International Journal of Population Data Science

Journal Website: www.ijpds.org

Data Note: Challenges when combining housing data from multiple sources to identify overcrowded households

Laura Scott[1,*], Yan Weigang[1], Marcella Ucci[2], and Jessica Sheringham[3]

Submission History

Submitted:	11/10/2024
Accepted:	07/04/2025
Published:	20/05/2025

[1] London Borough of Islington, Public Health Directorate, London N1 2UD, England
[2] University College London, Bartlett School Env, Energy & Resources, London WC1E 6BT, England
[3] University College London, Institute of Epidemiology and Healthcare, London WC1E 6BT, England

Abstract

Background

This project in one urban local authority in London (England) sought to assess the feasibility of generating locally-derived indices of overcrowding using data available to local councils on the population and their homes.

We merged data at household level using the Unique Property Reference Number from publicly available Energy Performance Certificates and commercial property platforms, with data available to councils on the population and their housing characteristics, drawn from multiple sources including council tax bands and council housing databases. Multiple imputation was used to address missing data. Using the dataset, it was possible to generate two indices of overcrowding for households with dependent children, based on the bedroom standard and the space standard, which could be compared with nationally derived estimates.

Data challenges

We encountered three challenges with data. 1. Individuals in the population were excluded through linkage with household-level data. 2. Definitions of overcrowding are ambiguous and variably applied. 3. Many local areas face high proportions of missing household data, particularly numbers of bedrooms. We discuss how we addressed such problems and illustrate with a local example how they could affect estimates of overcrowding prevalence.

Lessons learned

Further clarity is needed in how bedrooms are defined to compare overcrowding prevalence generated locally and nationally. Access to national records on bedroom numbers would facilitate local areas to identify overcrowding in their own populations. Despite these challenges, we demonstrate it is feasible to generate overcrowding indices that can be useful for researchers and local policy makers seeking to develop or evaluate strategies to address household overcrowding.

Keywords

administrative data; data linkage; public health; housing; overcrowding

*Corresponding Author:
Email Address: j.sheringham@ucl.ac.uk (Jessica Sheringham)

https://doi.org/10.23889/ijpds.v6i1.2927

Introduction

Local areas are seeking to identify household overcrowding in their populations to address limitations with national data [1]. In the UK, national overcrowding data based on census surveys undertaken every 10 years may become inaccurate over time [2]. Local policy makers also need more granular data to understand which households are affected to inform the design, location or targeting of services. There is little information available for a typical council about the data requirements and processes needed to generate local overcrowding indices, and what could affect the usefulness and interpretation of any index produced.

Our aim was to assess the feasibility of combining publicly available and council-held data on households and their housing conditions, to identify households in overcrowding. Because children and their families can be especially affected by overcrowding [3, 4], we sought to focus particularly on households with dependent children. We conducted this work in Islington, an urban borough in London, England.

Overcrowding: definitions

Overcrowding broadly encompasses situations where the number of occupants exceeds the capacity of the dwelling space available [5]. It is variously defined. Guidance in England recommends the application of the 'bedroom standard' for councils to allocate housing [6]. Homes with fewer bedrooms than a household requires are considered overcrowded. Overcrowding can also be identified by the space available to residents. The English Government has produced minimum space standards for new homes that take account of the type of building and the number of inhabitants [7].

Data sources

To determine if a household is overcrowded, we needed data on the population within a household (household composition), and characteristics of their home (housing characteristics).

Household composition of the population

Islington council linked key council administrative data such as council tax, electoral register and data from council services such as Adult Social Care, Education, Early Help and Housing. This created a unique record for every resident, which is a record created from 19 large databases held in Islington that contain person records. After substantial matching and cleaning using the commercial data matching tool, Holistix Data Hub (QES) [8], a unique record per person was created. We verified these data to ensure they were current residents, or children of current residents by confirming for every record that an adult has a live relationship with the council (using ten criteria e.g. current council tax account), and that each child is a relation of one of these adults. Each person was assigned to their current address on the Local Land and Property Gazetteer. Total numbers of adults and numbers of children with age and gender were then extracted for this research.

Housing characteristics: number of bedrooms and floor area

We list in Supplementary Data (Table S1) the sources we identified as potentially relevant to provide data directly on bedroom number or floor area of residential homes or sources that could be used to impute these variables.

Methods

All analysis except where otherwise specified was conducted in Stata 18 [9].

Data integration

We selected LLPG records that related to all "residential" properties in Islington (n=117,707), to which available data from the council on bedroom number (n = 59,199), council tax band (n = 107,549) and housing tenure (n = 117,707) were added. The authors merged these data at individual household level using Unique Property Reference Number (UPRN), a unique identifier at dwelling level to selected EPC variables (habitable rooms, total floor area, building construction age, type of property (e.g. flat or house)). The authors then merged property records to a household-level file of the Islington population register (Figure 1).

Cleaning of records

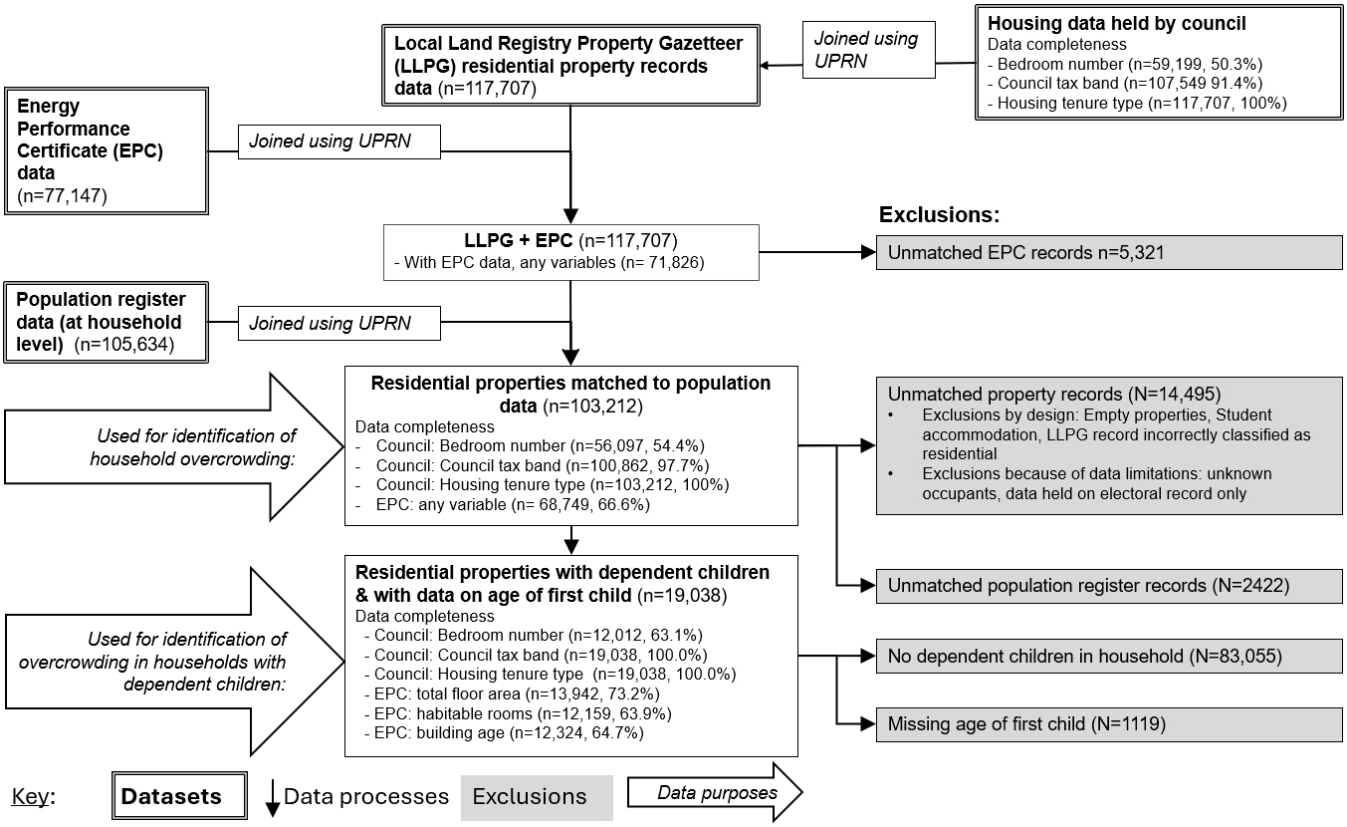
The cleaning of the records held within the council housing database, other council systems and the LLPG are conducted on an ongoing basis by the council as part of its data quality programme.

We examined the quality and credibility of EPC records by describing the range of values for each variable to identify outliers, and to identify non-credible values. Values were considered non-credible if they were outside of the expected range for the data (e.g. floor area recorded as $1\,m^2$) or where data did not align with other data sources, (e.g. where total numbers of habitable rooms were less than numbers of recorded bedrooms). Where there was a discrepancy between council-held data and EPC, we considered council-held data as more likely than EPC to be correct because the council relies on the accuracy and currency of this information to conduct housing repairs and to allocate housing to new tenants. Local intelligence was used in some cases (e.g. knowledge that a building had been recently converted to flats) which enabled the authors to infer that EPC habitable room and floor area values corresponded to an entire building, rather than one of the dwellings within it. Spot checks on floor plans for dwellings listed on publicly available housing data sites were also carried out to verify or correct EPC-recorded habitable rooms data where possible. As a result of these processes, a total of 2,500 (3%) of EPC values were either removed or corrected.

Processing

After merging, cleaning and restricting to households with dependent children and known age of first child, there were 19,038 households in the dataset with bedroom number known for 63% (Figure 1).

Figure 1: Flowchart describing the process for joining datasets, data completeness and data exclusions to identify overcrowding at household level



We therefore sought to impute missing data. Research has shown bedroom number is associated with building age, type of property, tenure, habitable room number and council tax bands [10, 11]. To inform a model for imputation, correlations between numerical candidate variables with bedroom number were explored. All were above the 'rule of thumb' of $\geq$0.5 recommended for inclusion in multiple imputation models (range: 0.55–0.91, Supplemetary Data, Table S2) [12].

MICE (multiple imputation by chained equation) [12] was applied to predict bedroom number using R 4.3.1. [13] The proportion of missing data was too high in the entire dataset to impute values for bedroom number for all properties (Figure 1) [14]. We therefore restricted imputation to households with dependent children where the data were more complete. The skewed variables - total floor areas, number of habitable rooms and bedroom number - were transformed by using the square root of these variables. The processed dataset had five imputations with 10 iterations. We sought to check the validity of our imputations using the URPN to identify the known addresses of a random selection of properties of imputed bedroom numbers. We searched publicly available data to compare reported bedrooms against imputed bedrooms (n = 50). Where exact bedroom data were available on publicly available data (55% of sample), imputed figures agreed in 75% of cases. Where publicly available information gave a range of possible bedroom numbers based on neighbouring properties (e.g. homes in the same street could have 1 or 2 bedrooms), imputed values were within the range of possible values in 100% of cases. A second validation check was undertaken to compare imputed values with known values for flats that were in the same housing block or houses in the same street where it might be inferred that bedroom numbers should be similar. This was accurate in 87.5% of cases. The imputed results also showed a similar distribution to observed bedroom number (Supplementary Data, Figure S1).

We used the data to identify household overcrowding using the ONS bedroom standard to illustrate the similarities and contrasts between overcrowding estimates using different data sources [2]. Locally-derived prevalence using known bedroom figures (13.1%) was higher but similar to nationally-derived figures (9.4%) for the council's population.

Discussion

We have described an approach to identifying overcrowding at household level and demonstrated how such estimates compare with nationally derived figures. We discuss below three problems we encountered that relate to three myths about data described by Christen and Schnell (2023) [15], i.e.:

1. The population is easily defined
2. Data definitions are unambiguous
3. Data are complete on all variables

The population covered in a database is not easily defined

We experienced problems in defining and identifying the denominator population (i.e. residential, households in

Islington with dependent children). We had to exclude 12% of records because the household LLPG file did not match the population register file. We describe below what led to exclusions of data.

Our local knowledge suggested that LLPG residential property coding overestimated the number of properties, where errors in the LLPG led to dwellings being classified as active residential dwellings when in fact they were empty or not used for residential purposes. Some exclusions upon merging with population data were therefore expected and intended. However, we may have 'lost' families living in properties that were incorrectly classified as non-residential.

We were also aware our population register did not capture all residents. Due to data protection permissions, we did not have access to data on residents on the closed electoral register. We relied on the open electoral register or where individuals were known to the council because of using other council services. This could lead to an underestimate of overcrowding because adults that are only on the closed electoral register cannot be counted in the total count of household members.

Definitions of overcrowding are varied and inconsistent

In the Introduction we described variations in definitions of overcrowding.

In operationalising the bedroom standard, we encountered ambiguities and inconsistencies in how a bedroom is defined. There are implications of this ambiguity. Firstly, it limits data integration possibilities. We had to exclude bedroom number data from privately rented stock because different definitions of bedroom were applied compared with council housing stock (Table S1). Secondly, it may explain why in our example (S3) locally derived figures were higher than figures from the census. The census survey 2021 defined a bedroom as "*any room that has been permanently converted for use as a bedroom*," which may include rooms originally built as living rooms. The ambiguity caused by living rooms used as bedrooms is recognised as more common in London and more common in smaller properties [16].

Any efforts to identify overcrowding need to be informed by local policies on housing allocation. In some councils, living rooms are explicitly included in the number of bedrooms, which would result in lower overcrowding prevalence [17]. ONS defines dependent children as anyone aged 20 or under; children over this age would be entitled to their own room. In contrast Islington Council defines dependent children as under 19 and some councils only include children aged under 17 years [18]. These local differences in bedroom entitlements would affect overcrowding prevalence estimates and ensuing policy.

Information was not available on key variables for all records in the database

Islington Council held data on bedroom number on council-held housing stock owned or previously owned by the council (e.g. properties sold under Right-to-Buy). However, this accounted for only 54.4% of residential properties (Figure 1) and largely did not cover housing association, privately-rented homes or owner-occupied properties.

We sought bedroom number data from other sources, but it was either not permitted to share such data or definitions of bedrooms were too different to be comparable to council-held data (Table S1). In particular, Valuation Office Agency (VOA) is the most complete source of bedroom number data nationally but VOA are not permitted to disclose this information to local councils.

We addressed missing data by using multiple imputation. This was made possible by the availability of EPC information on 66.6% of Islington property records. EPC records are becoming an increasingly useful data source to understand populations at household level. However, as others have reported, the quality of EPC records varies [19]. This is in part due to the age of reports; some variables have only been required on more recent reports and are therefore missing in older reports and in some cases, a property has been changed since an EPC inspection was conducted. In our experience, extensive checking and local knowledge was required to identify implausible EPC values and correct values that were not congruent with council-held data.

Imputed figures need further validation before using to inform strategies that require accurate bedroom numbers at household-level. Imputation may not be feasible for other local authorities with a smaller proportion of housing with known bedroom numbers.

Conclusions: recommendations and lessons

We have described an approach to generating locally-derived indices of overcrowding and demonstrated how such estimates compare with nationally derived figures. Methodological differences between generating local and national estimates using the bedroom standard estimates affect estimates of overcrowding, and – in turn – policies informed by them.

We encountered challenges in identifying denominator populations, variations in definitions of overcrowding and missing data on the population's household characteristics. Despite these challenges, the project illustrates the value of 'real-world' data sources for local population research and evaluation.

Below we provide recommendations and lessons learnt for other areas.

1. Locally-derived indices of overcrowding are feasible to generate, with some caveats, and can offer distinct benefits to nationally-derived indices, e.g. provide sufficient granularity to inform local policy. Nationally derived data are useful to enable comparison with other geographies.

2. Further clarity is needed locally and nationally on what counts as a bedroom for calculating occupancy using the bedroom standard.

3. Publicly available Energy Performance Certificate data is a valuable resource for calculating household overcrowding. Examining the ranges of EPC-recorded values to identify outliers and comparing values across multiple sources where possible is essential to identify

and remove non-credible values to make EPC data fit for purpose.

4. Sharing data between housing providers or making available nationally-derived data on bedroom number (e.g. collated by the Valuation Office Agency in England) would enable councils to derive timely estimates of overcrowding to inform the development and evaluation of local policies to improve wellbeing.

Household overcrowding in the UK (and other developed nations) is only one of the consequences of the lack of supply of affordable homes. Other consequences include rising numbers of people becoming homeless or living in temporary accommodation, both of which are harmful to individuals and are extremely costly for the public sector [20]. Accurate, comprehensive and accessible information on dwelling sizes (room numbers and floor area, with standardised definitions relevant to wellbeing), in relation to household numbers and demographic composition, is needed for overcrowding research. It is also essential for wider housing and wellbeing research that can inform policies to address problems caused by a lack of housing supply.

Acknowledgements

We thank Mahnaz Shaukat, Head of Data and Insights, Islington Council for valuable advice throughout the project and on this paper.

We would like also to extend thanks to attendees of a knowledge exchange workshop from local authorities and housing associations held in September 2024 how helped us understand how identifying overcrowding could be useful in their context.

Funders

This project is funded by the NIHR Public Health Research (PHR) Award: NIHR154776 and by ActEarly, a grant from the UK Prevention Research Partnership (MR/S037527/1), which is funded by the British Heart Foundation, Cancer Research UK, Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Health and Social Care Research and Development Division (Welsh Government), Medical Research Council, National Institute for Health Research, Natural Environment Research Council, Public Health Agency (Northern Ireland), The Health Foundation and Wellcome.

JS is funded by an NIHR Population Health Career Scientist Award (NIHR303616).

The project was supported by NIHR ARC North Thames. The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

Ethics statement

This project did not require ethical approval as the data referred to were either publicly available or were routinely processed for previous local authority-based research. It did require a specific data protection impact assessment submission which was approved by Islington Council in September 2023.

Conflict of interests

None declared

Data availability statement

Data are not available for further analysis as this is not permitted under the data protection impact assessment approvals for the project.

References

1. Wilk M, Harper G, Liverani S, Firman N, Simon P, Dezateux C. "Inequalities in household overcrowding in an ethnically diverse urban population: a cross-sectional study using linked health and housing records," *International Journal of Population Data Science*, vol. 10, no. 1, 01/23 2025. https://doi.org/10.23889/ijpds.v10i1.2408

2. Office for National Statistics. Overcrowding and under-occupancy by household characteristics, England and Wales: Census 2021 ONS website 2023 [Online] Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/housing/articles/overcrowdingandunderoccupancybyhouseholdcharacteristicsenglandandwales/census2021 (accessed 11 October 2024).

3. Marsh R, Salika T, Crozier S, Robinson S, Cooper C, Godfrey K, et al. The association between crowding within households and behavioural problems in children: Longitudinal data from the Southampton Women's Survey. Paediatric and Perinatal Epidemiology. 2019;33(3):195-203. https://doi.org/10.1111/ppe.12550

4. Solari CD, Mare RD. Housing crowding effects on children's wellbeing. Social Science Research. 2012;41(2):464–76. https://doi.org/10.1016/j.ssresearch.2011.09.012

5. Shannon H, Allen C, Clarke M, Dávila D, Fletcher-Wood L, Gupta S, et al. Web Annex A: Report of the systematic review on the effect of household crowding on health. WHO housing and health guidelines. 2018. Available: https://apps.who.int/iris/handle/10665/275838.

6. Ministry of Housing, Communities and Local Government. Allocation of accommodation: guidance for local housing authorities in England. In: Ministry of Housing, Communities and Local Government, editor. London2021, updated 2024. Available: https://www.gov.uk/guidance/allocation-of-accommodation-guidance-for-local-authorities.

7. Ministry of Housing, Communities and Local Government. Technical housing standards – nationally described space standard. In: Ministry of Housing, Communities and

Local Government, editor. 2015. [Online] Available: https://www.gov.uk/government/publications/technical-housing-standards-nationally-described-space-standard/technical-housing-standards-nationally-described-space-standard#using-the-space-standard.

8. Anon. QES Data Hub 2024. [Online] Available from: https://www.qes-online.co.uk/data-hub (accessed 11 October 2024).

9. *Stata Statistical Software: Release 18*. (2023). StataCorp LP, TX.

10. Özer S, Jacoby S. Dwelling size and usability in London: a study of floor plan data using machine learning. Building Research & Information. 2022;50(6):694-708. https://doi.org/10.1080/09613218.2022.2070452

11. Fone DL, Dunstan F, Christie S, Jones A, West J, Webber M, et al. Council tax valuation bands, socio-economic status and health outcome: a cross-sectional analysis from the Caerphilly Health and Social Needs Study. BMC Public Health. 2006;6(1):115. https://doi.org/10.10.1186/1471-2458-6-115

12. Nguyen CD, Carlin JB, Lee KJ. Practical strategies for handling breakdown of multiple imputation procedures. *Emerg Themes Epidemiol*, 2021. 18(1):5. https://doi.org/10.1186/s12982-021-00095-3

13. *R: A language and environment for statistical computing*. (2023). R Foundation for Statistical Computing, Vienna, Austria. [Online]. Available: https://www.R-project.org/.

14. Jakobsen JC, Gluud C, Wetterslev J, Winkel P. When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts. BMC Medical Research Methodology. 2017;17(1):162. https://doi.org/10.1186/s12874-017-0442-1

15. Christen P, Schnell R. Thirty-three myths and misconceptions about population data: from data capture and processing to linkage. International Journal of Population Data Science. 2023;8(1). https://doi.org/10.23889/ijpds.v8i1.2115

16. Office for National Statistics. Estimating the number of rooms and bedrooms in the 2021 Census: An alternative approach using Valuation Office Agency data. Undated. Available from: https://www.ons.gov.uk/census/planningforcensus2021/questiondevelopment/housingandcommunalestablishments (accessed 5 October 2024).

17. anon. Overcrowding undated [cited 2024 6 October 2024]. Available from: https://www.southwarkhomesearch.org.uk/content/Information/Overcrowding. (accessed 6 October, 2024).

18. anon. Bedroom entitlement 2022. [Online] Available from: https://www.peterborough.gov.uk/residents/housing/social-housing/bedroom-entitlement (accessed 6 October 2024).

19. Jenkins D, Simpson S, Peacock A. Investigating the consistency and quality of EPC ratings and assessments. Energy. 2017;138:480–9. https://doi.org/10.1016/j.energy.2017.07.105

20. Cromarty H, Greaves F, Rankl F, Barton C. Affordable housing in England. (2024). [Online] Available: https://commonslibrary.parliament.uk/affordable-housing-in-england/ (accessed 10 March 2025).

21. GeoPlace. What is an LLPG? 2024. Available from: https://www.geoplace.co.uk/local-authority-resources/guidance-for-custodians/how-to/about-the-role/what-is-an-llpg (accessed 5 October 2024).

22. Department for Levelling Up Housing, and Communities. "Official Statistics from DLUHC." https://opendatacommunities.org/home (accessed 19 December, 2024).

23. The Valuation Office Agency. How domestic properties are assessed for Council Tax bands 2016, updated 2024 [5 October 2024]. Available from: https://www.gov.uk/guidance/understand-how-council-tax-bands-are-assessed. (accessed 5 October 2024).

24. Department for Levelling Up Housing, and Communities, "English Housing Survey 2020 to 2021: headline report," 9 Dec 2021. [Online]. Available: https://www.gov.uk/government/statistics/english-housing-survey-2020-to-2021-headline-report (accessed 5 October 2024).

25. MetaStreet. Untitled 2024. Available from: https://metastreet.co.uk/features.html accessed 5 October 2024).

26. Ministry of Housing, Communities and Local Government. Selective licensing in the private rented sector: a guide for local authorities. 2023. [Online] Available: https://www.gov.uk/government/publications/selective-licensing-in-the-private-rented-sector-a-guide-for-local-authorities/selective-licensing-in-the-private-rented-sector-a-guide-for-local-authorities (accessed 5 October 2024).

Abbreviations

EPC:	Energy Performance Certificates
LLPG:	Local Land Property Gazetteer
ONS:	Office for National Statistics
UPRN:	Unique Property Reference Number
VOA:	Valuation Office Agency
LA:	local authority
LSOA:	lower level super output area
NIHR:	National Institute of Health and Care Research

Supplementary Data

Table S1: Potential data sources for property/dwelling information

Data source	Description	Purpose in this study
Definition of households		
Local Land and Property Gazetteer (LLPG) (n = 117,707 residential properties)	Each council is mandated to maintain an LLPG, a list of all addresses in its area (including houses, flats, parking spaces). Addresses are assigned a unique property reference, classified using a nationally standardised list, and collated in a national gazetteer (NLPG). [21]	Derive the sample frame, i.e. all residential properties within a local authority area
Bedroom number data		
Data that could be used		
Islington Council housing database (n = 38,836)	Data on currently and previously owned council stock in Islington: used for allocating properties to residents and maintaining homes.	'Gold standard' data on bedroom number. (A bedroom is defined as rooms designated as bedrooms when property was built)
Commercial property websites: historical data (n = 8,355)	Islington properties advertised on the platform for sale	Historical data on bedroom number (i.e. pre 2018)
Other dwelling characteristics		
Energy Performance Certificates (EPC) (n = 71,826)	Inspections conducted when a property is due to be rented or sold. The requirement for EPCs was introduced in stages from 2007. Apart from a few exempted buildings, a building must have an EPC when constructed, sold or let. Alongside energy performance, they include up to 92 variables related to characteristics of the dwelling and its surroundings. [22]	The following variables were selected to enable imputation of missing bedroom numbers and estimate space within homes: total floor area number of habitable rooms type of property building age tenure
Council-held: Council tax band (n = 100,862)	Each household is liable to pay a fee to their council for local services. The fee is in part determined by the valuation 'band' for their home. The Valuation Office Agency assesses the band of a property. Banding is based on a number of characteristics of the home, include size and layout. [23]	To enable the imputation of missing bedroom numbers.
Council: household tenure (n = 103,212)	Tenure covers whether a household rents or owns their property. For rented property it also covers whether the landlord is the council or a housing association or whether the household rents privately. Tenure is associated with overcrowding, and is associated with the type of properties and characteristics of households within them. [24]	Current data on tenure
Data we investigated but were not able to use or permitted to access		
Nethouseprice website download tool	Properties sold over a number of years.	Current data on bedroom number. Our request to access these data was turned down.
Commercial property websites: current data	Islington properties advertised on commercial property platforms for sale.	Downloading of data on bedroom number is no longer permitted from platforms

Continued

Table S1: Continued

Data source	Description	Purpose in this study
MetaStreet: Private rental sector register [25]	Database of the characteristics of Islington housing licensed for private rental under selective licensing scheme. [26]	Current data on bedroom number in private rental sector properties. Unable to use because bedroom count included living rooms and therefore could not be directly compared with council housing stock data.
Valuation Office Agency	The Valuation Office Agency collects and collates data on properties to provide the UK government with information needed to support taxation and benefits (e.g. council tax band, see above). This includes the numbers of bedrooms in a property. [23]	Current data on bedroom number. Our request to access these data was turned down because VOA were not permitted to disclose this information to local councils unless it is used in connection with a function of VOA or if there is a legal gateway established by an Act of Parliament.

Table S2: Correlation coefficients: independent numerical variables and bedroom number, used to confirm suitability of candidate variables for multiple imputation

	Number of bedrooms	Total floor area	Number of habitable rooms	Council tax band	Property type
Number of bedrooms	1.00				
Total Floor Area	0.80	1.00			
Number of Habitable Rooms	0.91	0.84	1.00		
Council Tax Band	0.56	0.69	0.59	1.00	
Property type	0.56	0.64	0.61	0.58	1.00

(n = 37,450)

Figure S1: Comparison of the distribution of imputed and actual bedroom number data