



A Dataset of Service Time and Related Patient Characteristics from an Outpatient Clinic

Haolin Feng ¹, Yiwu Jia ², Siyi Zhou ², Hongyi Chen ^{2,3} and Teng Huang ^{1,*}

¹ School of Business, Sun Yat-sen University, Guangzhou 510275, China

² Lingnan College, Sun Yat-sen University, Guangzhou 510275, China

³ Hangu TCM Innovation and Research Institute, Guangzhou 510627, China

* Correspondence: huangt258@mail.sysu.edu.cn

Abstract: Outpatient clinics' productivity largely depends on their appointment scheduling systems. It is crucial for appointment scheduling to understand the intrinsic heterogeneity in patient and service types and act accordingly. This article describes an outpatient clinic dataset of consultation service time with heterogeneous characteristics. The dataset contains 6637 consultation records collected from 381 half-day sessions between 2018 and 2019. Each record includes encrypted session and patient IDs, consultation start and (approximated) end times, the month and day of the week, whether it was on a holiday, the patient's visit count for a specific medical condition, gender, whether the consultation was cancer-related, and the distance from the patient's mailing address to the clinic. These features can be used to classify patients into heterogeneous groups in studies of appointment scheduling. Therefore, this dataset with rich, heterogeneous patient characteristics provides a valuable opportunity for healthcare operations management researchers to develop, test, and benchmark the performance of their models and methods. It can also be used for studying appointment scheduling in other service industries. More generally, it provides pedagogical value in areas related to management science and operations research, applied statistics, and machine learning.

Dataset: <https://github.com/fenghaolin/HanguData> (DOI:10.5281/zenodo.7444721)

Dataset License: CC-BY-NC-SA 4.0

Keywords: healthcare systems; appointment scheduling; outpatient clinic; patient heterogeneity; data-driven methods; machine learning



Citation: Feng, H.; Jia, Y.; Zhou, S.; Chen, H.; Huang, T. A Dataset of Service Time and Related Patient Characteristics from an Outpatient Clinic. *Data* **2023**, *8*, 47. <https://doi.org/10.3390/data8030047>

Academic Editor: Rüdiger Pryss

Received: 23 December 2022

Revised: 16 February 2023

Accepted: 23 February 2023

Published: 25 February 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Summary

Due to the rising demand for healthcare services and increased healthcare spending, healthcare organizations are under constant pressure to improve the efficiency of their operations [1,2]. Empirical evidence shows that operational efficiency is critical to the quality of patients' experience [3]. In addition, with the growing emphasis on patient-centered care, clinics, and hospitals are expected to improve the quality of care and guarantee access to care [4]. To address such challenges, various techniques have been proposed to optimize healthcare operations management (HOM) decisions such as nurse staffing and surgery planning [5], as well as outpatient appointment scheduling [6–9].

Outpatient appointment scheduling (AS) is a crucial aspect of delivering effective and efficient healthcare services to patients [10]. It focuses on finding appointment rules by optimizing a specific measure such as the weighted sum of patients' waiting time, physician overtime, and system idle time [11,12]. A well-managed appointment scheduling system can strike a balance between the productivity of patients and that of the healthcare providers [13].

Outpatient AS has been widely studied since the 1950s [14,15]. Some studies do not rely on observational data of service time but on theoretical probability distributions [16–19].

For those that do use data [20–23], their datasets are mostly withheld for privacy concerns [24]. As a result, publicly available datasets that could benefit the entire HOM research community are scarce. According to [24], no other U.S. healthcare system had made the wait times publicly available before them. Similarly, there is a shortage of publicly accessible real datasets on outpatient consultation service time. In fact, we have not found any besides the one we are making available.

We believe that real-world datasets describing the operations of healthcare service systems will benefit HOM researchers and practitioners, so we have made public this Hangu dataset containing service time and related feature information. This dataset holds practical value for management science and operations research (MS/OR) scholars working on healthcare appointment scheduling and the machine learning community. Here, we explain the value of the dataset:

- Accurate estimation of the service time is critical in designing efficient healthcare service systems. MS/OR studies on the operations of healthcare service systems typically model the uncertainty in service time by theoretical probability distributions or by approximating the empirical ones. Public datasets that are useful for the latter are rare. Our dataset is a realistic test bench for MS/OR researchers to evaluate their appointment scheduling algorithms and help develop new ones.
- Leveraging the rich feature information in the Hangu dataset, researchers can explore various ways of dealing with heterogeneity among patients, which helps determine the service time they need. Research shows that the larger the number of subgroups, the smaller the variability within each subgroup, thus the more predictable the service time becomes [21–23]. However, too many subgroups complicate the appointment scheduling decisions and make such decisions difficult to use in practice. One could explore the Hangu dataset together with similar ones and develop innovative methods for outpatient appointment scheduling.
- Researchers in the machine learning community can use the dataset to develop and test techniques for properly distinguishing patient types, as well as for accurately predicting their service time [25,26]. These are critical for analyzing the efficiency of a healthcare system [27].

2. Data Description

The data were obtained from our partner clinic, Hangu, a private outpatient clinic specializing in traditional Chinese medicine (TCM). Hangu was established by two experienced physicians formerly employed at a large public hospital in Guangzhou, China. It operates at multiple facilities in Guangzhou and Shenzhen, the two megacities in Guangdong province, China. While TCM physicians have their own specialties, it is typical that they treat patients with a variety of medical conditions, from minor illnesses to severe chronic conditions. Despite the differences in medical philosophy, the operation of Hangu mirrors that of modern Western medicine clinics. Patients make appointments through phone calls or online, and a clinic staff assigns them each an appointment slot for consultation. Currently, Hangu's appointment scheduling is simply on a first-come-first-appoint basis. Such a simple approach does not utilize heterogeneous patient characteristics, and it is a practice commonly adopted by many clinics [12,28,29].

This dataset pertains to the consultations provided by a stellar physician at a Guangzhou facility in 381 half-day sessions. The data cover all consultations provided by this physician in 2018 and 2019. Note that although Hangu operates six days a week, the physician in question, who is also a cofounder of Hangu, does not see patients every day due to other responsibilities within the clinic. Each session includes records of multiple patient consultations, resulting in a total of 6637 records in the final dataset stored in `Data.csv`. The details of data collection and processing are described in Section 3.

Table 1 provides a preview of the full dataset stored in `Data.csv`, and it contains all the consultations that occurred in a half-day session. Each row of Table 1 represents

one consultation record, and each column is a variable. We briefly explain the meaning of the records using the first row of Table 1 as an example:

- This record describes a consultation for a patient with ID 'HAA052B7CD'.
- The value of *Visit.No* is 7, indicating that the patient had seen the physician for the same medical condition six times before this visit.
- *M.Cancer* being 'TRUE' means the main condition to consult for was a type of cancer.
- *S.Cancer* being 'FALSE' means that the patient did not have other types of cancer other than the main condition.
- *StartTime* refers to the starting moment of the consultation.
- After being seen by the physician, the patient would make on-site payments at the front desk, and *PayTime* records the payment time. In the example row, the payment was made at 8:44:28.
- *Address* documents the patient's mailing address. In this example, the *Address* is 'Out of city', indicating that the patient resides outside the city of Guangzhou but within Guangdong province.
- *ServTime*, measured in seconds, is the derived service duration of the consultation. This example row shows that the derived service duration for this consultation was 691 s. The details of its derivation are provided in Section 3.

The meanings of the other variables are straightforward, and Table 2 provides the definition of all the variables in the data.

Table 1. Data of a sample session.

| ID | Session | Month | DayOfWeek | WorkingDay | AM_PM | Visit.No | Gender | M.Cancer | S.Cancer | StartTime | PayTime | Address | ServTime |
|------------|---------|---------|-----------|------------|---------|----------|--------|----------|----------|-----------|----------|-----------------|----------|
| HAA052B7CD | 1 | January | Wednesday | TRUE | morning | 7 | F | TRUE | FALSE | 8:31:40 | 8:44:28 | Out of city | 691 |
| HA18BDDC46 | 1 | January | Wednesday | TRUE | morning | 6 | F | FALSE | FALSE | 8:43:11 | 9:07:31 | In the city | 614 |
| HFC7DD5A0B | 1 | January | Wednesday | TRUE | morning | 2 | F | FALSE | FALSE | 8:53:25 | 9:08:38 | NA | 559 |
| HE10BEEB38 | 1 | January | Wednesday | TRUE | morning | 2 | M | FALSE | FALSE | 9:02:44 | 9:15:13 | Out of city | 749 |
| HBf11B62B6 | 1 | January | Wednesday | TRUE | morning | 10 | F | FALSE | FALSE | 9:26:19 | 10:01:57 | Out of city | 450 |
| H70AA1DE11 | 1 | January | Wednesday | TRUE | morning | 14 | M | FALSE | FALSE | 9:33:49 | 10:03:01 | Out of city | 744 |
| H019BB3DBB | 1 | January | Wednesday | TRUE | morning | 2 | M | FALSE | FALSE | 9:46:13 | 10:00:51 | Out of city | 295 |
| H12CF5C343 | 1 | January | Wednesday | TRUE | morning | 66 | F | FALSE | FALSE | 9:51:08 | 10:22:32 | NA | 1187 |
| HBDDb1EF1D | 1 | January | Wednesday | TRUE | morning | 8 | M | FALSE | FALSE | 10:10:55 | 10:40:17 | NA | 1461 |
| HE4EC70471 | 1 | January | Wednesday | TRUE | morning | 3 | F | FALSE | FALSE | 10:35:16 | 10:49:33 | NA | 403 |
| H834774031 | 1 | January | Wednesday | TRUE | morning | 2 | M | FALSE | FALSE | 10:41:59 | 11:07:48 | In the city | 572 |
| HD91C08D7D | 1 | January | Wednesday | TRUE | morning | 1 | M | FALSE | FALSE | 10:51:31 | 11:13:40 | NA | 1108 |
| H96BE60365 | 1 | January | Wednesday | TRUE | morning | 3 | M | TRUE | FALSE | 11:09:59 | 11:21:37 | In the city | 656 |
| HC26EECD08 | 1 | January | Wednesday | TRUE | morning | 6 | M | FALSE | FALSE | 11:20:55 | 11:33:30 | NA | 452 |
| H370DD4B95 | 1 | January | Wednesday | TRUE | morning | 2 | F | FALSE | FALSE | 11:28:27 | 11:43:23 | In the city | 500 |
| H9D1CC2F93 | 1 | January | Wednesday | TRUE | morning | 1 | F | TRUE | FALSE | 11:36:47 | 12:11:27 | Out of city | 1291 |
| H927913EA7 | 1 | January | Wednesday | TRUE | morning | 1 | F | FALSE | FALSE | 11:58:18 | 12:26:23 | NA | 1264 |
| H01EDA98AD | 1 | January | Wednesday | TRUE | morning | 7 | F | FALSE | FALSE | 12:19:22 | 12:39:53 | Out of Province | 1231 |

Table 2. Variable descriptions.

| Variable | Type | Description |
|------------|-------------|--|
| ID | Categorical | Encrypted version of the patient ID assigned by Hangu, unique for each patient |
| Session | Categorical | Identifier of a session that preserves actual chronological order |
| Month | Categorical | The calendar month of the record |
| DayOfWeek | Categorical | The day of the week of the record |
| WorkingDay | Logical | Whether it is a working day. Normally, weekdays are working days, while weekends are not. However, several exceptions exist due to public holidays' arrangement in China, which makes some weekends working days and some weekdays nonworking days. Please see Section 3 for more discussions. |
| AM_PM | Categorical | Whether it is a morning or afternoon session |
| Visit.No | Numerical | Indicate that this is the patient's x^{th} visit to this physician to treat the same medical condition |
| Gender | Categorical | Gender of the patient |
| M.Cancer | Logical | Whether the main purpose of this consultation is to treat cancer |
| S.Cancer | Logical | Other than the main purpose of the consultation, whether the patient has cancer |
| StartTime | Numerical | The start time of the consultation. It was recorded by the information system when the physician clicked the patient's electronic record and started the consultation. |
| PayTime | Numerical | It was recorded by the information system when the patients pay for the prescription (after the consultation). |
| Address | Categorical | Patient mailing address, can be 'In the city' (i.e., in Guangzhou), 'Out of city' (i.e., not in Guangzhou but in Guangdong province), or 'Out of province' (i.e., not in Guangdong province). 'NA' is used if no such information is provided to the clinic. |
| ServTime | Numerical | The service duration of the consultation measured in seconds |

3. Methods

We now describe how we obtained, processed, and analyzed the data. Section 3.1 explains the operational rather than clinical nature of the data, and it explains the usage approval obtained. Section 3.2 describes the collection and preprocessing of the raw data. Section 3.3 details the process of computing the derived service time. Section 3.4 introduces how to handle outliers. Section 3.5 presents the exploratory data analysis of the prepared data and insights about the results.

3.1. Data Usage Approval

The current study is an operational study rather than a clinical one. The data were collected retrospectively, and no medical records other than whether the visit was cancer-related were included in the data. The data were fully anonymized and handled within the regulations set by the clinic's administration. Because of the encrypted IDs, all identifiable information was removed, and it becomes fully untraceable. The administration of the clinic reviewed the research plan and the data involved, and they found no violation of any ethical principle in the study. Therefore, they granted the research team the approval to conduct the research and to use the data extracted from their information system (approval document No.20230119001).

3.2. Create the Raw Data Files

Authorized clinic personnel directly extracted the clinic's stellar physician's 2018 and 2019 consultation records from the clinic's information system. Then, the following steps were taken to protect patients' private information:

- The records were reviewed by the clinic and marked whether the consultation was related to cancer (the *M.Cancer* and *S.Cancer* in Table 2), and then the original diagnosis information was removed.

- Each patient was assigned a unique ID by Hangu, which was then encrypted during the creation of the raw dataset. We ensured that each patient's encrypted ID is unique, yet patients' true identities are deidentified after such encryption.
- The date of each session was removed. Potentially useful information related to the date was preserved in the following way:
 - The date information was turned into the variables *Month* and *DayOfWeek*.
 - We added a variable indicating whether a session occurred during a working day or a holiday. The dates of public holidays in China depend on the lunar calendar and the solar one; thus, the national holiday arrangement changes yearly. When tagging each session, we checked the 2018 and 2019 official public holiday arrangements [30,31].
 - Each session was given a unique ID between 1 and 381, which preserved the actual chronological order. Future data users can recreate all the sessions of the focal physician and the order and length of each patient's consultation.

Raw_1.csv was created following the above steps. In total, 6853 consultation-related records of the 381 half-day sessions are included in *Raw_1.csv*, and Table 3 provides a preview. Like the final dataset in *Data.csv*, each row represents one consultation record, and the meaning of the variables are explained both in Table 2 and in the text by example (c.f. Section 2).

The clinic also keeps records of a subset of patients' mailing addresses. *Raw_1.csv* includes 2469 unique patients, 1199 of whom have mailing addresses in the clinic's information system. We utilized JioNLP [32], a python package for Chinese NLP preprocessing, to automatically extract the city and province information from the original address. Then, we converted such information into the variable *Address* that takes one of the three values: {'In the city', 'Out of city', or 'Out of province'}. The meaning of these values is provided in Table 2. The encoded address records are stored in *Raw_2.csv*, and Table 4 is a preview of it.

3.3. Compute the Service Time

Note that Hangu's information system does not record the actual service duration. Here, we provide a way to compute it given available information in *Raw_1.csv*.

At the start of each consultation, the physician opens the patient's electronic record, which records the actual starting time, which may differ from the appointment time. We took the difference between the starting times of consecutive patients' consultations and named this quantity *D1*. When the following two assumptions are satisfied, *D1* is an accurate measurement of the service time:

1. The next patient was ready when the physician completed the previous patient's consultation.
2. The physician did not leave for other business or personal tasks during a session.

While patients' true arrival times were not tracked by the clinic, we can still assess whether the first assumption is a reasonable one. According to Hangu's staff, almost all patients arrived no later than their appointment time. This might be because after making an appointment, the patients get reminders one day before the appointment date and at an earlier time on the same day. Empirical studies also report that an overwhelming majority of patients arrive early rather than late [33]. Moreover, it was very rare that the physician would leave the consultation room for other tasks. Therefore, it is plausible to use *D1* as the service time.

Table 3. A preview of the raw data file Raw_1.csv.

| ID | Session | Month | DayOfWeek | WorkingDay | AM_PM | Visit.No | Gender | M.Cancer | S.Cancer | StartTime | PayTime |
|------------|---------|---------|-----------|------------|---------|----------|--------|----------|----------|-----------|----------|
| HAA052B7CD | 1 | January | Wednesday | TRUE | morning | 7 | F | TRUE | FALSE | 8:31:40 | 8:44:28 |
| HA18BDDC46 | 1 | January | Wednesday | TRUE | morning | 6 | F | FALSE | FALSE | 8:43:11 | 9:07:31 |
| HFC7DD5A0B | 1 | January | Wednesday | TRUE | morning | 2 | F | FALSE | FALSE | 8:53:25 | 9:08:38 |
| HE10BEEB38 | 1 | January | Wednesday | TRUE | morning | 2 | M | FALSE | FALSE | 9:02:44 | 9:15:13 |
| HBf11B62B6 | 1 | January | Wednesday | TRUE | morning | 10 | F | FALSE | FALSE | 9:26:19 | 10:01:57 |
| H70AA1DE11 | 1 | January | Wednesday | TRUE | morning | 14 | M | FALSE | FALSE | 9:33:49 | 10:03:01 |
| H019BB3DBB | 1 | January | Wednesday | TRUE | morning | 2 | M | FALSE | FALSE | 9:46:13 | 10:00:51 |
| H12CF5C343 | 1 | January | Wednesday | TRUE | morning | 66 | F | FALSE | FALSE | 9:51:08 | 10:22:32 |
| HBDDb1EF1D | 1 | January | Wednesday | TRUE | morning | 8 | M | FALSE | FALSE | 10:10:55 | 10:40:17 |
| HE4EC70471 | 1 | January | Wednesday | TRUE | morning | 3 | F | FALSE | FALSE | 10:35:16 | 10:49:33 |
| H834774031 | 1 | January | Wednesday | TRUE | morning | 2 | M | FALSE | FALSE | 10:41:59 | 11:07:48 |
| HD91C08D7D | 1 | January | Wednesday | TRUE | morning | 1 | M | FALSE | FALSE | 10:51:31 | 11:13:40 |
| H96BE60365 | 1 | January | Wednesday | TRUE | morning | 3 | M | TRUE | FALSE | 11:09:59 | 11:21:37 |
| HC26EECD08 | 1 | January | Wednesday | TRUE | morning | 6 | M | FALSE | FALSE | 11:20:55 | 11:33:30 |
| H370DD4B95 | 1 | January | Wednesday | TRUE | morning | 2 | F | FALSE | FALSE | 11:28:27 | 11:43:23 |
| H9D1CC2F93 | 1 | January | Wednesday | TRUE | morning | 1 | F | TRUE | FALSE | 11:36:47 | 12:11:27 |
| H927913EA7 | 1 | January | Wednesday | TRUE | morning | 1 | F | FALSE | FALSE | 11:58:18 | 12:26:23 |
| H01EDA98AD | 1 | January | Wednesday | TRUE | morning | 7 | F | FALSE | FALSE | 12:19:22 | 12:39:53 |

Table 4. First few rows of the address data.

| ID | Address |
|------------|-------------|
| HE5CAF8BAF | In the city |
| H0CCD91062 | In the city |
| H49AFCD775 | In the city |
| H70AEA9C2A | In the city |
| H7AB0B1885 | In the city |
| H1CD5DE959 | In the city |
| HCF080FFA6 | In the city |
| H238985184 | In the city |
| H55D7E3946 | Out of city |

However, when either assumption is violated, $D1$ would be an overestimation of the actual service time. For example, when the next patient was late, the first patient's $D1$ would contain the actual service time and the time that the physician waited for the next patient. To account for such events and improve the estimation accuracy, we further use the payment time, recorded in the variable $PayTime$, which only happens after the consultation. We calculated the difference between a patient's $PayTime$ and $StartTime$ and named it $D2$. When both $D1$ and $D2$ are available, the smaller one is a more accurate estimation of the service time. We define the service time as Equation (1).

$$ServTime = \min(D1, D2) \quad (1)$$

$D1$ is not available for the last consultation in each half-day session, since there would not be a subsequent patient. Occasionally, $PayTime$ is unavailable in the system, leading to missing values of $D2$. In either case, $ServTime$ was set to be equal to either $D1$ or $D2$ (depending upon their availability). $ServTime$ was encoded as 'NA' if both $D1$ and $D2$ were unavailable, and there are 28 such occurrences.

After adding $ServTime$ to the raw data of consultation records, we merged the resulting data with the address data in `Raw_2.csv` via the patients' encrypted ID. Missing mailing addresses were encoded as 'NA'. The merged data were saved in `Merged_ServiceTime.csv`.

3.4. Final Processing

In this subsection, we document how we identified the outliers and handled the missing values.

First, we identified 20 records with a derived service time longer than one hour ($ServTime > 3600$ seconds). Most of them are the last consultation in the half-day sessions. The clinic's manager explains that the last patient will occasionally make the payment later because the front desk leaves for lunch. Therefore, without a follow-up patient's starting time, a possibly delayed payment leads to a very large $ServTime$. In addition, 167 $ServTime$'s are less than three minutes. These short consultations may result from occasional abnormalities, such as the physician accidentally clicking a wrong patient's record or getting interrupted by another patient. We dealt with the abnormalities following the advice of the clinic's manager, i.e., service times larger than 60 min or less than 3 min were considered outliers. There are 187 records identified. Filtering out such outliers and the 28 missing values in $ServTime$, the remaining data have 6638 records left.

Lastly, we removed one consultation record with a missing value in the variable $Visit.No$.

After the above processing procedure, there are 6637 records remaining in the final dataset `Data.csv`.

Note that, in the final dataset, 2392 records have a missing value in Address (encoded as 'NA'). The reason for keeping the records with no address is that not providing a mailing address is a patient preference, and thus it carries information related to patient characteristics.

3.5. Exploratory Data Analysis

This subsection presents the exploratory data analysis results of the prepared dataset in `Data.csv`. All the numbers and figures can be reproduced by the Jupyter Notebook script `Data_process_stat.ipynb` in the Supplementary Materials. The same script can also reproduce the calculation of the derived service time in Section 3.3.

First, we provide a basic summary of each categorical/logical variable.

The number of consultations per session: The mean and median are 17.42 and 18, respectively, while the minimum and the maximum are 4 and 32, respectively.

Number of sessions per day of the week: Tuesday: 6; Wednesday: 177; Friday: 10; Saturday: 188.

The number of sessions on working vs. non-working days: 203 sessions happened on working days and 178 on nonworking days.

The number of morning- vs. afternoon- sessions: Morning: 191; Afternoon: 190.

The number of consultations grouped by *Gender*: Female: 3943; Male: 2694.

The number of consultations grouped by *M.Cancer*: 620 consultations have the variable *M.Cancer* being 'TRUE' and 6017 'FALSE'.

The number of consultations grouped by *S.Cancer*: 67 out of 6637 consultation records correspond to *S.Cancer* being 'TRUE'.

Next, we look at the numerical variables *Visit.No* and *ServTime*.

The mean and median of *Visit.No* are 6.83 and 2, and the minimum and the maximum are 1 and 170. Figure 1 visualizes ***Visit.No*'s distribution**. It is highly skewed to the right. While the majority of the records have their $Visit.No \leq 2$, there are more than 1000 records that have their $Visit.No \geq 10$. This means that a sizable amount of consultation records come from recurrent consultations for the same medical condition.

The distribution of *ServTime* can be found in Figure 2. The boxplot shows that its mean (a dot above the median line) is slightly larger than its median, and the distribution is moderately skewed to the right. While it is fairly common in the appointment scheduling literature that the service time is assumed to be exponentially distributed [17,18,34], the histogram in Figure 2 suggests otherwise, at least in the outpatient clinic in question. It will be interesting and value-adding to test the performance of those methods, developed under stylized assumptions, using realistic datasets such as ours.

We are interested in the service times and informative characteristics distinguishing them. Therefore, we explore the distribution of *ServTime* grouped by different categorical/logical variables. Figures 3–9 depict the derived service-time distributions grouped by *Month*, *WorkingDay*, *AM_PM*, *Gender*, *M.Cancer*, *S.Cancer*, and *Address*, respectively. From these figures, we can find that they have noticeable effects on the distributions of *ServTime*.

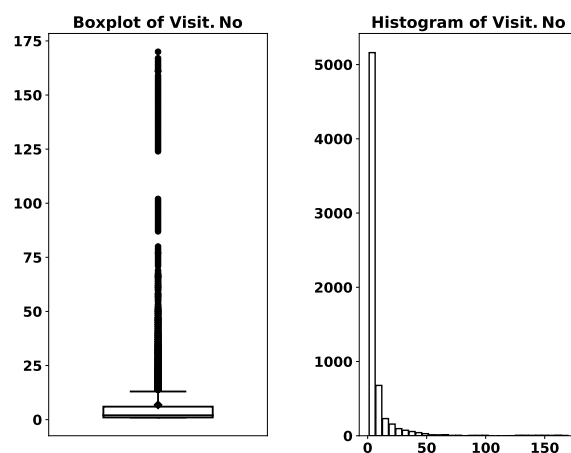


Figure 1. The boxplot and the histogram of *Visit.No*.

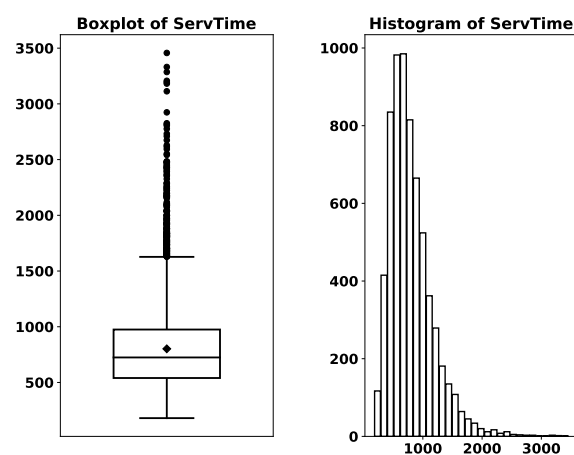


Figure 2. The boxplot and histogram of *ServTime*.

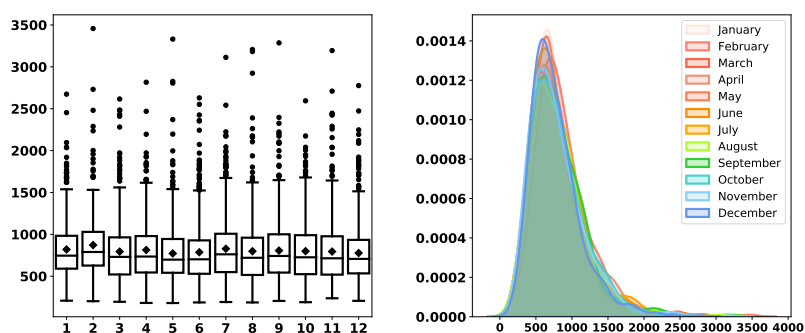


Figure 3. *ServTime* distribution grouped by *Month*.

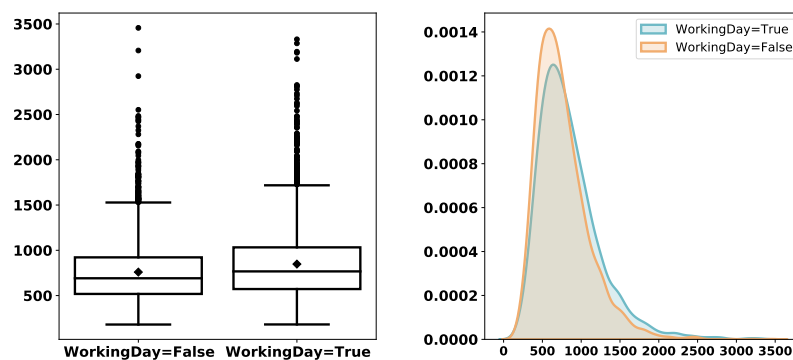


Figure 4. *ServTime* distribution grouped by *WorkingDay*.

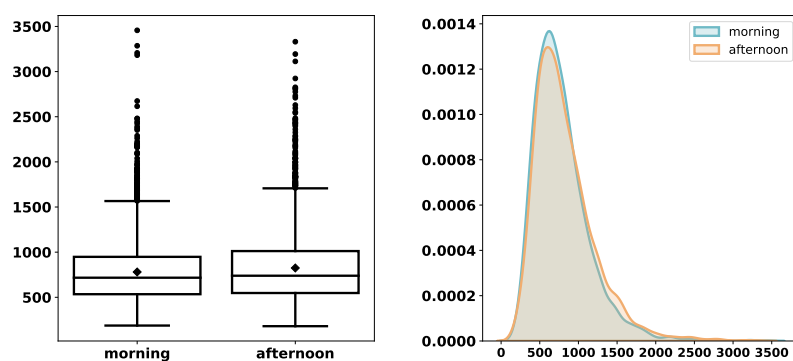


Figure 5. *ServTime* distribution grouped by *AM_PM*.

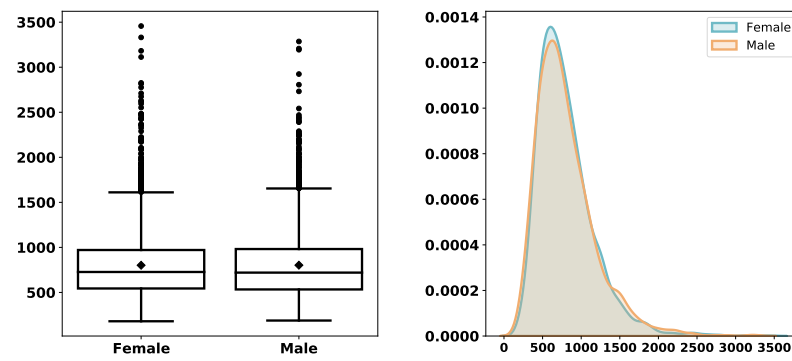


Figure 6. *ServTime* distribution grouped by Gender.

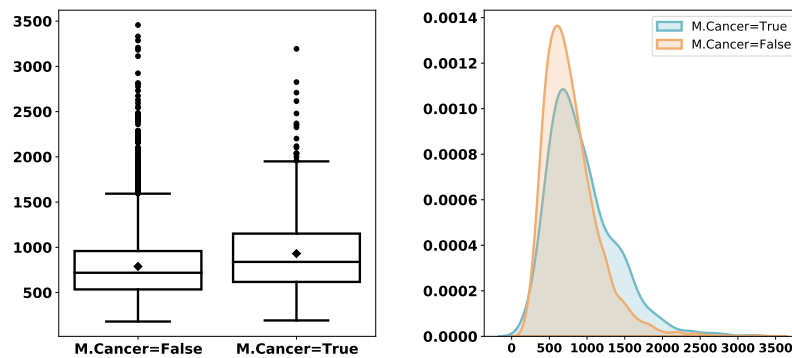


Figure 7. *ServTime* distribution grouped by *M.Cancer*.

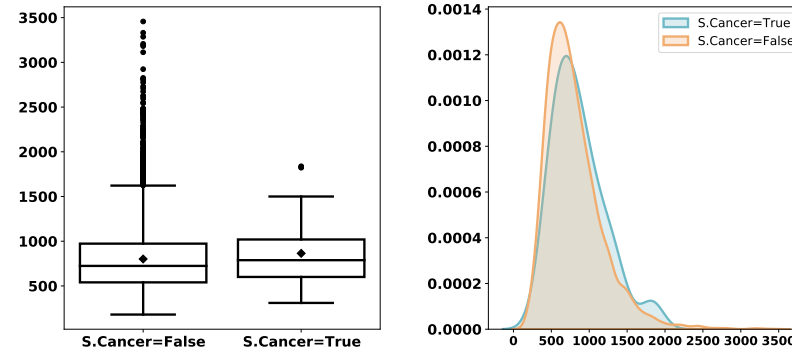


Figure 8. *ServTime* distribution grouped by *S.Cancer*.

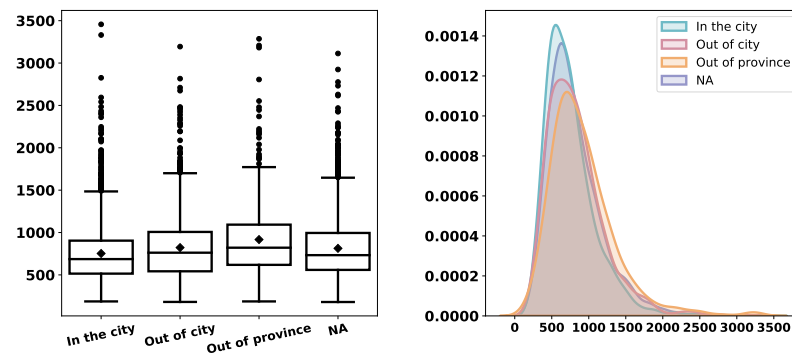


Figure 9. *ServTime* distribution grouped by Address.

Extant studies constantly find that the service time of a first-time visit and that of a follow-up one follow different distributions. Hence, we explore the distribution of *ServTime* based on whether *Visit.No* is equal to 1 or not. This essentially converts *Visit.No* into a

logical variable whose value depends on whether *Visit.No* is equal to 1. Figure 10 illustrates that first-time visits overall take a longer service time. Table 5 further compares the *ServTime* of *Visit.No* = 1 and that of *Visit.No* > 1 with different grouping. Overall, both means and medians of *ServTime* are larger in the case of *Visit.No* = 1, no matter which categorical variable is used for grouping.

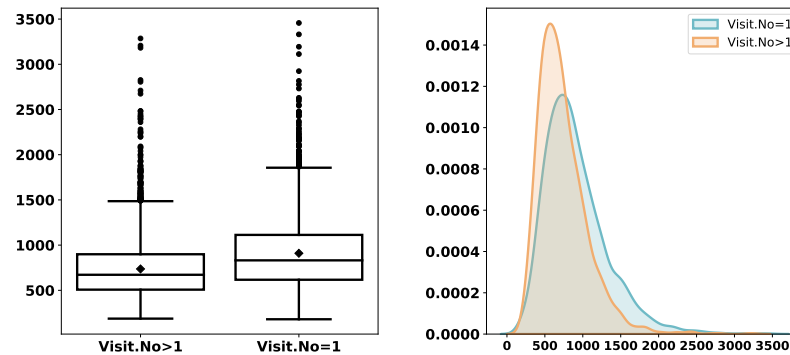


Figure 10. *ServTime* distribution grouped by *Visit.No*.

Table 5. Means and medians of *ServTime* in different groups.

| | Mean | | | Median | | |
|---------------------------|---------|---------------------|---------------------|---------|---------------------|---------------------|
| | Overall | <i>Visit.No</i> = 1 | <i>Visit.No</i> > 1 | Overall | <i>Visit.No</i> = 1 | <i>Visit.No</i> > 1 |
| WorkingDay = False | 759.3 | 858.8 | 698.0 | 691.0 | 799.0 | 642.0 |
| WorkingDay = True | 847.7 | 965.3 | 777.5 | 767.0 | 883.0 | 715.0 |
| Gender = F | 801.6 | 915.4 | 736.0 | 727.0 | 837.0 | 673.0 |
| Gender = M | 802.4 | 901.9 | 737.4 | 720.0 | 820.5 | 672.0 |
| AM_PM = afternoon | 824.7 | 924.8 | 753.0 | 739.0 | 840.0 | 680.0 |
| AM_PM = morning | 780.4 | 892.2 | 722.9 | 717.0 | 821.0 | 667.0 |
| M.Cancer = False | 788.6 | 885.4 | 726.5 | 719.0 | 816.0 | 661.0 |
| M.Cancer = True | 930.7 | 1282.8 | 815.4 | 839.0 | 1282.0 | 749.0 |
| S.Cancer = False | 801.3 | 908.2 | 735.9 | 724.0 | 831.0 | 671.0 |
| S.Cancer = True | 864.2 | 1194.4 | 784.7 | 788.0 | 1254.0 | 736.0 |
| Address = In the city | 752.8 | 846.3 | 711.0 | 686.5 | 764.0 | 650.0 |
| Address = Out of city | 822.4 | 941.6 | 776.0 | 762.0 | 868.0 | 713.0 |
| Address = Out of province | 917.2 | 1083.3 | 843.0 | 821.0 | 1012.5 | 762.5 |
| Address = NA | 813.0 | 914.2 | 704.5 | 733.0 | 840.0 | 649.5 |

Lastly, we regress *ServTime* on the rest of the variables. Table 6 summarizes the regression result. The coefficients of *Visit.No*, *M.Cancer*, *Address*, and *AM_PM* are statistically significant at the 0.001 level of significance. The coefficient of *Visit.No* is -1.12 . This means that an extra previous visit for the same condition, on average, reduces the service time by 1.12 seconds. The coefficient of *M.Cancer* is 137.51. This means that consultations mainly for cancer on average are 137.51 seconds longer than those mainly not for cancer. *Address* is a factor with four levels: {'In the city', 'Out of city', 'Out of province', and 'NA'}. 'In the city' is the baseline. The regression result suggests that the consultation of a patient with no address provided, on average, takes 46.67 seconds more than the baseline; the consultation of a patient from outside of the city but within the province, on average, takes 61.47 seconds more than the baseline; and the consultation of a patient from another province, on average, takes 152.43 seconds more than the baseline. The *Address* variable may be related to customer loyalty or illness severity. An intuitive interpretation is that a patient is unlikely to travel a long distance to consult for some minor health issue. The coefficient of *AM_PM_morning* is -49.37 . This suggests that, if other things are equal, on average, a consultation in the morning is 49.37 seconds shorter than its afternoon counterpart. Of course, all these interpretations are based on the linear regression model. Future users of the data should conduct a more comprehensive analysis to gain more insights.

Table 6. Regression results ($R^2 = 0.049$).

| | Coefficient | Std Err | p-Value |
|-------------------------|-------------|---------|-------------------------|
| Intercept | 724.99 | 47.75 | $<1 \times 10^{-3}$ *** |
| WorkingDay | 80.67 | 29.61 | 6×10^{-3} ** |
| Visit.No | −1.12 | 0.30 | $<1 \times 10^{-3}$ *** |
| M.Cancer | 137.51 | 15.79 | $<1 \times 10^{-3}$ *** |
| S.Cancer | −6.04 | 45.62 | 8.95×10^{-1} |
| Gender_M | −6.25 | 9.19 | 4.97×10^{-1} |
| Month_August | −5.60 | 23.07 | 8.08×10^{-1} |
| Month_December | −32.47 | 22.29 | 1.45×10^{-1} |
| Month_February | 44.57 | 26.72 | 9.5×10^{-2} |
| Month_January | 9.34 | 22.66 | 6.8×10^{-1} |
| Month_July | 15.86 | 22.81 | 4.87×10^{-1} |
| Month_June | −6.70 | 22.76 | 7.68×10^{-1} |
| Month_March | −7.97 | 22.12 | 7.19×10^{-1} |
| Month_May | −28.49 | 22.78 | 2.11×10^{-1} |
| Month_November | −12.83 | 22.94 | 5.76×10^{-1} |
| Month_October | −14.15 | 22.89 | 5.36×10^{-1} |
| Month_September | 4.01 | 23.23 | 8.63×10^{-1} |
| DayOfWeek_Saturday | 22.21 | 42.19 | 5.99×10^{-1} |
| DayOfWeek_Tuesday | 125.23 | 52.25 | 1.7×10^{-2} * |
| DayOfWeek_Wednesday | 25.79 | 31.40 | 4.11×10^{-1} |
| Address_NA | 46.67 | 10.83 | $<1 \times 10^{-3}$ *** |
| Address_Out of province | 152.43 | 18.32 | $<1 \times 10^{-3}$ *** |
| Address_Out of city | 61.47 | 12.28 | $<1 \times 10^{-3}$ *** |
| AM_PM_morning | −49.37 | 9.06 | $<1 \times 10^{-3}$ *** |

Note: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

4. User Notes

A major advantage of this dataset is the service time records that come with heterogeneous characteristics regarding patient demographics, medical conditions, and previous visit information. The exploratory data analysis here sheds some light on how some characteristics may affect the service time. Indeed, more in-depth analysis and feature engineering are demanded to make more accurate predictions of the service time. Making this dataset public provides valuable opportunities for HOM researchers. Here, we suggest some of them:

- To showcase the effectiveness and efficiency of newly developed models and methods for outpatient AS. For example, ref. [35] propose a data-driven approach to handle service-time heterogeneity in outpatient AS, in which this dataset is utilized to demonstrate the effectiveness of the proposed method.
- To enable research reproducibility. Studies focusing on developing new optimization methods for outpatient AS can report their methods' performance on public datasets such as this one in their numerical experiments, so that follow-up studies can reproduce and extend them.
- To enable fair performance comparison. Despite the rich literature on outpatient AS, existing studies typically withhold the data in use. It is difficult to make fair performance comparisons among different methods. Making real datasets publicly available can help solve this issue.

Note that the patients' scheduled appointment times and their no-show records are not available in this dataset. In the literature [6,9,12,16,17,33], researchers assume that a patient either does not show up or arrives punctually at the scheduled appointment time, and patients are served in the order of their scheduled appointments. One can assume that patients in our records exhibit similar behaviors to those mentioned in the literature. When studying no-shows, researchers make assumptions about hypothetical no-show rates in the absence of real no-show data. For example, ref. [34] simulates patient call-in sequences and assumes heterogeneous no-show rates in experiments.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/data8030047/s1>.

Author Contributions: H.F.: conceptualization, supervision, investigation, and formal analysis, writing—original draft; Y.J.: data curation and writing—original draft; S.Z.: data curation, formal analysis, visualization, and writing—original draft; H.C.: investigation, data curation; T.H.: project administration and writing—original draft. All authors have read and agreed to the published version of the manuscript.

Funding: Haolin Feng’s work is partially supported by the National Natural Science Foundation of China [grant numbers 72071217, 71721001]. Teng Huang’s work is supported by the National Natural Science Foundation of China [grant number 72101277] and the China Postdoctoral Science Foundation [grant number 2021M703664].

Institutional Review Board Statement: The work does not involve human subjects, animal tests, or cell lines. Experiments involve no human participants and are only numerical/computational. The data are fully anonymized and untraceable, and the data and the study have been reviewed and approved by the clinic’s administration (No.2023011901).

Informed Consent Statement: It is a retrospective study with the data extracted from the clinic’s information system. All identifiable information was removed, and it becomes fully untraceable. After reviewing the data being shared, the administration of the clinic confirmed that the data were fully anonymized and handled within the regulations set by the clinic’s administration. Therefore, the clinic approves the usage of the data (approval document No.20230119001).

Data Availability Statement: The data can be obtained from <https://github.com/fenghaolin/HanguData> (accessed on 22 February 2023).

Acknowledgments: The authors would like to thank Hangu’s administration for allowing the data to be shared publicly.

Conflicts of Interest: The funding support for the work has been acknowledged in the ‘Funding’ section. While the data reported here were collected during a collaborated project led by the first author (Haolin Feng) and the clinic investigating the appointment scheduling of the system, the authors do not receive any funding from the clinic. The authors declare that they have no known competing financial interests or personal relationships which have, or could be perceived to have, influenced the work reported in this article.

Abbreviations

The following abbreviations are used in this manuscript:

| | |
|-----|----------------------------------|
| HOM | Healthcare operations management |
| AS | Appointment scheduling |
| TCM | Traditional Chinese medicine |
| MS | Management science |
| OR | Operations research |

References

1. Marynissen, J.; Demeulemeester, E. Literature review on multi-appointment scheduling problems in hospitals. *Eur. J. Oper. Res.* **2019**, *272*, 407–419. [\[CrossRef\]](#)
2. Wang, L.; Demeulemeester, E. Simulation optimization in healthcare resource planning: A literature review. *IIE Trans.* **2022**, *in press*. [\[CrossRef\]](#)
3. Ko, D.G.; Mai, F.; Shan, Z.; Zhang, D. Operational efficiency and patient-centered health care: A view from online physician reviews. *J. Oper. Manag.* **2019**, *65*, 353–379. [\[CrossRef\]](#)
4. Youn, S.; Geismar, H.N.; Pinedo, M. Planning and scheduling in healthcare for better care coordination: Current understanding, trending topics, and future opportunities. *Prod. Oper. Manag.* **2022**, *31*, 4407–4423. [\[CrossRef\]](#)
5. Erdogan, S.A.; Denton, B.T.; Cochran, J.; Cox, L.; Keskinocak, P.; Kharoufeh, J.; Smith, J. Surgery planning and scheduling. In *Wiley Encyclopedia of Operations Research and Management Science*; Wiley: Hoboken, NJ, USA, 2011.
6. Klassen, K.J.; Yoogalingam, R. Improving performance in outpatient appointment services with a simulation optimization approach. *Prod. Oper. Manag.* **2009**, *18*, 447–458. [\[CrossRef\]](#)
7. Véricourt, F.d.; Jennings, O.B. Nurse staffing in medical units: A queueing perspective. *Oper. Res.* **2011**, *59*, 1320–1331. [\[CrossRef\]](#)

8. Zhang, Z.; Xie, X. Simulation-based optimization for surgery appointment scheduling of multiple operating rooms. *IIE Trans.* **2015**, *47*, 998–1012. [\[CrossRef\]](#)
9. Feng, H.; Li, Z.; Alvarado, M.M.; Colón-Morales, C.M. A simulation study of outpatient surgery clinic with stochastic patient re-entrance. In Proceedings of the 2020 Winter Simulation Conference (WSC), Orlando, FL, USA, 14–18 December 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 910–921.
10. Gupta, D.; Denton, B. Appointment scheduling in health care: Challenges and opportunities. *IIE Trans.* **2008**, *40*, 800–819. [\[CrossRef\]](#)
11. Cayirli, T.; Yang, K.K.; Quek, S.A. A universal appointment rule in the presence of no-shows and walk-ins. *Prod. Oper. Manag.* **2012**, *21*, 682–697. [\[CrossRef\]](#)
12. Lee, S.J.; Heim, G.R.; Sriskandarajah, C.; Zhu, Y. Outpatient appointment block scheduling under patient heterogeneity and patient no-shows. *Prod. Oper. Manag.* **2018**, *27*, 28–48. [\[CrossRef\]](#)
13. Cayirli, T.; Veral, E. Outpatient scheduling in health care: A review of literature. *Prod. Oper. Manag.* **2003**, *12*, 519–549. [\[CrossRef\]](#)
14. Bailey, N.T. A study of queues and appointment systems in hospital out-patient departments, with special reference to waiting-times. *J. R. Stat. Soc. Ser. B (Methodol.)* **1952**, *14*, 185–199. [\[CrossRef\]](#)
15. Welch, J.; Bailey, N. Appointment systems in hospital outpatient departments. *Lancet* **1952**, *259*, 1105–1108. [\[CrossRef\]](#) [\[PubMed\]](#)
16. Denton, B.; Gupta, D. A sequential bounding approach for optimal appointment scheduling. *IIE Trans.* **2003**, *35*, 1003–1016. [\[CrossRef\]](#)
17. Kaandorp, G.C.; Koole, G. Optimal outpatient appointment scheduling. *Health Care Manag. Sci.* **2007**, *10*, 217–229. [\[CrossRef\]](#)
18. Hassin, R.; Mendel, S. Scheduling arrivals to queues: A single-server model with no-shows. *Manag. Sci.* **2008**, *54*, 565–572. [\[CrossRef\]](#)
19. de Kemp, M.A.; Mandjes, M.; Olver, N. Performance of the smallest-variance-first rule in appointment sequencing. *Oper. Res.* **2021**, *69*, 1909–1935. [\[CrossRef\]](#)
20. Feng, H.; Alvarado, M.M.; Konda, S.; Lawley, M. Sequential clinical scheduling with stochastic patient re-entrance: Case of Mohs micrographic surgery. *Unpublished manuscript*, last modified on 15 November 2022.
21. Mandelbaum, A.; Momčilović, P.; Trichakis, N.; Kadish, S.; Leib, R.; Bunnell, C.A. Data-driven appointment-scheduling under uncertainty: The case of an infusion unit in a cancer center. *Manag. Sci.* **2020**, *66*, 243–270. [\[CrossRef\]](#)
22. Salah, H.; Srinivas, S. Predict, then schedule: Prescriptive analytics approach for machine learning-enabled sequential clinical scheduling. *Comput. Ind. Eng.* **2022**, *169*, 108270. [\[CrossRef\]](#)
23. Samorani, M.; Harris, S.L.; Blount, L.G.; Lu, H.; Santoro, M.A. Overbooked and overlooked: Machine learning and racial bias in medical appointment scheduling. *Manuf. Serv. Oper. Manag.* **2022**, *24*, 2825–2842. [\[CrossRef\]](#)
24. Feyman, Y.; Legler, A.; Griffith, K.N. Appointment wait time data for primary & specialty care in veterans health administration facilities vs. community medical centers. *Data Brief* **2021**, *36*, 107134. [\[PubMed\]](#)
25. Harutyunyan, H.; Khachatryan, H.; Kale, D.C.; Ver Steeg, G.; Galstyan, A. Multitask learning and benchmarking with clinical time series data. *Sci. Data* **2019**, *6*, 96. [\[CrossRef\]](#) [\[PubMed\]](#)
26. Song, H.; Rajan, D.; Thiagarajan, J.; Spanias, A. Attend and diagnose: Clinical time series analysis using attention models. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
27. Ang, E.; Kwasnick, S.; Bayati, M.; Plambeck, E.L.; Aratow, M. Accurate emergency department wait time prediction. *Manuf. Serv. Oper. Manag.* **2016**, *18*, 141–156. [\[CrossRef\]](#)
28. Cayirli, T.; Veral, E.; Rosen, H. Designing appointment scheduling systems for ambulatory care services. *Health Care Manag. Sci.* **2006**, *9*, 47–58. [\[CrossRef\]](#)
29. Cayirli, T.; Veral, E.; Rosen, H. Assessment of patient classification in appointment system design. *Prod. Oper. Manag.* **2008**, *17*, 338–353. [\[CrossRef\]](#)
30. The Stat Council, the People's Republic of China. Public Holidays 2018. Available online: http://english.www.gov.cn/policies/latest_releases/2017/12/01/content_281475960779708.htm (accessed on 13 July 2022).
31. The Stat Council, the People's Republic of China. Public Holidays 2019. Available online: http://english.www.gov.cn/policies/latest_releases/2018/12/06/content_281476422057966.htm (accessed on 13 July 2022).
32. Cui, C. JioNLP. GitHub Repository. 2020. Available online: <https://github.com/dongrixinyu/JioNLP> (accessed on 22 February 2023).
33. Millhiser, W.P.; Veral, E.A. A decision support system for real-time scheduling of multiple patient classes in outpatient services. *Health Care Manag. Sci.* **2019**, *22*, 180–195. [\[CrossRef\]](#)
34. Muthuraman, K.; Lawley, M. A stochastic overbooking model for outpatient clinical scheduling with no-shows. *IIE Trans.* **2008**, *40*, 820–837. [\[CrossRef\]](#)
35. Feng, H.; Jia, Y.; Zhou, S.; Chen, H.; Huang, T. Outpatient appointment scheduling with heterogeneous service time: A data-driven approach. *Unpublished manuscript*, last modified on 15 February 2023.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.