2023		95	60	95	95
2024		95	55		93

These imputed PIR values are then passed to the PIR regression model (see Our new methods), which converts the number of people directly employed to the number of estimated filled posts. These figures become our estimates for Location C, see Table 13.

Table 13. Location C's estimates using the new and old methods over the past six years

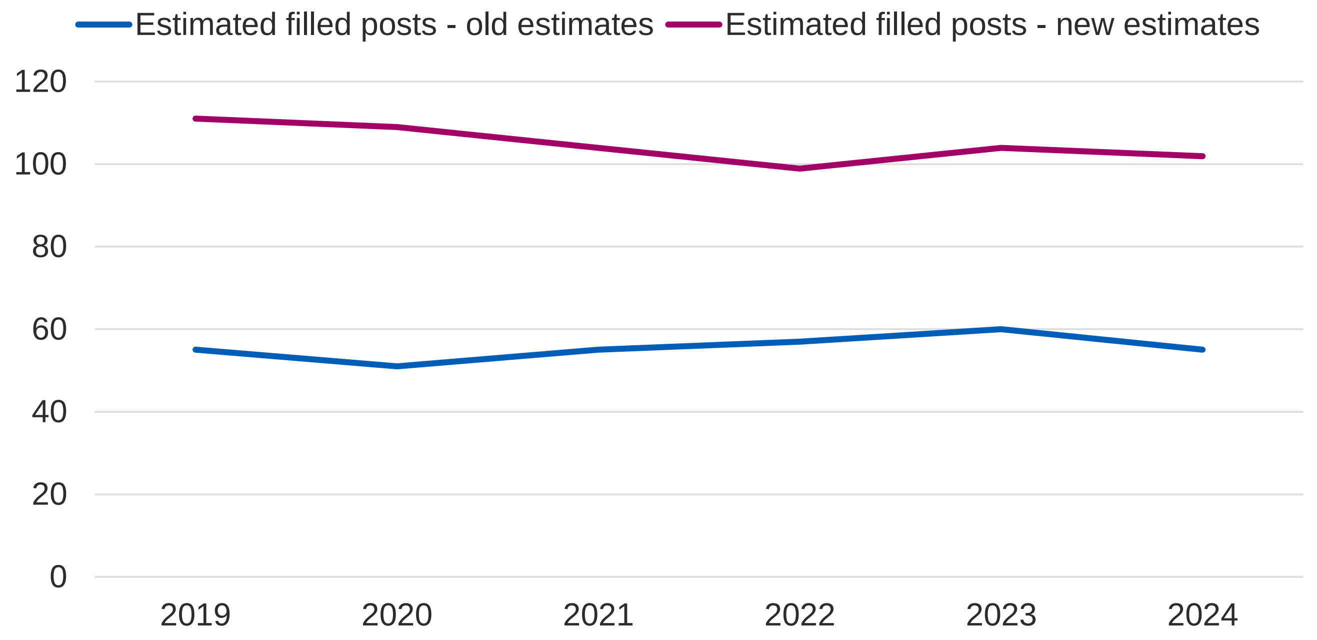
Source: Example data

Year	ASC-WDS raw	Old estimates	PIR raw	PIR deduplicated	PIR imputed	PIR regression model
2019			55		102	111
2020		100	51	100	100	109
2021		100	55		95	104
2022		90	57	90	90	99
2023		95	60	95	95	104
2024		95	55		93	102

Using PIR data in this way allows us to provide a more accurate estimate than if we knew nothing about Location C. This leads to a noticeable improvement for non-residential locations where we have PIR data at some point in time (see Chart 3), as it allows us to "anchor" the estimates, because we already have some information about the general size of the location.

We can then use that to create a more accurate estimate than if we had used the overall average.

Chart 3. Location C's estimates using the new and old methods over the past six years
Source: Example data



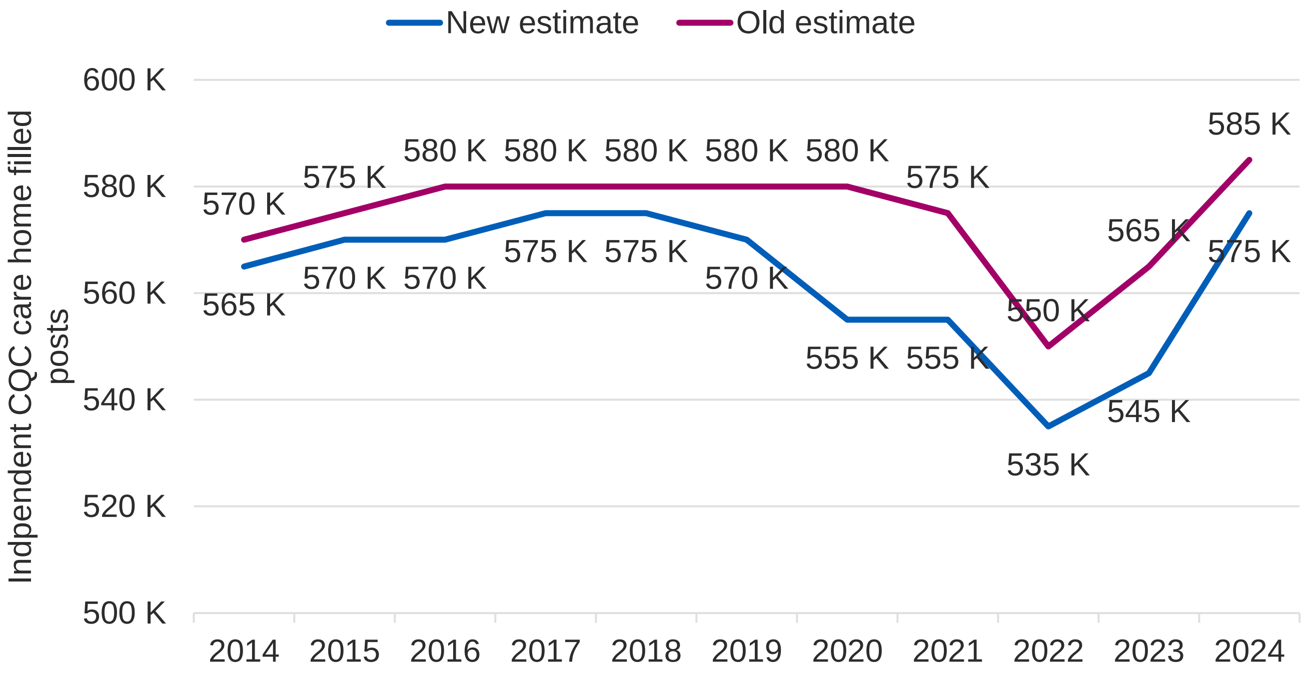
6. Changes to our historical estimates

As these new estimates are different to our previously published estimates, we have undertaken a thorough analysis of these changes.

6.1 Differences between our new care home estimates and previous annual figures

Chart 4 shows our old and new rounded care home estimates every March since 2014 to March 2024.

Chart 4. A comparison of our new and old annual independent CQC-regulated care home estimates, March 2014 to March 2024
Source: Skills for Care estimates



In March 2022, our new rounded estimate of the number of filled posts at independent CQC-regulated care homes in England was 15,000 filled posts lower than previously. Table 14 shows the aggregated differences in estimates by method of estimation.

The biggest change overall was where we were able to use imputation using ASC-WDS data from a different time period rather than a regression model (with the new estimate being 5578 lower than previously estimated figures for these locations). Table 2 demonstrates how estimates made using imputation are very accurate and therefore the difference from the previous estimates will be an improvement.

Another reason for the difference is that the new process identifies dual registrations in the CQC Locations API and removes them (5,577 posts). Dual registered locations are locations registered with more than one provider and are included twice, with two different location IDs, in the CQC API. These duplicates were not identified by the old process.

Table 14. Difference between new and old independent CQC-regulated care home estimates by source and service type, March 2022
Source: Skills for Care estimates

New estimate source	Old estimate source	Number of locations	Difference: New estimate minus previous estimate	Average difference per location
Current ASC-WDS data	ASC-WDS data	1203	253	0.2
	Regression model	95	951	10.0
Machine learning model	ASC-WDS data	86	230	2.7
	Regression model	4274	-2072	-0.5
Imputed ASC-WDS data	ASC-WDS data	4656	-626	-0.1
	Regression model	4290	-5578	-1.3
CQC dual registrations		63	-5577	-88.5
Total difference		**14667**	**-13541**	**-0.9**

The largest difference in our estimates can be seen to be in 2020 (see Chart 4). This was around and after the pandemic when we started producing a monthly estimate for how filled posts were changing within the year.

From 2021 onwards we started to utilise these monthly figures to assist with our annual estimates and the differences since that point were smaller (it allowed us to do some of the imputation steps present in the new process). One of the benefits of our new automated process is that we can apply the current improvements consistently across the history of our data.

6.2 Differences between our new estimates and previous monthly estimates

Our previous monthly estimates only looked at a relatively small subset of locations that had updated ASC-WDS data prior to March 2024 and after March 2024. As such they were presented as a temperature check on the direction of travel since our annual estimates and not representative statistics.

The new method represents an improvement for monthly tracking because it:

- Makes use of all the data collected in ASC-WDS over time and creates an estimate for every location

- Includes the impact of locations opening and closing
- Uses precise update dates, rather than comparing 'before and after March 2024'.

Conclusion

Key Strengths

We are now able to:

- Provide monthly filled posts estimates for the number of filled posts at care homes in the CQC-regulated independent sector
- Provide monthly filled posts estimates for the number of filled posts at non-residential establishments in the CQC-regulated independent sector
- Provide monthly estimates of all locations in the CQC-regulated independent sector, not just those locations in ASC-WDS
- Provide stable trends over the history of our dataset
- Provide more accurate estimates of non-residential establishments compared to our previous model
- Examine our data over time, not as isolated years

We will soon be able to:

- Provide monthly filled posts estimates at regional, ICB, and local authority levels
- Improve our historical estimates as new information comes to light (e.g. a new establishment entering data into ASC-WDS for the first time will have improved estimates within a month of submission)

In the long term we want to:

- Produce more frequent updates for more of our data across more sectors

Limitations

- Historical estimates will improve over time as more data is submitted, so historical figures will be changing regularly.
- We are still refining our process. These estimates are still statistics under development so processes and therefore figures may change between now and when we use them for our publications in 2025.

Skills for Care
West Gate
6 Grace Street
Leeds
LS1 2RP

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

skillsforcare.org.uk

twitter.com/skillsforcare
facebook.com/skillsforcare
linkedin.com/company/skills-for-care