Contents lists available at ScienceDirect

The Lancet Regional Health - Western Pacific

journal homepage: www.elsevier.com/locate/lanwpc

Research paper

Health-adjusted life expectancy (HALE) in Chongqing, China, 2017: An artificial intelligence and big data method estimating the burden of disease at city level

Xiaowen Ruan[a], Yue Li[b], Xiaohui Jin[c], Pan Deng[a], Jiaying Xu[a], Na Li[d], Xian Li[a], Yuqi Liu[d], Yiyi Hu[c], Jingwen Xie[c], Yingnan Wu[d], Dongyan Long[a], Wen He[d], Dongsheng Yuan[c], Yifei Guo[c], Heng Li[a], He Huang[e], Shan Yang[e], Mei Han[f], Bojin Zhuang[a], Jiang Qian[a], Zhenjie Cao[f], Xuying Zhang[b,*], Jing Xiao[a,*], Liang Xu[a,*]

[a] Ping An Technology (Shenzhen) Co., Ltd., Ping'an International Financial Center, Futian District, Shenzhen 518001, China
[b] China Population and Development Research Center, 12 Dahuisi Road, Haidian District, Beijing 100801, China
[c] Ping An Technology (Shenzhen) Co., Ltd., No. 316, Laoshan Road, Pudong New District, Shanghai 200122, China
[d] Ping An Technology (Shenzhen) Co., Ltd., Ping An International Finance Centre, No. 3, South Xinyuan Road, Chaoyang District, Beijing 100011, China
[e] Chongqing Municipal Health Commission, No. 232 Renmin Road, Yuzhong District, Chongqing 400015, China
[f] Ping An Technology (Shenzhen) Co., Ltd., Ping An Tech, US Research Lab, Suite 150, 3000 El Camino Real, Palo Alto, CA 94306, United States

ARTICLE INFO

Article history:
Received 18 August 2020
Revised 25 January 2021
Accepted 3 February 2021
Available online 2 March 2021

ABSTRACT

Background: A universally applicable approach that provides standard HALE measurements for different regions has yet to be developed because of the difficulties of health information collection. In this study, we developed a natural language processing (NLP) based HALE estimation approach by using individual-level electronic medical records (EMRs), which made it possible to calculate HALE timely in different temporal or spatial granularities.

Methods: We performed diagnostic concept extraction and normalisation on 13•99 million EMRs with NLP to estimate the prevalence of 254 diseases in WHO Global Burden of Disease Study (GBD). Then, we calculated HALE in Chongqing, 2017, by using the life table technique and Sullivan's method, and analysed the contribution of diseases to the expected years "lost" due to disability (DLE).

Findings: Our method identified a life expectancy at birth (LE$_0$) of 77•9 years and health-adjusted life expectancy at birth (HALE$_0$) of 71•7 years for the general Chongqing population of 2017. In particular, the male LE$_0$ and HALE$_0$ were 76•3 years and 68•9 years, respectively, while the female LE$_0$ and HALE$_0$ were 80•0 years and 74•4 years, respectively. Cerebrovascular diseases, cancers, and injuries were the top three deterioration factors, which reduced HALE by 2•67, 2•15, and 1•19 years, respectively.

Interpretation: The results demonstrated the feasibility and effectiveness of EMRs-based HALE estimation. Moreover, the method allowed for a potentially transferable framework that facilitated a more convenient comparison of cross-sectional and longitudinal studies on HALE between regions. In summary, this study provided insightful solutions to the global ageing and health problems that the world is facing.

Funding: National Key R and D Program of China (2018YFC2000400).

Research in Context

Evidence before this study

Health-adjusted life expectancy (HALE) is a summary measure for quantifying the population health that accounts for the years of life by different health status. HALE reflects

* Corresponding authors.
E-mail addresses: zhxy88480371@163.com (X. Zhang), XIAOJING661@pingan.com.cn (J. Xiao), XULIANG867@pingan.com.cn (L. Xu).

https://doi.org/10.1016/j.lanwpc.2021.100110

the human life quality more comprehensively and has been used to measure differences among regions and changes over time. In addition, both the Chinese government and World Health Report recently included HALE as a significant indicator of the population's health. During the 10 years after HALE began to gain attention, the most commonly used approach calculated HALE by using the life table technique and Sullivan's method based on the population survey data. In fact, as electronic medical records (EMRs) became widespread in China and the Chinese Central Office pointed out clearly that EMR data could be used to promote the intelligent decision-making application services, our EMR data appeared as an alternative for HALE estimation. Moreover, while most of the previous studies estimated HALE at the national or regional levels, few have accessed HALE at the city level in China, which provides city-specific characteristics.

Added value of this study

We used the population information of Chongqing in 2017 and the electronic medical records (EMRs) to estimate the Life Expectancy (LE), expected years lost due to disability (DLE), and HALE. We supplemented the mortality rate with the statistics from Chongqing Center for Disease Control and Prevention (CDC Chongqing). To deal with the difficulties of health information collection for disease prevalence estimation, we developed a novel method of disease concept extraction and normalisation using techniques of natural language processing (NLP). The method relies on the multi-centered (958 hospitals) EMRs, covering 955300 patients between 2013 and 2017. We used a recurrent neural network (RNN) to identify the disease surface forms from free text diagnostics. Distributed word representation models were trained using 13•99 million medical narratives and combined with a rule-based matching approach, standardising the detected surface forms to the disease names defined by GBD 2004. The LE and HALE obtained using the proposed method were consistent with the estimations released by official authorities and previous studies, indicating its effectiveness to serve as an alternative HALE estimation approach. In terms of DLE, we found four major contributors to health loss, namely cancer, injuries, cerebrovascular disease, and chronic obstructive pulmonary disease. In addition, a gender-wise analysis showed that given a particular disease type, the health loss of males and females varied. Our findings therefore provided a targeted interpretation to the population's health status and could help the local government with the formulation of health policies.

Implications of all the available evidence

HALE has been included as one of the most important indicators by China in the recently published outline of the plan – 'Healthy China 2030' and 'Healthy China Action 2019–2030'. Although China has been dedicated to the improvement of HALE levels, the methodologies of measurement and data acquisition remain underdeveloped. Moreover, HALE has rarely been estimated at the city level because of the complexity of harmonising health status valuations across cities and the lack of resources to conduct large population surveys. As a result, it is less likely to help local authorities in making better decisions. The large volume of unutilised EMRs can potentially serve as an alternative data source, where the extraction and analysis of rich disease-specific information can be automated with artificial intelligence techniques. An AI method based on EMR data may serve as a new instrument to provide more reliable and valid measurements. While the nation-wide surveys are usually performed every five years, our method allows for the seasonal or yearly estimations of HALE, thus providing considerably more flexibility. In addition, the diseases listed as the top disability contributors in this study should be considered to be priorities while formulating health policies, such as elderly health management, chronic disease management, and medical resource allocation in Chongqing. Gender-specific health policies should be formulated on the basis of the differences in the contributors and consequences of health loss.

1. Introduction

Population health evaluation has drawn increasing attention with the increase in life expectancy, as merely an increase in the number of years that one could survive does not necessarily mean an extension of a quality life experience. Healthy life expectancy comprehensively reflects the human life quality by taking the healthy component of life into consideration. World Health Report 2000 explicitly used health-adjusted life expectancy (HALE) as a performance indicator for quantifying the population health status [1]. Moreover, the European Union has been reporting the monitoring value of annual HALE since 2004. Countries including the United Kingdom, France, Sweden, the United States of America, and Japan have incorporated HALE improvement into their policy objectives [2]. In addition, China published 'Healthy China 2030' and 'Healthy China Action 2019–2030' in 2016 and 2019, both including HALE as one of the most important policy indicators of the population health condition.

HALE is calculated by subtracting the years of life lost due to disability from the estimated value of life expectancy. The most commonly used methods are based on specialised health surveys. The indicators of these methods included activity restriction by Sullivan, disability statistics by the Washington team, GALI index proposed by the REVES research team, Index of Independence in Activities of Daily Living by Katz et al., and Index of Independence in Instrumental Activities of Daily Living by Lawton et al [3–7]. In the specialised health survey questionnaires, the designed questions were adapted to the relevant health concepts. Thus, the population health status could be measured from the sampled surveys [8,9].

However, studies relying on the use of questionnaires are time-consuming and labor-intensive [10]. The quality of collected data varies considerably depending on the design of the sampling method, the representativeness of the sample, the organisation and implementation of the survey, and other factors that might affect the data collection process. Considerable resources and personnel are required to obtain robust and comparable health status data during the surveys, so as to overcome the high turnover of people, harmonise self-reported information from different participants, and ensure the progress in a standardised manner [10]. Without a big investment, the health status data cannot be continuously collected and could be out-of-date after the survey years. The lack of resources and personnel could result in poor data quality in some countries or regions [10]. Despite the fact that the method to calculate the years of life lost relies on the statistical data of diseases released by governments and can better link various diseases to HALE, the issues with respect to timeliness, complexity, specificity, and comparability among different regions still exist.

Issues related to the objectiveness, representativeness, and timeliness of data lead to uncertainty in the estimation of the prevalence and the health states at the population level, and have limited the application of HALE indicators [11]. Data sources with higher quality and more accurate years of life lost calculation with reduced time and effort are in urgent required to tackle these issues [12–14].

In this study, we developed an innovative and less cohort survey- dependent approach that calculates the years of life lost and HALE by combining electronic medical records (EMRs) with artificial intelligence and natural language processing (NLP) techniques. EMRs are digitised multi-dimensional sequential data generated from individual medical activities. The continuous health status of the population can be obtained from such a de-facto data source at a fraction of the cost of prospective cohort studies [15]. Data collected by surveys at a single time point make it difficult to reflect the changes in the population health status unless more surveys are done at different time points, which is expensive and time-consuming. The health information recorded in EMRs covers multi-sourced patient data collected from clinical examinations and diagnoses, instead of self-reports, through different points of time. In spite of the heterogeneity in EMR systems between countries and regions, disease-specific metrics are necessarily collected at each clinic visit for each patient. This makes the EMR-based HALE calculation method adaptable to different regions. Moreover, the greater frequency of clinic visits in EMRs ensures the timeliness in the acquisition of the health status data, making periodic computation of HALE practical. Thus, EMR data may provide an alternative data source of population health status for HALE calculation in a time-saving and effort-saving manner.

Despite all the abovementioned advantages, the unstructured data in EMRs have been impairing their application for automated data computation and analysis tasks. The recent development of NLP techniques allows machines to read and process clinical narratives in EMR data. In particular, named entity recognition (NER), a sub-field of NLP, has made medical concept identification and extraction from free text possible, and therefore, it has been prevalently used in biomedical language processing [16–20]. Meanwhile, word embeddings capture useful semantic properties and relationships between words, which can be applied for measuring the semantic relatedness between medical terms in downstream tasks such as medical concept normalisation in EMR [21–22].

In this study, we used data from Chongqing, one of the largest and most populated cities in western China, to show the feasibility of HALE estimation with EMR and the NLP technique. The LE_0 and $HALE_0$ values of the entire population and the gender-wise population in Chongqing in 2017 were calculated, and the contributions of disease and injuries to the expected years 'lost' due to disability (DLE) were analysed with the proposed method. As only the disease relevant metrics of EMRs were considered for the prevalence estimation, and NLP was adopted as a universally applicable analytic workflow to handle disease concepts extraction and normalisation from multi-centered and heterogeneous EMR data in Chongqing, our method therefore has great potential to be reapplied to different regions. The results of this study demonstrated that the proposed method holds promise for tackling health inequality issues and for developing health reform policies, and thus, benefiting the population's health.

2. Methods

2.1. Overview

A consistent and comparative description of the burden of diseases and injuries, and the risk factors that cause them, is an important input to health decision-making and planning processes. Information that is available on mortality and health in populations in some regions of the world is fragmentary and sometimes inconsistent. Thus, a framework for integrating, validating, analysing, and disseminating such information is needed to assess the comparative importance of diseases and injuries in causing premature death, loss of health, and disability in different populations. The first Global Burden of Disease Study (GBD) quantified the health effects of more than 100 diseases and injuries for eight regions of the world in 1990 [23–25]. It generated comprehensive and internally consistent estimates of mortality and morbidity by age, sex, and region [26].

The workflow of the proposed NLP-based HALE estimation method is shown in Fig. 1. Disease-relevant information was extracted by NLP algorithms from unstructured EMR data and normalised to be aligned with the diseases covered by the 2004 GBD disease system. We calculated the prevalence rate of different diseases in different age groups in Chongqing on the basis of the resulting diseases in alignment with GBD 2004. The years of life lost in the health parameter was then estimated from the calculated prevalence rate and life expectancy with Sullivan's life table method.

2.2. Data

The required data for the HALE calculations in this study were obtained from the Chongqing family population information system (FIS) and EMR systems operated by Chongqing Municipal Health Commission. The personal information of patients was de-identified, and the use of data was approved by the Chongqing Municipal Health Commission.

2.2.1. Population data

The population information of Chongqing (covering 40 districts and counties, including Liangjiang New Area and Wansheng Economic Development Zone) was fetched from FIS. The total population was 41•7181 million, including permanent residents, household registered population, and death toll. We estimated the age structure of the living population in 2017 by using the birth, mortality, and migration temporal information from FIS. Further adjustments were made on the basis of the Chongqing official statistical yearbook. We estimated that Chongqing had an average living population of 30•6153 million in 2017 and that the median age of the population was 39 years (interquartile range, IQR: 20–54 years) with a male–female ratio of 105•95 (female = 100).

We supplemented the mortality rate in 2017 with the death certificate statistics provided by CDC Chongqing and applied Chiang's method to calculate the probability of death at all ages [27]. The fraction of the year at death turned out to be 0•25 ($a_0 = 0 \bullet 25$). We used the human mortality database (HMD) with the log-quadratic model and model life table method to adjust the mortality rate of 0-year-old infants and 0-to-4-year-old children, respectively [27–29]. Furthermore, we applied Kannisto's model to adjust the mortality rate of the population aged 80 years and above [30].

2.2.2. Health data

EMR data, instead of the collected data from health surveys or the released statistical data, were introduced to obtain the required health-status data for the HALE calculations in this study. EMRs are digitised multidimensional sequential data that was generated from individual medical activities, including symptoms, diagnoses, and interventions. An ICD-10 disease code was assigned during the information recording procedure upon a patient visit. In addition, abundant disease information was embedded inside the descriptor of the disease name as well as the diagnostic notes, which were written in free text, containing diverse writing styles and formats. Compared with structured data on disease diagnosis in EMR, free text notes were able to provide more information on patient's medical history and chronic diseases.

This study leveraged both structured data fields such as disease names and diagnostic codes (ICD-10), and unstructured diagnostic notes for the disease prevalence rate estimation. The use of both the structured and the unstructured EMR data mentioned above was beneficial for acquiring a more complete picture of the current and the historical health status of a patient [31]. The EMR data covered 13•99 million medical records collected from 957 health

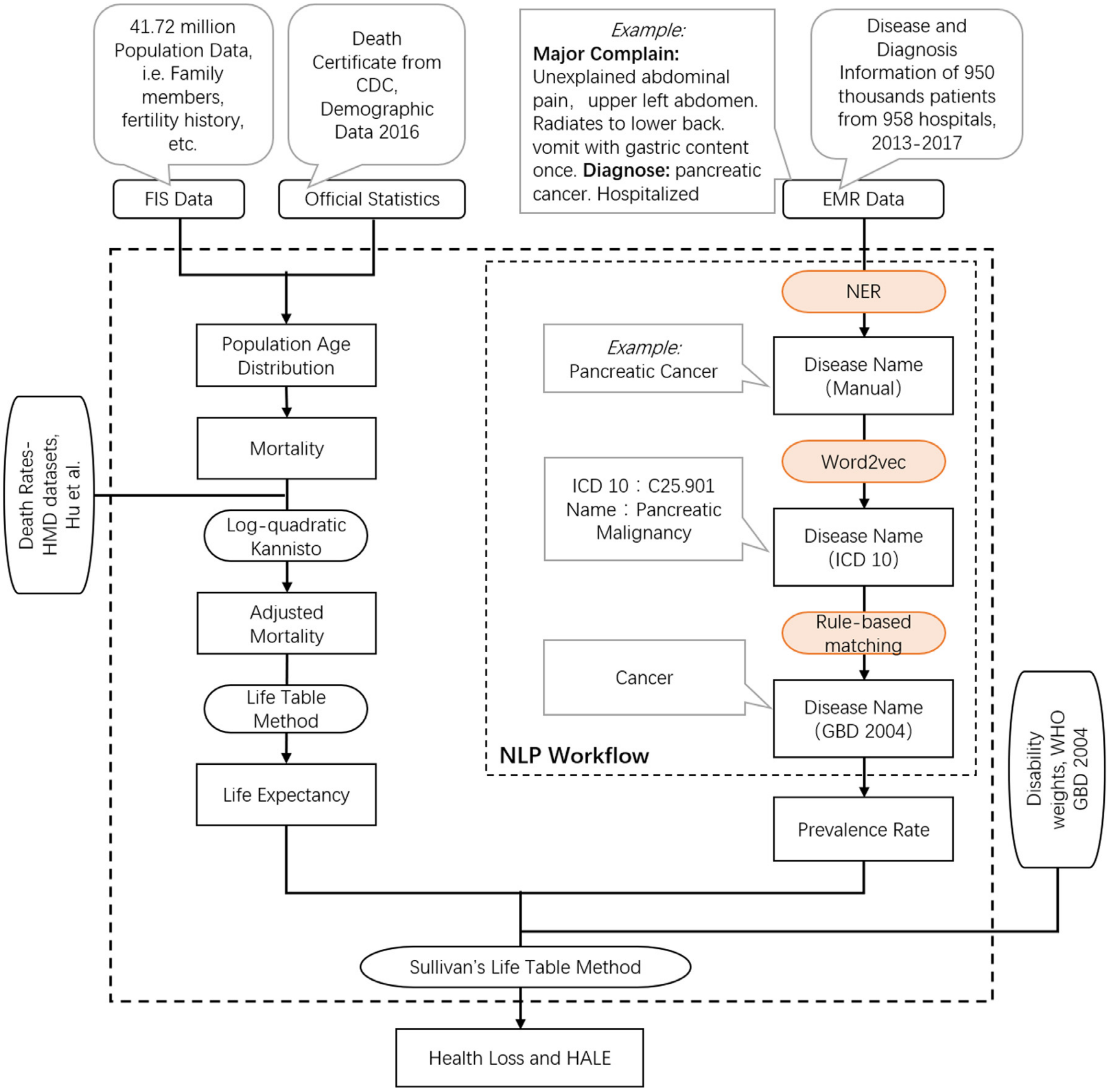
41.72 million Population Data, i.e. Family members, fertility history, etc.
Death Certificate from CDC, Demographic Data 2016
Example:
Major Complain:
Unexplained abdominal pain, upper left abdomen. Radiates to lower back. vomit with gastric content once. Diagnose: pancreatic cancer. Hospitalized
Disease and Diagnosis Information of 950 thousands patients from 958 hospitals, 2013-2017
FIS Data
Official Statistics
EMR Data
NER
Population Age Distribution
Example:
Pancreatic Cancer
Disease Name (Manual)
Mortality
Word2vec
Death Rates-HMD datasets, Hu et al.
Log-quadratic Kannisto
ICD 10：C25.901
Name：Pancreatic Malignancy
Disease Name (ICD 10)
Adjusted Mortality
Rule-based matching
Cancer
Life Table Method
Disease Name (GBD 2004)
NLP Workflow
Disability weights, WHO GBD 2004
Life Expectancy
Prevalence Rate
Sullivan's Life Table Method
Health Loss and HALE

Fig. 1. Workflow of NLP-based HALE estimation. Left: Age-wise death rate calculation with FIS data (micro data), official statistical data (macro data), and HMD database (macro). Right: Age-wise disease prevalence estimation with EMR (micro) and GBD disability weights (macro).

institutions (including 77 hospitals, 142 community health centres, 682 health centres, 25 clinics, 20 maternal and child care service centres, and 11 other health-related agencies) in Chongqing. After data cleaning and quality control including de-duplication and abnormality detection, we obtained the disease information covering 955,300 individuals in Chongqing. This is an immense sampling of the population compared to the data scale in regular surveys.

We calculated the fractions of different types of diseases grouped by 22 ICD-10 annotated categories in our data (Table 1). We compared Ratio, the number of ICD codes of each disease category in the total number of ICD codes in all records of all patient visits, with Ratio for reference from 2016 Chongqing Health and Family Planning Statistical Yearbook, using the statistics in Table 1. The results of Spearman rank correlation test indicated that the general health status from our EMR-derived data was statistically correlated with the health status of the population in Chongqing (ρ=0•850, p=2•13E-6). The subsequent HALE calculations depended on the prevalence rate of the 22 ICD-10 chapters and FIS data, and thus the predicted values are expected to correspond to the HALE of the population in Chongqing.

In particular, the calculation of the health loss caused by various diseases and conditions could be split into two steps. (1) We calculated the prevalence rate of various diseases per year on the basis of the EMRs. (2) We combined multiple indicators such as population prevalence rate, WHO disability weights, and life expectancy measured, to calculate the years of life lost due to various diseases and conditions by using the life tables. For a given age group, the total number of years of life lost in health was the sum of health loss caused by all the diseases or conditions.

The WHO disability weights provided by a GBD study in 2004 covered 254 types of diseases and conditions, including 199 types of diseases and conditions (except cancers and injuries), 18 types of malignant neoplasms and their long-term sequelae, and 37 types of injuries [33]. In our study, disability referred to any short-term or long-term health loss other than death described in the WHO disability weight coefficient.

2.2.3. NLP-based model construction and disease information retrieval

As shown in Fig. 1, we established a three-step NLP workflow to extract the diagnostic entities (free text mentions regarding dis-

Table 1
Statistics of disease prevalence grouped by 22 chapters in ICD-10 derived from EMR data.

ICD-10 Chapter	Coding Range	Occurrence	Ratio	Ratio for reference[32]
Certain infectious and parasitic diseases	A00-B99	130,350	2•86%	3•07%
Neoplasms	C00-D48	490,782	10•77%	4•70%
Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	D50-D89	84,808	1•86%	0•65%
Endocrine, nutritional and metabolic diseases	E00-E90	332,391	7•29%	2•33%
Mental and behavioural disorders	F00-F99	26,920	0•59%	1•40%
Diseases of the nervous system	G00-G99	107,051	2•35%	2•95%
Diseases of the eye and adnexa	H00-H59	65,122	1•43%	2•59%
Diseases of the ear and mastoid process	H60-H95	22,999	0•50%	0•85%
Diseases of the circulatory system	I00-I99	657,067	14•42%	14•08%
Diseases of the respiratory system	J00-J99	643,580	14•12%	19•38%
Diseases of the digestive system	K00-K93	472,767	10•37%	10•52%
Diseases of the skin and subcutaneous tissue	L00-L99	69,728	1•53%	0•93%
Diseases of the musculoskeletal system and connective tissue	M00-M99	256,890	5•64%	6•35%
Diseases of the genitourinary system	N00-N99	422,711	9•28%	7•15%
Pregnancy, childbirth and the puerperium	O00-O99	173,790	3•81%	8•33%
Certain conditions originating in the perinatal period	P00-P96	33,652	0•74%	1•74%
Congenital malformations, deformations and chromosomal abnormalities	Q00-Q99	10,104	0•22%	0•45%
Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified	R00-R99	163,842	3•59%	1•50%
Injury, poisoning and certain other consequences of external causes	S00-T98	201,847	4•43%	6•84%
Codes for special purposes	U00-U85	4	0•00%	4•19%(diseases for other medical services)
External causes of morbidity and mortality	V01-Y98	6077	0•13%	
Factors influencing health status and contact with health services	Z00-Z99	185,020	4•06%	

Occurrence: total number of the EMR entries in which the diagnosis of one disease in the corresponding ICD-10 chapter was made. Ratio represents the number of ICD codes of each disease category in total number of ICD codes in all records of all patient visits. Ratio for reference: the disease prevalence data from the statistical yearbook of Chongqing.

eases) from EMRs and match all the entities to the disease entries defined in GBD 2004: (1) Extraction of a patient's historical disease mentions from notes to supplement the existing disease names found in the disease name field. (2) Normalisation of both extracted mentions and disease names stored in EMRs to standardised names under ICD-10 taxonomies with word embedding techniques and retrieval of their corresponding ICD-10 codes. The retrieved disease codes were then used to supplement the existing ICD-10 codes in the EMR data. (3) Alignment of all the ICD-10 codes to the GBD 2004 disease system with mapping rules curated by domain experts. Finally, the disease prevalence in Chongqing could be derived.

2.2.4. Named entity recognition-based disease information extraction

In our study, the diagnostic concepts in the unstructured EMR notes, such as disease names and symptoms, were referred to as entities. Annotating and extracting the target entities from unstructured text, or NER, formed the basis of the NLP analysis. In this step, we applied the BiLSTM-CRF model to recognise the disease-related entities in EMRs with the incorporation of contextual semantics. BiLSTM refers to the family of bi-directional recurrent neural networks with long short-term memory that encodes the contextual semantic properties in both the forward and the backward directions. A conditional random field (CRF) layer on top of the BiLSTM model was adopted to optimise the decoding process of the entity labels. We annotated 3000 randomly sampled diagnostic notes written in Chinese by using the BIO tagging scheme at the character level. The annotated text contained 1626648 ground-truth labels, including 7584 disease/symptom terms (each diagnostic text included 2•52 labeled terms on average). We trained a character level Bi-LSTM-CRF model, which was adapted for Chinese medical text entity recognition, with cross validation and grid search. We split the annotated data into training and test sets, and performed a five-fold cross validation on the training set, with four-folds for training and the rest one-fold for performance validation, where F1-scores were assessed on the development set in each fold and were averaged as the performance indicator. Grid search was applied to find the optimal model hyper-parameters, given the averaged F1-score derived from cross validation. Finally, we retrained the model with the optimised set of hyper-parameter values and evaluated on the separate test set to derive the final F1-score.

2.2.5. Word2vec method-based ICD-10 synonym identification

One particular disease might be described with different expressions such that polysemy occurs frequently in the extracted disease entities. To normalise these disease expressions, we applied a word2vec model to encode the disease entities detected in the NER step into vector representations, and mapped them to the corresponding ICD10 disease names. We trained an unsupervised word representation model on 0•955 million diagnostic notes, 1•08 million word entries from a publicly available knowledge base and 69K sentences from medicine books. Similar to previous studies that utilised ord2vec for synonym extraction relying on semantic space similarity, we employed word representations for the alignment of surface terms detected by NER to standard ICD terms that were the most semantically relevant. The disease mentions extracted from EMRs and the standard disease names in ICD-10 were first encoded into lower dimensional embeddings. The terms in ICD-10 are usually longer than the extracted entities from free text. Thus, we first applied Chinese word segmentation to the ICD-10 term, and then transformed segmented words into dense vectors, which were weight-averaged to derive the final embedding of the ICD-10 term. To improve the quality of result representations, the principal component (derived by performing PCA on the word embedding matrix) was subtracted from the word embeddings in order to eliminate the dominating directions and the mean values of the vector space, thus making the embeddings stronger [34]. We then measured the cosine similarity between the dense vectors of the extracted entities and the ICD-10 term. To define the best cut-off threshold in similarity measure, we created ground-truth data by aligning 6795 of the 7584 terms detected by NER

(789 unaligned) in the previous step to ICD-10 names. We automated a testing procedure to evaluate our trained model on these ground-truth data and adjusted a threshold (0•68) that yielded optimum F1-score. The disease name in ICD-10 with the highest similarity score above the adjusted threshold of 0•68 was selected as a normalisation outcome. More detailed explanations of the above two sections can be found in the Supplementary Information.

2.2.6. Regular expression-based WHO disease weight system (2004) code matching

Using expert-curated regular expressions rules, we mapped the ICD-10 codes to the 254 diseases in the GBD 2004 disease system. Constraints were set on the patients' gender, age, and disease lasting period for different diseases to discard samples containing mismatched information. For instance, samples of male patients with gestational hypertension were discarded. We constructed a coding table, mapping 11226 ICD-10 disease codes to 254 GBD disease categories, for instance, mapping pancreatic carcinoma (ICD code: C25901) to type Cancer.

2.3. Role of the funding source

The sponsors of this study played no role in the design of the HALE calculation method, collection, or analysis of data; interpretation of results; or in the writing of this manuscript. The corresponding authors had full access to all the data in the study and had final responsibility for the decision to submit for publication.

3. Results

3.1. Life expectancy and health-adjusted life expectancy estimation

Using EMR, we estimated the life expectancy (LE) and the health-adjusted life expectancy (HALE) of the population of Chongqing in 2017 by age and gender. The results showed that the entire population's life expectancy at birth was 77•9 years and the HALE at birth was 71•7 years. To be more specific, the LE_0 and $HALE_0$ of the male population in Chongqing were 76•3 years and 68•9 years, respectively, while the E_0 and $HALE_0$ of the female population were 80•0 years and 74•4 years, respectively (Table 2). The calculated values with more detailed age groups are included in Appendix Table A1.

Fig. 2 shows the estimates of HALE, DLE, LE, and HALE/LE (%), at 10-year age intervals, of the male and female populations of Chongqing in 2017. DLE represents the healthy years lost due to disability. In the year 2017, the discrepancy between LE_0 and $HALE_0$ was 7•4 years for males and 5•6 years for females, which implied that $HALE_0$ decreased by 9•6% and 7%, respectively, compared with the total life expectancy considering the quality of life. As the population aged, the proportion of the years of life in the full health state decreased in the total years of life, while the proportion of the years of life with disability increased in the total years of life. The discrepancy between LE and HALE in a 90-year-old decreased to 0•9 for males and 0•5 for females, and the HALE values only accounted for 83•5% and 86•5% of the total life expectancy at the age of 90, respectively. In conclusion, females had both a higher average LE and HALE than males in Chongqing. Additionally, men spent more years in poor health than women and had a higher ratio of life years with disability than women in this city.

Table 2
Age-specific LE and HALE estimation for Chongqing population in 2017 with 5-year age interval.

Age	All		Male		Female	
	LE	HALE	LE	HALE	LE	HALE
0	77•9	71•7	76•3	68•9	80•0	74•4
5	74•7	68•5	73•2	65•8	76•7	71•1
10	69•8	63•7	68•2	61•1	71•7	66•3
15	64•9	59•0	63•3	56•4	66•8	61•5
20	60•0	54•3	58•5	51•7	61•8	56•8
25	55•0	49•6	53•6	47•2	56•9	51•9
30	50•2	44•8	48•7	42•6	52•0	47•1
35	45•3	40•2	43•9	38•2	47•1	42•4
40	40•5	35•7	39•2	33•8	42•2	37•7
45	35•9	31•3	34•6	29•6	37•5	33•3
50	31•3	27•2	30•2	25•6	32•8	29•0
55	27•0	23•3	26•1	21•9	28•2	24•8
60	22•8	19•6	22•1	18•6	23•7	20•8
65	18•7	16•1	18•3	15•4	19•3	16•9
70	15•0	12•9	14•9	12•5	15•2	13•3
75	11•6	9•9	11•8	9•8	11•4	10•0
80	8•6	7•3	9•1	7•6	8•2	7•1
85	6•3	5•4	7•1	5•9	5•7	4•9
90	4•5	3•8	5•3	4•4	3•9	3•3
95	2•9	2•5	3•4	2•9	2•5	2•1
100	0•5	0•4	0•5	0•5	0•5	0•4

The values at more detailed age groups are included in Appendix Table A1.

3.2. Contributors to DLE

Our NLP-based HALE estimation method relied on the Chongqing EMR database and WHO's disability database, which helped in the analysis of disease species (groups) that accounted for the healthy years lost due to disability. Table 3 shows that the top four contributors to DLE in Chongqing were cancers, injuries, cerebrovascular disease, and chronic obstructive pulmonary disease, which caused an overall health loss of 2•32, 1•40, 0•52, and 0•37 years, aggregated across age and sex. These four contributors accounted for 37•4%, 22•5%, 8•4%, and 6•0% of DLE. Cerebrovascular diseases included cerebral haemorrhage, encephalic angioma, stroke, and the corresponding sequelae. Injuries included fractures, amputations, open injuries, sprains, and burns. The major contributors found in this research were consistent with the leading causes of life loss reported by Chinese and other global institutions [11,33]. Note that chronic obstructive pulmonary disease (COPD) induced severe health loss for the population aged 40 years and above, indicating that COPD significantly impacted the quality of life for the middle-aged and senior populations with a relatively high prevalence rate.

The fifth to tenth top contributors to DLE were lower back pain, upper respiratory infections, nephrosis, drug use disorders, liver cirrhosis, and hypertensive heart diseases. Drug use disorders included non-alcoholic addictions, psychoactive substance addictions, and acute intoxication. The specific basin climate and dense population distribution of Chongqing might account for the high rank of upper respiratory infections, a type of acute infectious diseases. The detailed ranking of the health loss contributors of Chongqing population in 2017 can be found in Appendix Table A2.

A gender-wise analysis (Fig. 3) showed that many diseases, such as cancers, injuries, COPD, liver cirrhosis, tuberculosis, and hepatitis B, imposed a far greater negative impact on men's health than on women's health, while diabetes led to more health loss in females than in males.

3.3. Survivorship function

Fig. 4 shows the DLE caused by different diseases in the different age groups that we analysed using the survivorship function. Consistent with the results above, the top four health loss contributors were cancers, injuries, cerebrovascular disease, and chronic obstructive pulmonary disease.

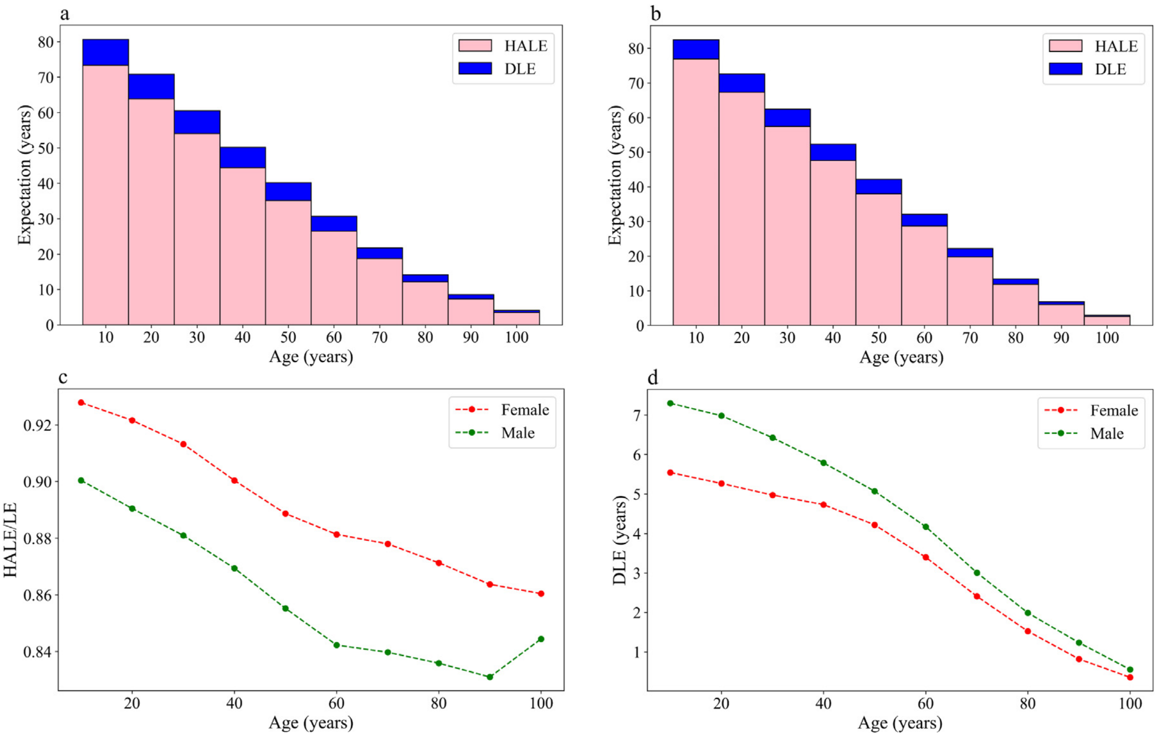

Fig. 2. HALE and DLE comparison between male and female populations of Chongqing in 2017. (a)**:** Female HALE and DLE by age groups. (b)**:** Male HALE and DLE by age groups. Overall heights are the expected ages of death. (c)**:** Changing curve for male and female HALE/LE (%) by age groups. (d)**:** Changing curve for male and female DLE by age groups.

Table 3
Ranking of health loss contributors of Chongqing population in 2017(Top 10).

Disease	All		Male	Female
	DLE	percentage	DLE	DLE
Cancers	2•3197	37•35%	2•5523	2•1436
Injuries	1•4003	22•55%	1•9530	1•0941
Cerebrovascular_disease	0•5248	8•45%	0•5522	0•5222
Chronic_obstructive_pulmonary_disease	0•3745	6•03%	0•4700	0•2793
Low_back_pain	0•1932	3•11%	0•2012	0•2051
Upper_respiratory_infections	0•1725	2•78%	0•1913	0•1721
Nephritis_and_nephrosis	0•1242	2•00%	0•1545	0•1084
Drug_use_disorders_Cases	0•1015	1•63%	0•1179	0•0974
Cirrhosis_of_the_liver	0•0923	1•49%	0•1356	0•0581
Hypertensive_heart_disease_Cases	0•0814	1•31%	0•0785	0•0893

4. Discussion

While the health life expectancy has become an important indicator to measure population health, the collection of health information remains a major challenge. This study developed an artificial intelligence and big data method to estimate the city-level HALE and to analyse the major causes of DLE on the basis of the EMR data.

The results suggested that the LE_0 and $HALE_0$ of the male population in Chongqing were 76•3 years and 68•9 years, respectively, while the LE_0 and $HALE_0$ of the female population were 80•0 years and 74•4 years, respectively. Note that because of the differences in the method, calibre, and granularity, it was not easy to make a direct comparison between the estimated values in this research and the other publicly available estimation values. As a result, we found a way to represent the estimated results by referring to China's total population. WHO estimated that the LE_0 and $HALE_0$ of China in 2015 were 76•1 years and 68•5 years, respectively [31]. Zhou et al. estimated that the LE_0 and $HALE_0$ of China's male population in 2015 was 73•2 years and 65•8 years, respectively, and of the female population was 80•0 years and 70•8 years, respectively [35]. Chen estimated the $HALE_0$ of China's male people and female people was 65•90 years and 70•32 years, respectively, in 2013 [36]. As can be seen, the LE_0 and $HALE_0$ values obtained from the EMRs in this study were reasonable in terms of the reference values . The life years lived with disability of the male population (76•3–68•9 = 7•4 years) in our study was the same as that in the study by Zhou (73•2–65•8 = 7•4 years) [35]. The HALE values of 2017 calculated in our study had a reasonable growth compared with the values in 2013 in Chen's study [36]. The conclusion that females had both an average higher LE and HALE values than those of males in our study also agreed with that of the previous studies. The results demonstrated the feasibility of the use of EMR data for HALE calculations and verified the effectiveness of the proposed NLP-based HALE estimation method.

As shown in Fig. 4, the ten leading contributors for DLE in Chongqing identified in this study were cancers, injuries, cerebrovascular disease, chronic obstructive pulmonary disease (COPD), lower back pain, upper respiratory infections, nephrosis, drug use

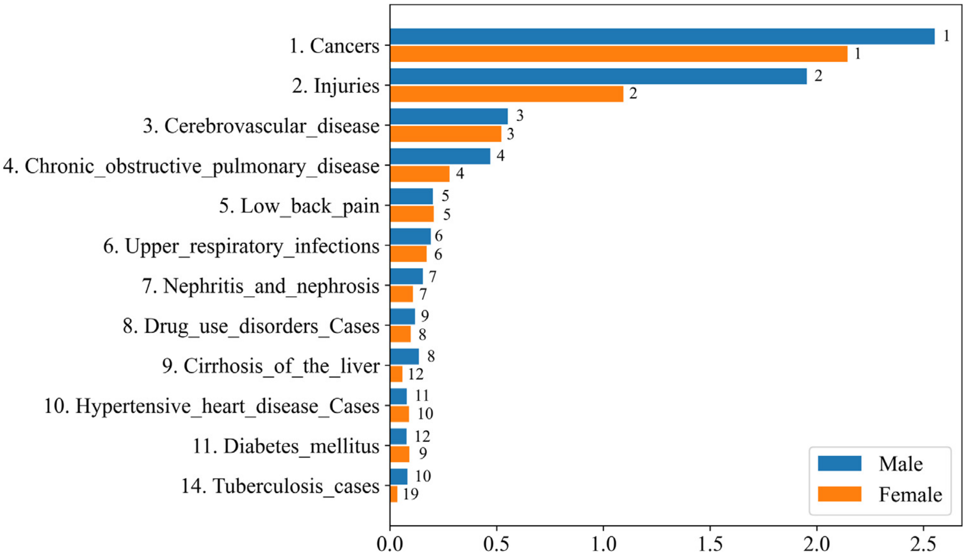

Fig. 3. TOP 10 health loss contributors of male and female populations of Chongqing, 2017, respectively. Diseases are ranked by total population health loss. Left-most numbers are the corresponding ranks in the total population. Numbers on the right of the bars are the gender-wise health loss ranks.

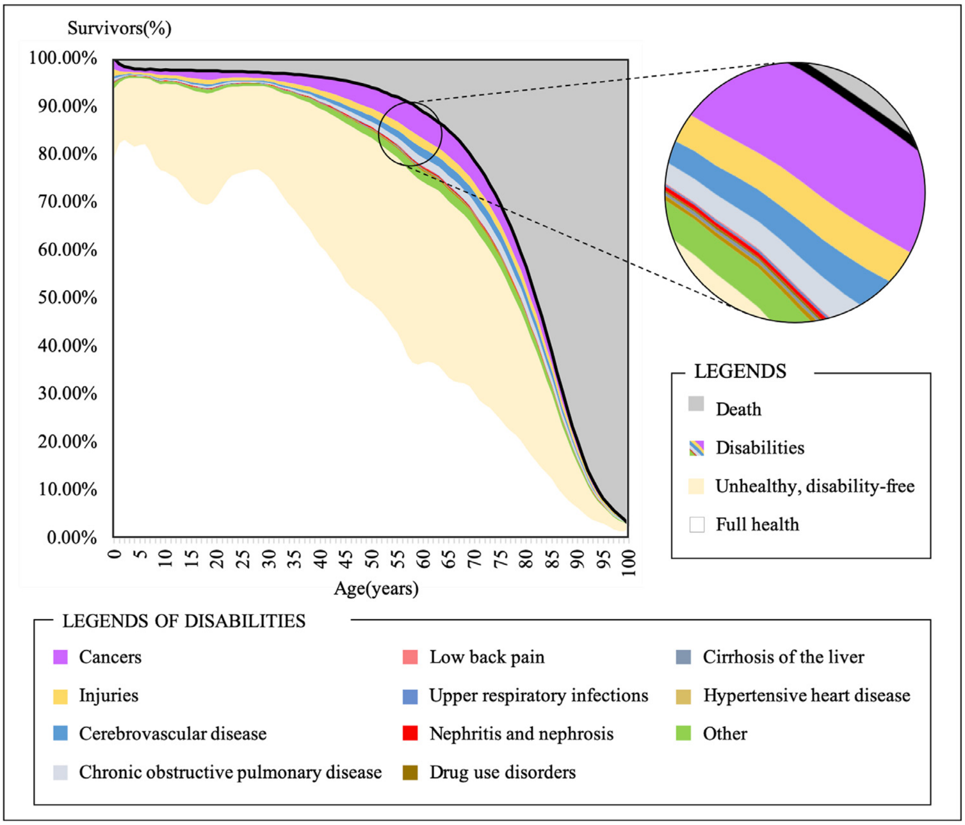

Fig. 4. Health states at different age groups analyzed by the survivorship function. The upper black curve in the figure is the survival curve, which means the survival probability of the single-year-old group, which is calculated from the life table. The other colored areas are the health loss rate of each disability in the legends on the right, and the calculation formula is $g(s)$, where $g(s) = s \bullet \textit{incidence rate} \bullet \textit{disability weight}$ where s is the survival rate from the survival curve for each age.

disorders, liver cirrhosis, and hypertensive heart diseases. The first four contributors accounted for as high as 74•4% of DLE. Among them cancers have been one of the most serious challenges to health and need to be paid more attention to [37–39]. Our study also identified injury as the second leading cause of DLE, which has also been found by previous studies [40,41]. Among the top ten contributors found in our study, cerebrovascular disease, and chronic obstructive pulmonary disease were both common non-communicable diseases. Chongqing Municipality is in southwest China with a high poverty incidence and a significant urban–rural income gap. Many researchers found that non-communicable diseases had a negative total effect in many low- and middle-income regions, corresponding to the target regions of Chongqing Municipality in this study [42,43].

This study found that many diseases, such as cancers, injuries, COPD, liver cirrhosis, tuberculosis, and hepatitis B, took a far greater toll on men's health than on women's health. This observation might account for the shorter HALE of men. Note that the health loss caused by liver cirrhosis in males was considerably higher than that in females, which was also found in Zhou's study. This might be attributed to the differences in alcohol addiction and hepatitis B disease prevalence [36]. Diabetes led to more health loss in females than in males, which might be attributed to hormone disorders during pregnancy and menopause. Meanwhile, previous studies show that diabetic women had higher risks of dying from vascular diseases than men [44]. These results suggested that gender-specific health policies need to be formulated depending on the differences in the factors and consequences of health loss, so as to reduce health loss and improve health-adjusted life expectancy and life expectancy more efficiently.

In this study, we first developed an AI-based method to analyse the disease- relevant information in the EMR data for HALE estimations in Chongqing. Using the massive repository of data points in EMRs enabled a larger sampling ratio of the real-world population than regular surveys. The EMR data used in our study included 955,300 individuals in Chongqing, covering 3% of the whole population, while the sample sizes of most of the other survey studies for HALE estimations account for less than 0.05% of the total population under study [45–47]. Leveraging the health-status statistics derived from the EMR data might be more representative of the real world than cohort studies and is commonly seen in other health-related applications, such as risk prediction [15]. In addition, the patients data observed with greater frequency in EMRs ensures the timeliness of the derived health information.

Moreover, the proposed method for the disease-relevant concept standardisation and HALE calculation with EMR data is transferable. Although the structure and format of EMRs across different countries and regions varies, metrics such as disease names, symptoms, diagnostic codes, and free text notes for prevalence estimation are necessarily collected at each clinic visit, which are processed by NLP programs. The method proposed in this study could be generalised to the HALE estimation of other cities without the need to change the entire workflow and framework. Trained NER and word2vec models under the current settings of institutions can be transferred to other institutions (in a different city) and further fine-tuned to achieve optimum performance. When applied to EMRs in different languages, the workflow is still applicable, however, with the NER model and the word2vec model re-trained with corpora in the target language. These make the HALE calculations based on EMR data at the different levels, such as city, province, and national levels, and the comparison of HALE values among different regions possible.

Despite all of the aforementioned improvements, EMR data come with many analytic challenges. Medical treatment seeking activities recorded in EMRs are often related to demographic and socio-economic factors, thus introducing potential analytical biases. Moreover, EMR data may not cover records in every institution and thus could be prone to missing patient information. We verified the estimated disease prevalence distribution against the broad heading prevalence data released by the authorities. The results indicated that the estimated prevalence distribution was substantially close to the actual prevalence distribution (Table 1), and hence, the possible deviation poses little impact on our research findings.

Although medical records have become increasingly available in the form of EMRs, there is a vast quantity of unstructured clinically relevant data, which remain underutilised, as text features cannot straightforwardly be utilised by conventional data analytic process [48]. In this study, the structured disease names in EMR were processed and from which, 4557502 disease names were extracted, covering 18967 types of diseases in total. 1094668 complementary disease names were captured in the clinical narrative data, covering 8842 types of diseases, among which, 1002 diseases were not seen in the original structured disease name field. This method is therefore beneficial to the acquisition of more complete real-world patient disease information. On this basis, it could contribute in the future to the full utilisation of EMR data when developing clinical decision support systems, so as to improve health care quality in general. To the best of our knowledge, this study is among the first studies that designed a transferable HALE estimation framework with NLP and the EMR data in China. The conclusions from our study displayed good consistency with previous studies, which demonstrated the feasibility of a health status evaluation with EMR data and the effectiveness of the proposed method for relevant medical concept standardisation and HALE calculation [35].

5. Research and policy implications

HALE has only been calculated at the national or city's regional level in almost all the studies on HALE. City-level HALE estimation poses challenges in harmonising outcomes collected from multiple regions, particularly for large cities like Chongqing. In contrast, nationwide and province-wide surveys conducted on the basis of years have a considerable demand for time and resources and the Chinese government does not provide city-level population health results or analysis. Therefore, the estimates can help the local government or officials to make better decisions for improving the citizens' health status, particularly in China where the population and health-status of different cities vary considerably [49]. The proposed AI-based method for HALE estimations with EMR data in this study provides a potentially useful and flexible tool for HALE calculation and health-status analysis at the city level.

The use of EMR data in our study enabled analysis of the contributions of different diseases to the health loss of the population. It showed that the top four contributors to DLE were cancers, injuries, cerebrovascular disease, and chronic obstructive pulmonary disease in Chongqing. In addition, it is worth noting that upper respiratory infections were identified as one of the top 10 contributors to the health loss in Chongqing. Moreover, this study quantified how individual diseases and injuries influenced gender-wise population health. Liver cirrhosis and diabetes were identified as the gender-specific causes of the health loss for males and females, respectively, in Chongqing. These results offer a targeted interpretation of the population health status and help with the formulation of health policies for the local government. Firstly, the decrease in the prevalence and incidence of these causes is expected to reduce the number of years lived with disability if other conditions remain unchanged. The diseases listed above should be considered as priorities while formulating health policies in Chongqing. Secondly, lessons should be learnt from the countries that suffered from both the positive mortality and the disability effects of cancer, chronic obstructive pulmonary disease, and cerebrovascular disease. Furthermore, on the basis of the differences in the factors and consequences of health loss between the males and females, the government could formulate the corresponding health policies for each gender. WHO and other stakeholders have recommended 'best buys' or other measures against these non-communicable diseases such as diabetes, which could serve as a basis for formulating policies [46].

6. Limitations of our study

Using the GBD system, we inherited its limitations in our research. Firstly, the universal weight system GBD can introduce deviations from the local situation where the indicators are measured. Taking the fatty liver in Chongqing as an instance, we found that fatty liver, one of the listed diseases in ICD-10, has not been included in the measurement system of GBD. In cases wherein the fatty liver disease is associated with the estimated value of HALE in Chongqing, it would be a missing variable. Due to the complexity and difficulty of estimating the disability weights as mentioned in the WHO report and the reported strong evidence of highly consistent results across the samples from different cultural environment, in this paper the adjustment for disease categorization and weight coefficient was not conducted on the basis of the actual prevalence situation in Chongqing [50,51]. Secondly, the possibility of inter-correlation between diseases has not been eliminated, such as the correlation between myocarditis and ischemic heart disease. With reference to the GBD weight of disease system by WHO and Sullivan's method, a possible bias resulting from multicollinearity was not further adjusted.

There also exist concerns regarding the quality and representability of EMRs. In the future studies, the patient information could be adjusted by combining the appropriate health survey data with the multi-dimensional data extracted from the EMR database to reduce the influence of the population distribution bias. In terms of the hypothesis test we used when comparing the observed and expected ratios, we should apply more detailed ICD10 categories for verification if more information can be accessed from the government-released yearbook in the future.

As the disease information was extracted using the NLP method from EMRs, there might be random errors from the algorithms. However, we tried to minimise these errors through multiple experiments and parameter adjustments (NER F1 = 0•884, synonym recognition F1 = 0•871). In the subsequent applications, the accuracy of disease extraction could be improved by combining the NLP method with manual reviews by medical experts.

Lastly, we did not implement negation detection in patient disease recognition, such that examples such as 'denied hypertension' should not be considered in the ICD mapping. We omitted the annotation of disease mentions that appeared within a negation context in the diagnostic text, in order to reduce the falsely predicted entities by the model. In the future studies, we may attempt to include a negation detection module in our NLP workflow to improve accuracy. Other than NER-based disease concepts recognition method, a promising alternative that relies on document and patient level phenotyping might be considered in the future studies. Both codified data and concepts extracted by NLP could be leveraged as features to build prediction models, thus mitigating potential errors introduced by local features such as negation at the sentence level [52].

Contributors

Ms. Xiaowen Ruan, Dr. Yue Li, and Ms. Xiaohui Jin are co-first authors.

Dr. Xuying Zhang, Dr. Jing Xiao, and Dr. Liang Xu are all corresponding authors.

Ms. Xiaowen Ruan is an AI expert who guided the design of AI algorithms and calculation schemes from the perspective of population and health big data applications, making great contributions to revising drafts.

Dr. Yue Li mainly made academic contributions in demography, conducted literature research and review, designed the calculation method, and was involved in the writing of the manuscript.

Ms. Xiaohui Jin not only contributed to the design of AI algorithms, the setting of the calculation process, and the implementation of data processing but also was responsible for manuscript writing and figure drawing.

Dr. Pan Deng contributed to data processing, AI algorithms design, figure drawing, and overall manuscript writing. Dr. Jiaying Xu had academic contributions in public health, and was involved in designing of the calculation method and the writing of the manuscript, by providing guiding suggestions from the perspectives of the project's policy contributions and public health management. Dr. Na Li implemented big data processing and statistical analysis, and participated in the project design, manuscript writing, and figure making. Dr. Xian Li contributed to the data processing implementation and AI algorithms design, and was responsible for essay writing and diagram modifications. Ms. Yuqi Liu researched the literature for the estimation of healthy life expectancy in the field of demography and designed the calculation plan accordingly. Mr. Yiyi Hu and Ms. Jingwen Xie performed data processing, conducted literature research, and made great contributions to the essay writing. Ms. Yingnan Wu made contributions to essay modifications and reference checking. Dr. Dongyan Long and Mr. Wen He communicated with Chongqing Municipal Health Commission, participated in the manuscript writing, and gave guidance in terms of the policy significance. Mr. Dongsheng Yuan performed data processing, made figures in the paper and participated in the essay writing. Ms. Yifei Guo performed data processing, conducted literature research, and participated in the essay writing. Mr. Heng Li made contributions to data analysis, figure making and manuscript revision. Ms. He Huang and Ms. Shan Yang provided guidance as health big data experts and participated in the manuscript writing. Dr. Mei Han, Dr. Bojin Zhuang, Dr. Jiang Qian and Mr. Zhenjie Cao provided guidance as AI algorithms experts and participated in the essay writing.

Dr. Xuying Zhang is a demographic expert who designed the calculation method from the perspective of China's population development and health management, participating in the core steps of designing the measurement plan and revising the manuscript. Dr. Jing Xiao is an AI expert who guided the design of AI algorithms and calculation schemes from the perspective of population and health big data applications, making great contributions to revising drafts. Dr. Liang Xu primarily made contributions to processing data, designing the AI algorithms and the calculation method, and participated in the essay writing. In addition, Dr. Liang Xu established a collaboration with Chongqing Health and Health Commission, and China Population and Development Research Center.

Data sharing statement

The data required for the HALE calculations in this study were obtained from the Chongqing family population information system and EMR systems operated by Chongqing Municipal Health Commission. The study is carried out under privacy protection. The data owner confirmed that the data cannot be made to the public due to privacy reasons.

Declaration of Competing Interest

We know that a conflict can be actual or potential, and full disclosure to the Editor of all relationships is a requisite. All authors confirm that none conflicts of interest exist and there is nothing to disclose.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.lanwpc.2021.100110.

Appendix

Tables A1 and A2.

Table A1
LE and HALE Estimation Result for Chongqing Population in Year 2017.

Age	All		Male		Female	
	LE	HALE	LE	HALE	LE	HALE
0	77·9	71·7	76·3	68·9	80·0	74·4
1	78·2	72·0	76·6	69·3	80·2	74·6
2	77·4	71·2	75·8	68·5	79·4	73·8
3	76·6	70·3	75·0	67·6	78·5	72·9
4	75·7	69·4	74·1	66·8	77·6	72·0
5	74·7	68·5	73·2	65·8	76·7	71·1
6	73·7	67·6	72·2	64·9	75·7	70·2
7	72·7	66·6	71·2	63·9	74·7	69·2
8	71·8	65·6	70·2	63·0	73·7	68·2
9	70·8	64·7	69·2	62·0	72·7	67·3
10	69·8	63·7	68·2	61·1	71·7	66·3
11	68·8	62·8	67·3	60·1	70·7	65·4
12	67·8	61·8	66·3	59·2	69·7	64·4
13	66·8	60·9	65·3	58·2	68·8	63·4
14	65·9	59·9	64·3	57·3	67·8	62·5
15	64·9	59·0	63·3	56·4	66·8	61·5
16	63·9	58·1	62·4	55·4	65·8	60·6
17	62·9	57·1	61·4	54·5	64·8	59·6
18	61·9	56·2	60·4	53·6	63·8	58·7
19	60·9	55·2	59·4	52·7	62·8	57·7
20	60·0	54·3	58·5	51·7	61·8	56·8
21	59·0	53·4	57·5	50·8	60·9	55·8
22	58·0	52·4	56·5	49·9	59·9	54·8
23	57·0	51·5	55·5	49·0	58·9	53·9
24	56·0	50·5	54·5	48·1	57·9	52·9
25	55·0	49·6	53·6	47·2	56·9	51·9
26	54·1	48·6	52·6	46·3	55·9	51·0
27	53·1	47·7	51·6	45·3	54·9	50·0
28	52·1	46·7	50·6	44·4	54·0	49·0
29	51·1	45·8	49·7	43·5	53·0	48·1
30	50·2	44·8	48·7	42·6	52·0	47·1
31	49·2	43·9	47·8	41·7	51·0	46·2
32	48·2	42·9	46·8	40·8	50·0	45·2
33	47·2	42·0	45·8	39·9	49·0	44·3
34	46·3	41·1	44·9	39·0	48·1	43·3
35	45·3	40·2	43·9	38·2	47·1	42·4
36	44·4	39·3	43·0	37·3	46·1	41·4
37	43·4	38·3	42·0	36·4	45·1	40·5
38	42·4	37·4	41·1	35·5	44·2	39·6
39	41·5	36·5	40·1	34·6	43·2	38·6
40	40·5	35·7	39·2	33·8	42·2	37·7
41	39·6	34·8	38·3	32·9	41·3	36·8
42	38·7	33·9	37·4	32·1	40·3	35·9
43	37·7	33·0	36·4	31·2	39·4	35·0
44	36·8	32·2	35·5	30·4	38·4	34·1
45	35·9	31·3	34·6	29·6	37·5	33·3
46	35·0	30·5	33·7	28·8	36·5	32·4
47	34·1	29·6	32·8	28·0	35·6	31·5
48	33·1	28·8	32·0	27·2	34·6	30·7
49	32·3	28·0	31·1	26·4	33·7	29·8
50	31·3	27·2	30·2	25·6	32·8	29·0
51	30·5	26·4	29·4	24·8	31·9	28·1
52	29·6	25·6	28·5	24·1	30·9	27·3
53	28·7	24·8	27·7	23·4	30·0	26·5
54	27·9	24·1	26·9	22·7	29·1	25·7
55	27·0	23·3	26·1	21·9	28·2	24·8
56	26·1	22·5	25·2	21·2	27·3	24·0
57	25·2	21·7	24·4	20·5	26·4	23·2
58	24·4	21·0	23·6	19·8	25·5	22·4
59	23·6	20·3	22·9	19·2	24·6	21·6
60	22·8	19·6	22·1	18·6	23·7	20·8
61	22·0	18·9	21·4	17·9	22·8	20·0
62	21·2	18·2	20·6	17·3	21·9	19·3
63	20·3	17·5	19·8	16·7	21·0	18·5
64	19·5	16·8	19·1	16·0	20·1	17·7
65	18·7	16·1	18·3	15·4	19·3	16·9
66	18·0	15·4	17·6	14·8	18·4	16·2
67	17·2	14·8	16·9	14·2	17·6	15·5
68	16·5	14·1	16·2	13·6	16·8	14·8
69	15·7	13·5	15·5	13·0	16·0	14·0
70	15·0	12·9	14·9	12·5	15·2	13·3
71	14·3	12·2	14·2	11·9	14·4	12·6

(*continued on next page*)

Table A1 (*continued*)

Age	All		Male		Female	
	LE	HALE	LE	HALE	LE	HALE
72	13•6	11•6	13•6	11•4	13•6	11•9
73	12•9	11•0	13•0	10•9	12•9	11•3
74	12•2	10•5	12•4	10•3	12•2	10•6
75	11•6	9•9	11•8	9•8	11•4	10•0
76	10•9	9•3	11•2	9•4	10•8	9•4
77	10•3	8•8	10•7	8•9	10•1	8•8
78	9•7	8•3	10•1	8•4	9•4	8•2
79	9•2	7•8	9•6	8•0	8•8	7•6
80	8•6	7•3	9•1	7•6	8•2	7•1
81	8•1	6•9	8•7	7•2	7•6	6•6
82	7•6	6•5	8•3	6•9	7•1	6•1
83	7•2	6•1	7•8	6•5	6•6	5•7
84	6•7	5•7	7•5	6•2	6•1	5•3
85	6•3	5•4	7•1	5•9	5•7	4•9
86	5•9	5•0	6•7	5•6	5•3	4•6
87	5•6	4•7	6•3	5•3	4•9	4•2
88	5•2	4•4	6•0	5•0	4•5	3•9
89	4•8	4•1	5•6	4•7	4•2	3•6
90	4•5	3•8	5•3	4•4	3•9	3•3
91	4•2	3•6	4•9	4•1	3•6	3•1
92	3•9	3•3	4•6	3•8	3•3	2•8
93	3•6	3•0	4•2	3•5	3•0	2•6
94	3•3	2•8	3•8	3•2	2•8	2•4
95	2•9	2•5	3•4	2•9	2•5	2•1
96	2•6	2•2	3•0	2•5	2•2	1•9
97	2•2	1•9	2•5	2•1	1•9	1•7
98	1•8	1•5	2•0	1•6	1•6	1•3
99	1•2	1•1	1•3	1•1	1•1	1•0
100	0•5	0•4	0•5	0•5	0•5	0•4

Table A2
Measurement of Health Loss Contributors in year 2017 Chongqing Population.

Disease	All		Male	Female
	DLE	percentage	DLE	DLE
Cancers	2•3197	37•35%	2•5523	2•1436
Injuries	1•4003	22•55%	1•9530	1•0941
Cerebrovascular_disease	0•5248	8•45%	0•5522	0•5222
Chronic_obstructive_pulmonary_disease	0•3745	6•03%	0•4700	0•2793
Low_back_pain	0•1932	3•11%	0•2012	0•2051
Upper_respiratory_infections	0•1725	2•78%	0•1913	0•1721
Nephritis_and_nephrosis	0•1242	2•00%	0•1545	0•1084
Drug_use_disorders_Cases	0•1015	1•63%	0•1179	0•0974
Cirrhosis_of_the_liver	0•0923	1•49%	0•1356	0•0581
Hypertensive_heart_disease_Cases	0•0814	1•31%	0•0785	0•0893
Diabetes_mellitus	0•0813	1•31%	0•0780	0•0909
Cardiomyopathy	0•0615	0•99%	0•0694	0•0566
Schizophrenia_Cases	0•0566	0•91%	0•0771	0•0488
Tuberculosis_Cases	0•0532	0•86%	0•0821	0•0350
Cataracts_Low_vision	0•0525	0•84%	0•0441	0•0639
Appendicitis_Episodes	0•0480	0•77%	0•0575	0•0447
Ischemic_heart_disease	0•0438	0•71%	0•0467	0•0434
Rheumatic_heart_disease_Cases	0•0412	0•66%	0•0279	0•0572
Iron_deficiency_anemia	0•0337	0•54%	0•0302	0•0371
Parkinson_disease_Cases	0•0292	0•47%	0•0297	0•0293
Alzheimer_and_other_dementias	0•0254	0•41%	0•0260	0•0265
Skin_diseases_Cases	0•0233	0•38%	0•0303	0•0191
Meningococcaemia	0•0217	0•35%	0•0262	0•0203
Asthma_cases	0•0185	0•30%	0•0183	0•0196
Epilepsy_Cases	0•0176	0•28%	0•0223	0•0152
Insomnia_primary_Cases	0•0169	0•27%	0•0154	0•0191
Hepatitis_B_Episodes	0•0152	0•24%	0•0237	0•0100
Lower_respiratory_infections	0•0124	0•20%	0•0156	0•0096
Glaucoma_Low_vision	0•0121	0•20%	0•0123	0•0135
Panic_disorder_Cases	0•0117	0•19%	0•0090	0•0150
Birth_asphyxia	0•0114	0•18%	0•0063	0•0147
Rheumatoid_arthritis_Cases	0•0113	0•18%	0•0076	0•0154
Unipolar_depressive_disorders	0•0105	0•17%	0•0089	0•0126
Multiple_sclerosis_Cases	0•0105	0•17%	0•0118	0•0099

(*continued on next page*)

Table A2 (*continued*)

Disease	All		Male	Female
	DLE	percentage	DLE	DLE
Diphtheria	0•0097	0•16%	0•0107	0•0096
Osteoarthritis	0•0094	0•15%	0•0085	0•0114
Refractive_errors_Low_vision	0•0086	0•14%	0•0069	0•0101
Congenital_heart_anomalies_Cases	0•0071	0•11%	0•0067	0•0088
Gout_Cases	0•0070	0•11%	0•0131	0•0028
Meningitis	0•0065	0•10%	0•0093	0•0039
Inflammatory_heart_disease	0•0064	0•10%	0•0070	0•0064
Benign_prostatic_hypertrophy_Symptomatic	0•0057	0•09%	0•0112	0•0000
HIV_AIDS	0•0042	0•07%	0•0072	0•0024
Anorectal_atresia_Cases	0•0042	0•07%	0•0075	0•0011
Dengue	0•0038	0•06%	0•0042	0•0033
Migraine_Cases	0•0033	0•05%	0•0031	0•0037
Renal_agenesis_Cases	0•0032	0•05%	0•0032	0•0034
Gonorrhoea	0•0032	0•05%	0•0022	0•0044
Macular_degeneration_Low_vision	0•0030	0•05%	0•0030	0•0033
Malaria	0•0028	0•04%	0•0025	0•0030
Dental_caries_Episodes	0•0023	0•04%	0•0029	0•0020
Peptic_ulcer_disease	0•0022	0•04%	0•0027	0•0019
Low_birth_weight	0•0017	0•03%	0•0019	0•0020
Otitis_media_Chronic_infection	0•0016	0•03%	0•0016	0•0016
Bipolar_affective_disorder	0•0015	0•02%	0•0014	0•0013
Hearing_loss_adult_onset_Mild	0•0014	0•02%	0•0016	0•0013
Tetanus_Episodes	0•0010	0•02%	0•0007	0•0014
Anencephaly_Cases	9•19E-04	0•01%	1•07E-03	7•12E-04
Pertussis	7•36E-04	0•01%	4•85E-04	1•11E-03
Hepatitis_C_Episodes	6•48E-04	0•01%	1•03E-03	3•84E-04
Alcohol_use_disorders	5•40E-04	0•01%	1•16E-03	1•29E-05
Syphilis	5•16E-04	0•01%	3•61E-04	7•10E-04
Periodontal_disease_Cases	4•96E-04	0•01%	5•30E-04	5•06E-04
Trachoma	3•77E-04	0•01%	2•77E-04	5•18E-04
Spina_bifida_Cases	3•01E-04	0•00%	3•73E-04	2•86E-04
Chlamydia	2•90E-04	0•00%	3•63E-04	2•92E-04
Maternal_haemorrhage	2•71E-04	0•00%	0•00E+00	3•92E-04
Trypanosomiasis	2•34E-04	0•00%	4•42E-04	4•12E-05
Measles_Episodes	2•29E-04	0•00%	3•33E-04	1•45E-04
Protein_energy_malnutrition	2•08E-04	0•00%	2•00E-04	2•25E-04
Cleft_palate_Cases	1•80E-04	0•00%	2•39E-04	1•43E-04
Cleft_lip_Cases	1•38E-04	0•00%	2•20E-04	9•96E-05
Oesophageal_atresia_Cases	1•22E-04	0•00%	2•30E-04	0•00E+00
Lymphatic_filariasis	9•53E-05	0•00%	5•49E-05	1•32E-04
Obsessive_compulsive_disorder_Cases	7•82E-05	0•00%	1•81E-04	3•30E-05
Down_syndrome_Cases	7•59E-05	0•00%	0•00E+00	1•75E-04
Japanese_Encephalitis	7•53E-05	0•00%	0•00E+00	1•22E-04
Obstructed_labour	6•07E-05	0•00%	0•00E+00	9•16E-05
Post_traumatic_stress_disorder_Cases	4•74E-05	0•00%	0•00E+00	1•02E-04
Ascariasis	4•33E-05	0•00%	4•23E-05	5•31E-05
Vitamin_A	3•38E-05	0•00%	3•65E-05	3•25E-05
Schistosomiasis	2•50E-05	0•00%	4•71E-05	6•54E-07
Leishmaniasis	1•95E-05	0•00%	6•94E-05	0•00E+00
Abortion	5•19E-06	0•00%	0•00E+00	7•21E-06
Hookworm_disease	6•54E-07	0•00%	1•32E-06	0•00E+00
Trichuriasis	5•23E-07	0•00%	5•07E-07	6•13E-07
Abdominal_wall_defect_Cases	0•00E+00	0•00%	0•00E+00	0•00E+00
Hypertensive_disorders_of_pregnancy	0•00E+00	0•00%	0•00E+00	0•00E+00
Iodine_deficiency	0•00E+00	0•00%	0•00E+00	0•00E+00
Leprosy	0•00E+00	0•00%	0•00E+00	0•00E+00
Maternal_sepsis	0•00E+00	0•00%	0•00E+00	0•00E+00
Mild_mental_retardation	0•00E+00	0•00%	0•00E+00	0•00E+00
Onchocerciasis	0•00E+00	0•00%	0•00E+00	0•00E+00

References

[1] Mathers CD, Vos T, Lopez AD, Salomon JA. National burden of disease Studies: a practical guide. Global Program on Evidence for Health Policy; 2001. 2001.

[2] Oortwijn WJ, Mathijssen J, Lankhuizen M, Cave J. Evaluating the Uptake of the Healthy Life Years Indicator: Final report. RAND Corporation; 2007.

[3] Sullivan DF. A single index of mortality and morbidity. HSMHA Health Rep 1971;86(4):347–54.

[4] Nguyen L, Saito Y, Phan H, Nguyen L. Health expectancy and its variations in Vietnam. In: Proceedings of the REVES meeting in Taichung, Taiwan; 2012. 2012.

[5] Robine J M, Jagger C. Creating a coherent set of indicators to monitor health across Europe - The Euro-REVES 2 project. . The European Journal of Public Health 2003;13(3 Suppl):6–14.

[6] Katz S C, Ford A B, Moskowitz R W, et al. Studies of illness in the aged. The index of Adl: a standardized measure of biological and psychosocial function. JAMA J Am Med Assoc 1963;185(12):914–19.

[7] Lawton M P, Brody E M. Assessment of older people: self-maintaining and instrumental activities of daily living. The Gerontologist 1969;9(3 Part 1):179–86.

[8] Mathers CD. Disability-free and handicap-free life expectancy in Australia 1981 and 1988. Canberra: Australian Institute of Health; 1991.

[9] van Oyen H, Van der Heyden J, Perenboom R, Jagger C. Monitoring population disability: evaluation of a new Global Activity Limitation Indicator (GALI). Soz Praventivmed 2006;51(3):153–61.

[10] Riley L, Guthold R, Cowan M, et al. The World Health Organization STEPwise approach to noncommunicable disease risk-factor surveillance: methods, challenges, and opportunities. Am J Public Health 2016;106(1):74–8.

[11] The global burden of disease: 2004 update. Irish Med J 2008;106(1):4.

[12] Piri SM, Saeedi Moghaddam S, Ghodsi Z, et al. Trend of appendicitis mortality at national and provincial levels in Iran from 1990 to 2015. Arch Iran Med 2020;23(5):302–11.

[13] Nauman J, Soteriades ES, Hashim MJ, et al. Global incidence and mortality trends due to adverse effects of medical treatment, 1990-2017: a systematic analysis from the global burden of diseases, injuries and risk factors study. Cureus 2020;12(3):e7265.

[14] Westergaard D, Moseley P, Sorup FKH, Baldi P, Brunak S. Population-wide analysis of differences in disease progression patterns in men and women. Nat Commun 2019;10(1):666.

[15] Goldstein BA, Navar AM, Pencina MJ, Ioannidis JP. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc 2017;24(1):198–208.

[16] Jagannatha AN, Yu H. Bidirectional RNN for medical event detection in electronic health records. Proc Conf 2016;2016:473–82.

[17] Liu K, Hu Q, Liu J, Xing C. Named entity recognition in Chinese electronic medical records based on CRF. In: Proceedings of the 14th web information systems and applications conference (WISA); 2017 11-12 Nov. 2017; 2017. p. 105–10.

[18] Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Informat Assoc 2010;17:507–13.

[19] Wu Y, Lei J, Wei W Q, et al. Analyzing differences between chinese and english clinical text: a cross-institution comparison of discharge summaries in two languages. Stud Health Technol Informat 2013;192(1):662–6.

[20] Yang JF, Guan Y, He B, Qu CY, Yu QB, Liu YX, Zhao YJ. Corpus construction for named entities and entity relations on Chinese electronic medical records. Ruan Jian Xue Bao/J Softw 2016;27(11):2725–46. in Chinese http://www.jos.org.cn/1000-9825/4880.htm .

[21] Zhang L, Li J, Wang C. Automatic synonym extraction using Word2Vec and spectral clustering. In: Proceedings of the 36th Chinese control conference (CCC); 2017 26-28 July 2017; 2017. p. 5629–32.

[22] Wang C, Cao L, Zhou B. Medical synonym extraction with concept space models. In: Proceedings of the 24th international conference on artificial intelligence. AAAI Press; 2015. p. 989–95.

[23] World Bank World development report 1993. Investing in health. New York: Oxford University Press for the World Bank; 1993.

[24] Murray CJL, Lopez AD. Evidence-based health policy – lessons from the Global Burden of Disease Study. Science 1996;274:740–3.

[25] Murray CJL, Lopez AD, editors. The global burden of disease: a comprehensive assessment of mortality and disability from diseases, injuries and risk factors in 1990 and projected to 2020. Cambridge. Harvard School of Public Health on behalf of the World Health Organization and the World Bank; 1996.

[26] Murray CJL, Lopez AD. Global health statistics. Cambridge, Harvard School of Public Health on behalf of the World Health Organization and the World Bank, 1996.

[27] Chiang CL. The life table and its applications. Malabar, FL, USA: Robert E Krieger Publishing Company; 1984.

[28] Hu SB, Wang F, Yu CH. Evaluation and estimation of the provincial infant mortality rate in China's Sixth Census. Biomed Environ Sci 2015;28(6):410–20.

[29] Samuel C, David S. Contemporary model life tables for developed countries an application of model-based clustering; 2011.

[30] Yi Z, Vaupel JW. Oldest-Old Mortality in China. Human longevity, individual life duration, and the growth of the oldest-old population. Robine J-M, Crimmins EM, Horiuchi S, Yi Z, editors. Dordrecht: Springer; 2007.

[31] International statistical classification of diseases and related health problems. Tenth revision. Statistical inference and simulation for spatial point processes. Chapman and Hall/CRC; 2016.

[32] 2016 Chongqing health and family planning statistical yearbook. Beijing: Chongqing Municipal Health and Family Planning Commission; 2017.

[33] WHO Global burden of disease 2004 Update: disability weights for diseases and conditions. Health statistics and information systems; 2004.

[34] Mu J, Bhat S, Viswanath P. All-but-the-Top: simple and effective postprocessing for word representations. arXiv:170201417 [cs, stat] 2018; published online March 19. http://arxiv.org/abs/1702.01417 (Accessed 11 September 2020).

[35] Zhou M, Li Y, Wang H, et al. Analysis of life expectancy and health adjusted life expectancy in China by province from 1990 to 2015. Zhonghua Liu Xing Bing Xue Za Zhi 2016;37(11):1439–43.

[36] He C. Decomposition of changes in health-adjusted life expectancy in China, 1990-2013. Popul Res 2020;44(1):26–38.

[37] Hu X, Sun X, Li Y, et al. Potential gains in health-adjusted life expectancy from reducing four main non-communicable diseases among Chinese elderly. BMC Geriatrics 2019;19(1):16.

[38] Manuel DG, Schultz SE, Kopec JA. Measuring the health burden of chronic disease and injury using health adjusted life expectancy and the Health Utilities Index. J Epidemiol Community Health 2002;56(11):843–50.

[39] Lang JJ, Alam S, Cahill LE, et al. Global burden of disease study trends for Canada from 1990 to 2016. Can Med Assoc J 2018;190(44):E1296.

[40] Zhou M, Wang H, Zhu J, et al. Cause-specific mortality for 240 causes in China during 1990-2013: a systematic subnational analysis for the Global Burden of Disease Study 2013. Lancet 2016;387(10015):251–72.

[41] Ma J, Guo X, Xu A, Zhang J, Jia C. Epidemiological analysis of injury in Shandong Province, China, 8. BMC Public Health; 2008. p. 122.

[42] Chen H, Chen G, Zheng X, Guo Y. Contribution of specific diseases and injuries to changes in health adjusted life expectancy in 187 countries from 1990 to 2013: retrospective observational study. BMJ 2019;364:l969.

[43] DALYs GBD, Collaborators H. Global, regional, and national disability-adjusted life-years (DALYs) for 359 diseases and injuries and healthy life expectancy (HALE) for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet 2018;392(10159):1859–922.

[44] Wang Y, O'Neil A, Jiao Y, et al. Sex differences in the association between diabetes and risk of cardiovascular disease, cancer, and all-cause and cause-specific mortality: a systematic review and meta-analysis of 5,162,654 participants. BMC Med 2019;17(1):136.

[45] Ferrari AJ, Charlson FJ, Norman RE, et al. Burden of depressive disorders by country, sex, age, and year: findings from the global burden of disease study 2010. PLoS Med 2013;10(11):e1001547.

[46] Steensma C, Loukine L, Orpana H, et al. Comparing life expectancy and health-adjusted life expectancy by body mass index category in adult Canadians: a descriptive study. Popul Health Metr 2013;11(1):21.

[47] Zhang T, Shi W, Huang Z, Gao D, Guo Z, Chongsuvivatwong V. Gender and ethnic health disparities among the elderly in rural Guangxi, China: estimating quality-adjusted life expectancy. Glob Health Action 2016;9:32261.

[48] Beeksma M, Verberne S, van den Bosch A, Das E, Hendrickx I, Groenewoud S. Predicting life expectancy with a long short-term memory recurrent neural network using electronic medical records. BMC Med Inform Decis Mak 2019;19(1):36.

[49] Jonker MF, Congdon PD, van Lenthe FJ, Donkers B, Burdorf A, Mackenbach JP. Small-area health comparisons using health-adjusted life expectancies: a Bayesian random-effects approach. Health Place 2013;23:70–8.

[50] World Health Organization WHO methods and data sources for global burden of disease estimates 2000-2015. Geneva, Switzerland: Department of Information, Evidence and Research WHO; 2017.

[51] Salomon JA, Vos T, Hogan DR, Gagnon M, Naghavi M, Mokdad A, et al. Common values in assessing health outcomes from disease and injury: disability weights measurement study for the Global Burden of Disease Study 2010. Lancet 2012;380:2129–43.

[52] Yu S, Liao KP, Shaw SY, et al. Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources. J Am Med Informat Assoc 2015;22:993–1000.