Datasheet for NHS RTT Waiting Times Dataset

Understanding Patient Access to Treatment in England

Chenika Bukes

November 26, 2024

Extract of the questions from Gebru et al. (2021).

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*
    - The dataset was created to monitor patient access to treatment within the NHS in England and support compliance with the 18-week waiting time target. It provides insights for healthcare planning, resource allocation, and policy development.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
    - The dataset was created and maintained by NHS England's statistical team under the Statistical Work Areas Division.
3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*
    - The dataset is funded and published by NHS England, as part of its public healthcare management responsibilities.
4. *Any other comments?*
    - This dataset is crucial for evaluating the efficiency of NHS services and identifying areas requiring intervention.

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

    - The instances represent monthly statistics on NHS patient waiting times for treatment in England. They include data on the number of patients waiting for treatment and the number who received treatment within the reporting period.

2. *How many instances are there in total (of each type, if appropriate)?*

- There are monthly instances of data on RTT waiting times from August 2007 until October 2024.

3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*

    - The dataset includes all NHS Trusts and Clinical Commissioning Groups (CGG) in England, making it comprehensive.

4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*

    - Each instance consists of structured data fields for 15 treatment types (e.g. cardiology, neurology) which include the number of patients waiting, the number treated within specific timeframes (e.g., within 1 week, 2 weeks, ..., up to 52 weeks), and the median waiting times in weeks for admission. These fields provide a detailed breakdown of patient wait times and treatment durations.

5. *Is there a label or target associated with each instance? If so, please provide a description.*

    - No.

6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*

    - Almost all months had reporting from all NHS locations. In October, November, and December 2022, data was missing from Frimley Health NHS Foundation Trust (RDU) and Manchester University NHS Foundation Trust (R0A). The reason for the missing data was not provided, but a supplemental time series instance was provided to account for the expected numbers from these NHS locations.

7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*

    - Not applicable.

8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*

    - No.

9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*

    - No.

10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

- Instances for years 2007 to 2010 are located on the national archive site which is publicly available.
    - Instances for years 2010 to 2024 the dataset is self contained on the NHS website.

11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*

    - No, all data is anonymized and aggregated to prevent identification of individuals.

12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*

    - No, the dataset contains operational statistics.

13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

    - No.

14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*

    - No, all data is aggregated and anonymized.

15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*

    - No.

16. *Any other comments?*

    - Not Applicable.

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

    - The data was collected via NHS Trusts and Clinical Commissioning Groups (CCGs), submitted monthly to NHS England.

2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*

    - NHS Trusts used internal systems to track referral-to-treatment pathways and submitted the data through standard reporting templates.

3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*

    - Not Applicable.

4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*

    - NHS staff at Trusts and CCGs submitted the data. RTT data collection was conducted by Unify2 until March 2018 and then by the NHS Digital's Strategic Data Collection Service.

5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*

    - Data collection began in August 2007 and continues monthly.

6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

    - Ethical review was not required as the dataset contains aggregated, anonymized data.

7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*

    - The data was collected directly by NHS Trusts and CCGs.

8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

    - No.

9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

    - No.

10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

    - No.

11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*

    - Yes, NHS England adheres to GDPR and data protection standards.

12. *Any other comments?*

    - The dataset is a vital tool for improving healthcare delivery and policy.

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*

    - Yes, the data was validated and cleaned by NHS England prior to publication. This process involved filtering treatment categories to ensure consistency and accuracy. The median waiting time was calculated, and the data was aggregated by waiting times in weekly intervals. Each column represents the number of patients who were treated within a specific number of weeks (e.g., 1 week, 2 weeks, ..., 52 weeks) during the reporting month, providing a structured and comprehensive view of patient wait times.

2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*

    - The raw data is not publicly available.

3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*

    - No.

4. *Any other comments?*

    - Not applicable.

Uses 1. *Has the dataset been used for any tasks already? If so, please provide a description.* - Yes, it is widely used for monitoring NHS performance and research. 2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.* - No 3. *What (other) tasks could the dataset be used for?* - Resource allocation, patient flow modeling, and healthcare policy analysis. 4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?* - No. 5. *Are there tasks for which the dataset should not be used? If so, please provide a description.* - It should not be used to infer individual patient outcomes. 6. *Any other comments?* - Not applicable.

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*

    - Yes, it is publicly available.

2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*

    - Via the NHS England website for public use (NHS 2024c).
    - Via secure email to ministers and those cleared to attain pre-releases of the data (NHS 2024d).

3. *When will the dataset be distributed?*

    - It is updated on the website monthly.

4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*

    - Yes, distributed under Open Government License.

5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*

    - No

6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*

    - No

7. *Any other comments?*

    - Ministers and selected officials receive pre-release access to official statistics once they are in their final form prior to publication (NHS 2024d).

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*

    - NHS England Statistical Work Areas Division.

2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*

- Via the NHS England website contact page or at the dmail: england.rtt@nhs.net

3. *Is there an erratum? If so, please provide a link or other access point.*

   - Errata are published with updates when necessary.
   - An example erratum for October, November, December 2022 is on the 2022-2023 RTT page (NHS 2024a).
   - All updates to guidance is tracked on the RTT-Guidance page (NHS 2024b).

4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*

   - Revisions are published periodically (usually every six months) in line with NHS England Analytical Service team's revisions policy.

5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*

   - No retention limits are specified.

6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*

   - Historical data until 2010 accessible on the site.
   - Archived data until 2007 is available on the national archive site (Health 2012).

7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*

   - No external contributions are accepted.

8. *Any other comments?*

   - Not applicable

References

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. "Datasheets for Datasets." *Communications of the ACM* 64 (12): 86–92.

Health, Department of. 2012. *Referral to Treatment Waiting Times Statistics*. https://webarchive.nationalarchives.gov.uk/ukgwa/20130107105354/http://www.dh.gov.uk/en/Publicationsandstatistics/Statistics/Performancedataandstatistics/ReferraltoTreatmentstatistics/index.html.

NHS. 2024a. *Consultant-Led Referral to Treatment Waiting Times Data 2022-23*. https://www.england.nhs.uk/statistics/statistical-work-areas/rtt-waiting-times/rtt-data-2022-23/ .

———. 2024b. *Consultant-Led Referral to Treatment Waiting Times Rules and Guidance*. https://www.england.nhs.uk/statistics/statistical-work-areas/rtt-waiting-times/rtt-guidance/ .

———. 2024c. *Referral to Treatment (RTT) Waiting Times*. https://www.england.nhs.uk/statistics/statistical-work-areas/rtt-waiting-times/.

———. 2024d. *Statistics Code of Practice Compliance*. https://www.england.nhs.uk/statistics/code-compliance/#Unifypolicy.