Available online at www.sciencedirect.com

ScienceDirect

Biomedical Journal

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

Original Article

Chang Gung Research Database: A multi-institutional database consisting of original medical records

Ming-Shao Tsai [a,b], *Meng-Hung Lin* [b], *Chuan-Pin Lee* [b], *Yao-Hsu Yang* [b,c,d,e], *Wen-Cheng Chen* [f,g], *Geng-He Chang* [a], *Yao-Te Tsai* [a], *Pau-Chung Chen* [d,h,*], *Ying-Huang Tsai* [i,j,**]

[a] *Department of Otolaryngology – Head and Neck Surgery, Chang Gung Memorial Hospital, Chiayi, Taiwan*
[b] *Center of Excellence for Chang Gung Research Datalink, Chang Gung Memorial Hospital, Chiayi, Taiwan*
[c] *Department of Traditional Chinese Medicine, Chang Gung Memorial Hospital, Chiayi, Taiwan*
[d] *Institute of Occupational Medicine and Industrial Hygiene, National Taiwan University College of Public Health, Taipei, Taiwan*
[e] *School of Traditional Chinese Medicine, College of Medicine, Chang Gung University, Taoyuan, Taiwan*
[f] *Department of Radiation Oncology, Chang Gung Memorial Hospital, Chiayi, Taiwan*
[g] *College of Medicine, Chang Gung University, Taoyuan, Taiwan*
[h] *Department of Environmental and Occupational Medicine, National Taiwan University Hospital and National Taiwan University College of Medicine, Taipei, Taiwan*
[i] *Division of Thoracic Oncology, Department of Pulmonary and Critical Care Medicine, Chang Gung Memorial Hospital, Chiayi, Taiwan*
[j] *Department of Respiratory Care, College of Medicine, Chang Gung University, Taoyuan, Taiwan*

ARTICLE INFO

Article history:
Received 27 May 2016
Accepted 8 August 2017
Available online 10 November 2017

Keywords:
Chang Gung Memorial Hospital
Medical center
Taiwan
Comorbidities
Coverage

ABSTRACT

Background: The Chang Gung Research Database (CGRD) is a de-identified database derived from original medical records of Chang Gung Memorial Hospital (CGMH), which comprises seven medical institutes located from the northeast to southern regions of Taiwan. The volume of medical services performed in CGMH is large, and clinical and scientific studies based on the CGRD are reported to be of high quality. However, the CGRD as a useful database for research has not been analyzed before. The objective of the study was to analyze the CGRD with regard to its characteristics and coverage of Taiwan's population.
Methods: We performed a nationwide cohort study using population-based data from the Taiwan National Health Insurance Research Database (NHIRD). All patients who had any medical record of outpatient visits or admission between January 1, 1997, and December 31, 2010, were included, and the sex ratio, age distribution, socioeconomic status, urbanicity, severity of illness, prevalence of specific disease, and coverage of the CGRD were analyzed.

* *Corresponding author*. Institute of Occupational Medicine and Industrial Hygiene, National Taiwan University College of Public Health, Room 733, 17, Syujhou Rd., Taipei 10055, Taiwan.
** *Corresponding author*. Division of Thoracic Oncology, Department of Pulmonary and Critical Care Medicine, Chang Gung Memorial Hospital, Chiayi. 6, Sec. W., Jiapu Rd., Puzi City, Chiayi County 613, Taiwan.
E-mail addresses: pchen@ntu.edu.tw (P.-C. Chen), chestmed@cgmh.org.tw (Y.-H. Tsai).
Peer review under responsibility of Chang Gung University.
http://dx.doi.org/10.1016/j.bj.2017.08.002

Results: The sex ratio, age distribution, socioeconomic status, and urbanicity of the population of the CGRD are different from those of Taiwan NHIRD and medical centers in Taiwan (all the pairwise $p < 0.05$). The severity of comorbidities, and prevalence of specific diseases of the population of the CGRD are significantly higher than those of Taiwan NHIRD and medical centers in Taiwan for both outpatient and inpatient samples (all the pairwise $p < 0.05$). The overall coverage of the CGRD was 21.2% for outpatients and 12.4% for inpatients. The disease-specific coverage of the CGRD was 27–34% for outpatients and 14–21% for inpatients.

Conclusions: The CGRD is a multi-institutional, original medical record-based research database with high overall and disease-specific coverage of Taiwan. The population of the CGRD has significantly higher severity of comorbidities, and prevalence of specific diseases than those of Taiwan NHIRD and medical centers in Taiwan.

At a glance commentary

Scientific background on the subject

The Chang Gung Research Database (CGRD) is a database derived from original medical records of Chang Gung Memorial Hospital (CGMH). The volume of medical services in CGMH is large, and studies based on the CGRD are reported to be of high quality. However, the CGRD has not been analyzed before.

What this study adds to the field

This study indicates the CGRD is a multi-institutional, original medical record-based research database with high overall and disease-specific coverage of Taiwan. The population of the CGRD has significantly higher severity of comorbidities, and prevalence of specific diseases than those of Taiwan NHIRD and medical centers in Taiwan.

The Chang Gung Research Database (CGRD) is a de-identified database derived from medical records of Chang Gung Memorial Hospital (CGMH), and it is systematically updated annually to include new data generated in CGMH. CGMH, founded in 1976, is currently the largest hospital system in Taiwan, and it comprises seven medical institutes, which are located from the northeast to southern regions of Taiwan: Keelung CGMH, Taipei CGMH, Linkou CGMH, Taoyuan CGMH, Yunlin CGMH, Chiayi CGMH, and Kaohsiung CGMH. CGMH has 10,070 beds and admits more than 280,000 patients each year. The outpatient department visits and emergency department visits to CGMH were over 8,500,000 and 500,000, respectively in 2015 [1]. In recent years, the CGRD promoted clinical and scientific studies to a considerable extent. In 2015, more than 1800 studies were conducted by CGMH staff, and the studies were published in a diverse range of reputed journals. Some of these studies are based on the CGRD as multicenter research studies with relatively large sample sizes [2,3].

Although the CGRD is a medical record database with large volumes of data that are useful to perform several research studies and analysis, its characteristics and coverage have never been reported before. The objective of this study was to analyze the CGRD with regard to its characteristics and coverage of Taiwan's population.

Methods

Study population

We performed a nationwide cohort study using population-based data from the Taiwan National Health Insurance Research Database (NHIRD), one of the largest administrative health-care databases worldwide. The National Health Insurance (NHI) program, implemented on March 1, 1995, is a single-payer compulsory and universal health insurance plan, and it now covers all forms of health-care services for over 99% of Taiwan's residents [4–8]. This high coverage rate (nearly 100%) enables studies based on the NHIRD to be nationwide and population-based. For researchers' convenience, the National Health Research Institutes (NHRI) of Taiwan sampled a representative database of 1 million patients from the year 2000 registry of all NHI enrollees ($n = 23{,}753{,}407$) by using a systematic and random sampling method (Longitudinal Health Insurance Database, LHID2000) [6]. In LHID2000, no statistically significant differences exist in age, sex, or health-care costs between the sample group and all enrollees, according to NHRI Reports [6]. In this study, we used these databases for estimating outpatient visits and admissions of the sample cohort, as both include information about patient characteristics including sex, date of birth, date of admission, date of discharge, dates of outpatient visits, and up to three outpatient visit diagnoses and five discharge diagnoses [based on the International Classification of Diseases, Ninth Revision (ICD-9) classification] [9]. These databases have previously been used for various scientific studies, and the information they provide about diagnoses, hospitalizations, and prescription use has been proved to be of high quality [4,10–12]. The identification numbers of all patients in the NHIRD were encrypted to protect their privacy. The Ethics Review Board of our institution approved the study (CGMH-IRB No. 201600346B0).

Study design

The flowchart of the patient enrollment process of study cohort is presented in Fig. 1. The study cohort was identified

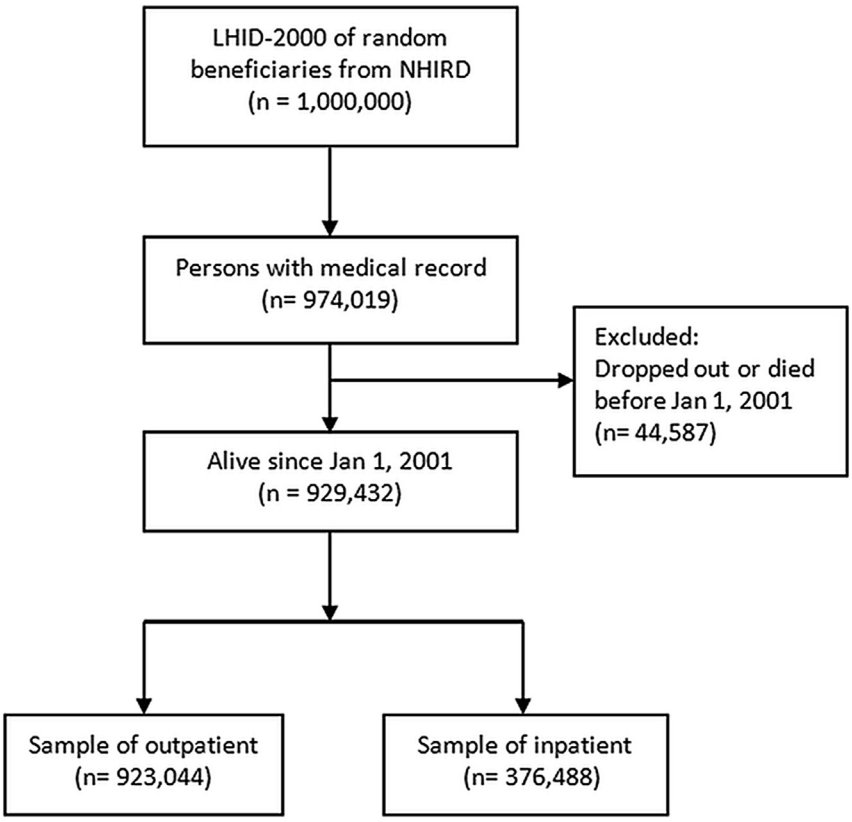

Fig. 1 Flowchart of the patient enrollment process of study cohort.

from LHID2000, which comprise 1 million beneficiaries. Complete data of all these beneficiaries were recorded from January 1, 1997, to December 31, 2010. We conducted a nationwide population-based cohort study that included all patients who had any medical record of outpatient visits or admission between January 1, 1997, and December 31, 2010 ($n = 974,019$). We excluded patients who died or dropped out from the NHI scheme before January 1, 2001, to allow for more than 4 years of study period. On the basis of this criterion, data on 44,587 patients were excluded. Finally, data on 929,432 patients were included in this study.

The claim data of each outpatient and inpatient visits sourced from the LHID2000 (a subset database of NHIRD) included the variables of "hospital grade" and "hospital ID". Patients with outpatient visits were categorized into three groups: Group 1, NHIRD (all patients); Group 2, Center [patients with record of outpatient visit(s) in any medical center]; Group 3, CGRD [patients with record of outpatient visit(s) in CGMH]. Accordingly, we also grouped patients with record of admission into Group 1 (NHIRD), Group 2 (Center), and Group 3 (CGRD).

Patients' sociodemographic information including age, sex, socioeconomic status, and urbanicity level was obtained from enrollment data files in LHID2000. The sociodemographic information such as age, socioeconomic status, and urbanicity level might change over time. Thus, to define the patients' sociodemographic information in the same period is necessary for fair comparison. In this study, we used a subset database of NHIRD – LHID2000 which contained one million enrollnees. All the enrollnees in LHID2000 had initial registry data of NHIRD during 1996–2000 which was used to define their sociodemographic information. The enrollee category (EC) of all patients was registered in National Health Insurance Research Database (NHIRD). This study used EC as a measure of socioeconomic status to classify beneficiaries into four subgroups (EC 1–EC 4): EC 1 (civil servants, full-time, or regularly paid personnel in governmental agencies and public schools, and self-employed people); EC 2 (employees of privately owned enterprises or institutions); EC 3 (other employees or paid personnel, and members of the farmers' or fishers' associations); and EC 4 (substitute service draftees, members of low-income families, and veterans). On average, the payroll-related amount for health insurance was highest for EC 1, followed by EC 2, EC 3, and EC 4 [13]. To determine the level of urbanicity among the study cohort, we classified all 359 townships in Taiwan into three categories: urban, suburban, and rural areas. The urbanicity classification was based on five indices: population density, percentages of residents with college or higher education, percentages of residents aged >65 years, percentages of residents who were agriculture workers, and the number of physicians per 100,000 people [14]. We used Deyo's Charlson Comorbidity Index (CCI) to evaluate the comorbidities of patients, and the CCI score was calculated from January 1, 1997, to December 31, 2010, using

the ICD-9 coding system [15]. The leading causes of death, most common catastrophic diseases and chronic diseases in Taiwan were also studied and defined as follows: cancer (ICD-9 codes as 140–208), liver cirrhosis (ICD-9 codes 571.2, 571.5, 571.6, 572.2, 572.3, 572.4, 572.8, and 573.0), stroke (ICD-9 codes 430–438), coronary artery disease (CAD; ICD-9 codes 410–414), chronic kidney disease (CKD; ICD-9 codes 582, 583, 585, 586, and 588), hypertension (ICD-9 codes 401–405), diabetes mellitus (DM; ICD-9 codes 250), hyperlipidemia (ICD-9 codes 272), chronic obstructive pulmonary disease (COPD; ICD-9 codes 491, 492, 496), asthma (ICD-9 codes 493), pneumonia (ICD-9 codes 480–486 and 507.0–507.8), hepatitis B (ICD-9 codes 070.2, 070.3, and V02.61), hepatitis C (ICD-9 codes 070.7, 070.41, 070.44, 070.51, 070.54, and V02.62), mental disease (ICD-9 codes 290–319) [16,17].

Statistical analysis

Basic demographic data were summarized as *n* (%) for categorical variables and mean with standard deviation for continuous variables. Distribution of comorbidities was tabulated as *n* (%). We used Chi-square test for comparison between three groups, and used Fisher's exact test for pairwise comparison between each group. SAS version 9.4 (SAS Inc., Cary, NC, USA) was used for statistical analysis.

Results

Patient characteristics

A total of 929,432 patients were included as the study cohort; among the patients, 923,044 had outpatient visits (outpatient samples) and 376,448 had admission to a hospital (inpatient samples). The demographic characteristics of the patients are presented in Table 1. The sex ratio, age distribution, socio-economic status, and urbanicity of the population of the CGRD are different from those of Taiwan NHIRD and medical centers in Taiwan for both outpatient and inpatient samples (all the pairwise $p < 0.05$).

For the outpatient samples, the case numbers for Group 1, Group 2, and Group 3 were 923,044, 552,451, and 195,529, respectively. The sex ratio (i.e., male-to-female ratio) was 50.7/49.3 in Group 1, 47.8/52.2 in Group 2, and 46.6/53.4 in Group 3. Both Groups 2 and 3 showed 3–4% female predominance compared with Group 1. The mean age of patients in Groups 1, 2, and 3 was 29.3 ± 19.8, 31.2 ± 19.9, and 31.9 ± 20.0 years, respectively. The mean age and age distribution in Group 2 were close to those in Group 3; compared with Group 1, the other two groups had a slight increase in patients in the 40–69-year age group (approximately 2%) and a decline in those in the 0–9-year age group (approximately

Table 1 Baseline characteristics for outpatient and inpatient in NHIRD, Center, and CGRD.

	Outpatient			Inpatient		
	NHIRD	Center	CGRD	NHIRD	Center	CGRD
	n (%)	*n* (%)	*n* (%)	*n* (%)	*n* (%)	*n* (%)
Total	923,044	552,451	195,529	376,448	164,945	46,520
Gender						
Male	468,155 (50.7)	264,304 (47.8)	91,048 (46.6)	180,732 (48.0)	82,795 (50.2)	23,092 (49.6)
Female	454,889 (49.3)	288,147 (52.2)	104,481 (53.4)	195,716 (52.0)	82,150 (49.8)	23,428 (50.4)
Age						
Mean (SD)	29.3 (19.8)	31.2 (19.9)	31.9 (20.0)	36.0 (21.1)	39.0 (21.1)	38.7 (21.0)
0–9	183,622 (19.9)	92,165 (16.7)	32,293 (16.5)	43,754 (11.6)	16,963 (10.3)	5219 (11.2)
10–19	123,242 (13.4)	75,376 (13.6)	22,804 (11.7)	42,053 (11.2)	14,170 (8.6)	3458 (7.4)
20–29	179,398 (19.4)	104,186 (18.9)	36,879 (18.9)	73,755 (19.6)	26,531 (16.1)	7257 (15.6)
30–39	168,971 (18.3)	98,695 (17.9)	35,279 (18.0)	60,079 (16.0)	26,627 (16.1)	7673 (16.5)
40–49	118,407 (12.8)	77,257 (14.0)	28,730 (14.6)	51,874 (13.8)	25,770 (15.6)	7398 (15.9)
50–59	64,808 (7.0)	45,407 (8.2)	18,068 (9.2)	38,510 (10.2)	20,043 (12.2)	6144 (13.2)
60–69	56,588 (6.1)	40,918 (7.4)	15,543 (8.0)	42,553 (11.3)	23,052 (14.0)	6494 (14.0)
70–79	23,380 (2.5)	16,056 (2.9)	5276 (2.7)	19,953 (5.3)	10,203 (6.2)	2508 (5.4)
≥80	4628 (0.5)	2391 (0.4)	657 (0.3)	3917 (1.0)	1586 (1.0)	369 (0.8)
EC						
1	67,561 (7.3)	46,060 (8.3)	14,662 (7.5)	25,458 (6.8)	13,204 (8.0)	2981 (6.4)
2	433,418 (47.0)	262,002 (47.4)	92,847 (47.5)	158,484 (42.1)	69,626 (42.2)	19,191 (41.3)
3	287,917 (31.2)	163,103 (29.5)	63,446 (32.5)	129,518 (34.4)	53,829 (32.6)	18,150 (39.0)
4	134,148 (14.5)	81,286 (14.7)	24,574 (12.6)	62,988 (16.7)	28,286 (17.2)	6198 (13.3)
Urbanicity						
Urban	282,100 (30.6)	199,443 (36.1)	64,091 (32.8)	106,263 (28.2)	57,842 (35.1)	12,486 (26.8)
Suburban	436,915 (47.3)	249,791 (45.2)	94,549 (48.4)	175,537 (46.6)	73,789 (44.7)	23,202 (49.9)
Rural	204,029 (22.1)	103,217 (18.7)	36,889 (18.9)	94,648 (25.1)	33,314 (20.2)	10,832 (23.3)

Abbreviation: SD: standard deviation.
Enrollee category (EC) as a proxy measure of socio-economic status (SES).
Chi-square test for comparison between three groups. All the *p*-value were <0.05.
Fisher's exact test for pairwise comparison between each group. All the pairwise $p < 0.05$.

3%). The urbanicity level in Group 1 was close to that in Group 3, and both groups had a reduction in urban population (3.3–5.5%), compared with Group 2.

Regarding the inpatient samples, the case numbers for Groups 1, 2, and 3 were 376,488, 164,945, and 46,520, respectively. The sex ratio (i.e., male-to-female ratio) was 48.0/52.0 in Group 1, 50.2/49.8 in Group 2, and 49.6/50.4 in Group 3. Group 1 showed 1–2% female predominance compared with Groups 2 and 3. The mean ages of patients in Groups 1, 2, and 3 were 36.0 ± 21.1, 39.0 ± 21.1, and 38.7 ± 21.0 years, respectively. The mean age and age distribution in Group 2 were close to those in Group 3; compared with Group 1, the other two groups had a slight increase in patients in the 40–69-year age group (nearly 2–3%) and a drop in those in the 10–29-year age group (approximately 3–4%). The urbanicity level in Group 1 was close to that in Group 3, and both groups had a slight increase in suburban and rural populations (2–5%) and a reduction in urban population (approximately 7–8%) compared with Group 2.

Severity and prevalence of comorbidities

The patient comorbidities are listed in Table 2. The severity of comorbidities, and prevalence of specific diseases of the population of the CGRD are significantly higher than those of Taiwan NHIRD and all medical centers in Taiwan for both outpatient and inpatient samples (all the pairwise $p < 0.05$). Regarding the outpatient samples, Groups 3 was determined to be tended to have a higher CCI score and higher prevalence of specific comorbidities than did Group 1 and Group 2. In Group 3, the prevalence rates of cancer, hepatitis C, and liver cirrhosis were 1.6 times higher than those in Group 1. Other diseases also showed higher prevalence rates (CKD was 1.5 times higher; stroke, COPD, asthma, DM, hyperlipidemia, CAD, and mental disease were 1.4 times higher; and hepatitis B, hypertension, and pneumonia were 1.3 times higher) in Group 3, compared with Group 1.

For the inpatient samples, Groups 3 was observed to be tended to have a higher CCI score and higher prevalence rates of specific comorbidities than did Group 1 and Group 2. In Group 3, the prevalence rates of liver cirrhosis, cancer, hepatitis C, and CKD were 1.7, 1.6, 1.5, and 1.4 times higher than those in Group 1. Other diseases also demonstrated higher prevalence rates (DM was 1.3 times higher; stroke, COPD, hyperlipidemia, hepatitis B, and hypertension were 1.2 times higher; and asthma, CAD, mental disease, and pneumonia were 1.1 times higher) in Group 3 compared with Group 1.

Population coverage

The overall and disease-specific coverage rates of the CGRD and medical centers are illustrated in Fig. 2. For the outpatient samples, the overall coverage rate of the CGRD was 21.2% and the disease-specific coverage rate was 27–34%. The disease with the highest coverage rates were cancer (34%), liver

Table 2 Comorbidities for outpatient and inpatient in NHIRD, Center, and CGRD.

	Outpatient			Inpatient		
	NHIRD	Center	CGRD	NHIRD	Center	CGRD
	n (%)	*n* (%)	*n* (%)	*n* (%)	*n* (%)	*n* (%)
Total	923,044	552,451	195,529	376,448	164,945	46,520
CCIs						
≤4	835,669 (90.5)	482,396 (87.3)	166,831 (85.3)	299,804 (79.6)	119,517 (72.5)	32,824 (70.6)
>4	87,375 (9.5)	70,055 (12.7)	28,698 (14.7)	76,644 (20.4)	45,428 (27.5)	13,696 (29.4)
Disease						
Cancer	49,537 (5.4)	42,212 (7.6)	17,043 (8.7)	44,141 (11.7)	30,248 (18.3)	8855 (19.0)
Liver cirrhosis	15,351 (1.7)	11,836 (2.1)	5196 (2.7)	13,753 (3.7)	8162 (5.0)	2874 (6.2)
Hepatitis C	12,630 (1.4)	9622 (1.7)	4242 (2.17)	9660 (2.6)	5135 (3.1)	1829 (3.9)
CKD	38,264 (4.2)	30,124 (5.5)	12,559 (6.4)	30,175 (8.0)	17,585 (10.7)	5354 (11.5)
Stroke	66,742 (7.2)	50,266 (9.1)	19,721 (10.1)	55,976 (14.9)	29,816 (18.1)	8653 (18.6)
COPD	85,196 (9.2)	63,019 (11.4)	25,210 (12.9)	59,705 (15.9)	30,173 (18.3)	8606 (18.5)
Asthma	85,704 (9.3)	60,809 (11.0)	25,584 (13.1)	49,955 (13.3)	22,814 (13.8)	7039 (15.1)
Diabetes mellitus	100,013 (10.8)	73,364 (13.3)	28,681 (14.7)	69,083 (18.4)	36,389 (22.1)	10,784 (23.2)
Hyperlipidemia	142,059 (15.4)	105,737 (19.1)	41,635 (21.3)	81,850 (21.7)	42,392 (25.7)	12,130 (26.1)
CAD	94,152 (10.2)	72,175 (13.1)	27,353 (14.0)	68,935 (18.3)	36,864 (22.4)	9677 (20.8)
Mental disease	220,716 (23.9)	164,655 (29.8)	65,045 (33.3)	128,946 (34.3)	61,536 (37.3)	17,423 (37.5)
Hepatitis B	33,091 (3.6)	24,987 (4.5)	9236 (4.7)	18,516 (4.9)	9684 (5.9)	2799 (6.0)
Hypertension	196,365 (21.3)	139,939 (25.3)	52,781 (27.0)	126,519 (33.6)	65,381 (39.6)	18,547 (39.9)
Pneumonia	90,154 (9.8)	63,868 (11.6)	24,405 (12.5)	70,359 (18.7)	33,913 (20.6)	9939 (21.4)

Abbreviations: CCI: Charlson Comorbidity Index; CAD: coronary artery disease; COPD: chronic obstructive pulmonary disease; CKD: chronic kidney disease.

Chi-square test for comparison between three groups. All the *p*-value were <0.05.

Fisher's exact test for pairwise comparison between each group. All the pairwise $p < 0.05$, except hypertension, hyperlipidemia, COPD, hepatitis B, Mental disease for the inpatient samples between Center and CGRD ($p > 0.05$).

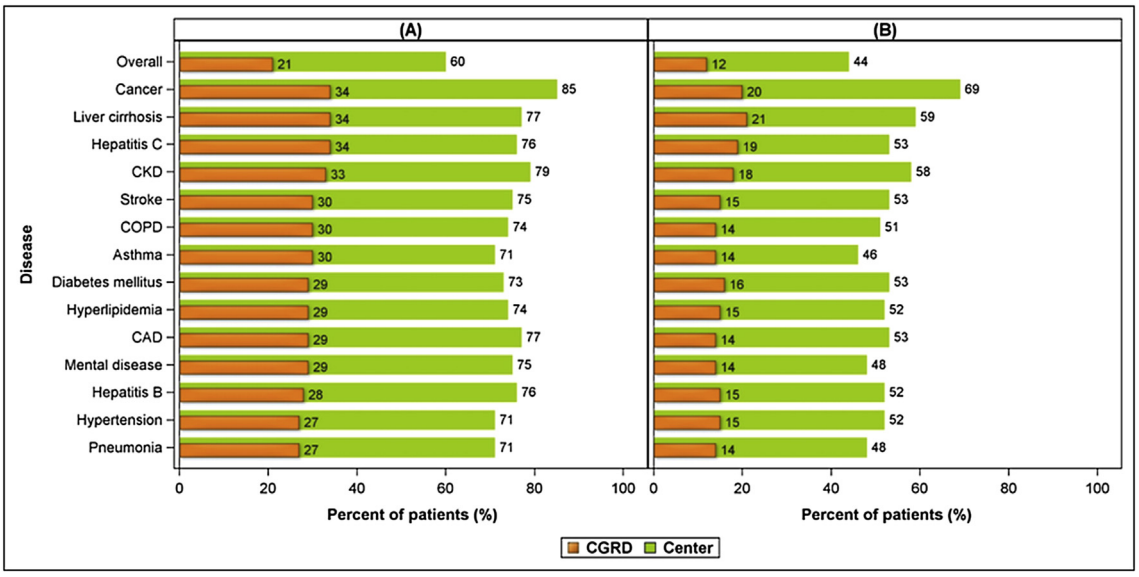

Fig. 2 The coverage of the CGRD and medical centers for the outpatients (A) and inpatients (B) samples.

cirrhosis (34%), hepatitis C (34%), and CKD (33%). Regarding the inpatient samples, the overall coverage rate of the CGRD was 12.4% and the disease-specific coverage rate was 14–21%. The disease with the highest coverage rates were liver cirrhosis (21%), cancer (20%), hepatitis C (19%), and CKD (18%).

Discussion

To the best of our knowledge, this is the first study to evaluate the characteristics and coverage of a database of a private hospital system. The study results reveal the sex ratio, age distribution, socioeconomic status, urbanicity, severity of comorbidities, and prevalence of diseases of the population of the CGRD are different from those of Taiwan NHIRD and medical centers in Taiwan. Population of CGRD is tended to have a higher CCI score and higher prevalence of specific comorbidities than those of NHIRD and medical centers in Taiwan. The findings reveal CGRD could be a suitable database for studying patients with more severe and complicated diseases. The overall coverage of the CGRD was 21.2% for outpatients and 12.4% for inpatients (Fig. 2). Among the total number of NHI enrollees in the year 2000 ($n = 23{,}753{,}407$), the CGRD included data on 5 million patients for the outpatient sample and 2.9 million patients for the inpatient sample. Cancer, liver cirrhosis, and hepatitis C had the highest outpatient coverage rate (34%), whereas their inpatient coverage rate was 20%. These proportions accounted for 8.1 million outpatient and 4.8 million inpatient samples in the CGRD. This large sample size of the CGRD, including data on cancer, catastrophic diseases, and chronic diseases facilitates a large number of studies in these areas.

Being a large database, our results supported the CGRD to be suitable for use in public health and epidemiologic studies. In addition, the CGRD enhances statistical power when studying rare diseases or infrequent outcomes. Because the CGRD is updated annually, researchers can perform longitudinal studies from 2000, the year from which electronic medical records of CGMH became available. Staff members of CGMH who fulfill the requirements of conducting research projects are eligible to apply for the CGRD. Because data in the CGRD are not collected for a specific study, the observer-expectancy effect is minimized and the objectiveness of collected data is strengthened. Being free of cost, the concern about grant support is less pressing. According to experience of conducting studies based on the Taiwan NHIRD and UK General Practice Research Database, the low cost required to access CGRD data can considerably promote scientific research. NHIRD studies have rapidly expanded in both quantity and quality since the first study was published in 2000 [18]. Hence, the CGRD, which has numerous advantages, is expected to be a highly competitive database.

In contrast to the NHIRD, which contains secondary data from billing order and charge codes, the CGRD contains original data, and is therefore superior to the NHIRD. A critical limitation of the NHIRD is that it lacks data on examination results and health behavior; however, these are documented and comprehensively detailed in the CGRD. Second, the accuracy of diagnosis in the NHIRD is often criticized. By contrast in the CGRD, the diagnosis can be proved or supported by pathology studies and laboratory or other examinations. Third, the NHIRD does not include any data on self-pay or trial settings, which are accessible in the CGRD [19]. These advantages render the CGRD superior to the NHIRD. Cancer studies on the basis of the NHIRD are often criticized because of the lack of stage and histological classifications, which could be studied more completely in CGRD-based studies [20,21]. In hepatitis studies, more precise diagnosis of samples can be established (hepatitis B, hepatitis C, or alcoholic hepatitis with or without liver cirrhosis) according to the history, unhealthy habits, serologic test results, and abdominal sonography report available in the CGRD.

Some challenges exist for the CGRD. First, although the government center of CGMH is expanding considerable effort to protect patients' privacy, it is necessary to keep pace with time and amount of data collected and stored and continue to

protect patients' privacy. Second, with the increasing utilization of the CGRD, more human resources with more specialized work will be required. Third, data collected before 2000 were traditional paper medical records and these are not currently available for researchers. Fourth, as long as the patients have ever visited other hospital for any medical issues, their medical information in CGRD won't be completed. Therefore, in the future, the CGRD should continue to improve with time and look forward to contribute more to human health studies worldwide.

Conclusions

The CGRD is a multi-institutional, original medical record-based research database with high overall and disease-specific coverage of Taiwan. The population of the CGRD have significantly higher severity of comorbidities, and prevalence of specific diseases than those of Taiwan NHIRD and medical centers in Taiwan.

Conflicts of interest

The authors declare that they have no competing interest.

Acknowledgement

The study was financially supported by grants from the Chang Gung Memorial Hospital, Chiayi, Taiwan, R.O.C. (CGRPG6G0011). The authors would like to thank Center of Excellence for Chang Gung Research Datalink (CORPG6D0161-2, CORPG6D0251-2) for the comments and assistance in data analysis.

REFERENCES

[1] CGMH. About CGMH – service overview: Chang Gung Medical Foundation. 2015. https://www.cgmh.org.tw/cgmh/about/about_04.htm.
[2] Chou WC, Liu KH, Lu CH, Hung YS, Chen MF, Cheng YF, et al. To operate or not: prediction of 3-month postoperative mortality in geriatric cancer patients. J Cancer 2016;7:14–21.
[3] Chou WC, Wang F, Cheng YF, Chen MF, Lu CH, Wang CH, et al. A simple risk stratification model that predicts 1-year postoperative mortality rate in patients with solid-organ cancer. Cancer Med 2015;4:1687–96.
[4] Cheng CL, Kao YH, Lin SJ, Lee CH, Lai ML. Validation of the National Health Insurance Research Database with ischemic stroke cases in Taiwan. Pharmacoepidemiol Drug Saf 2011;20:236–42.
[5] Cheng CL, Lee CH, Chen PS, Li YH, Lin SJ, Yang YH. Validation of acute myocardial infarction cases in the national health insurance research database in taiwan. J Epidemiol 2014;24:500–7.
[6] NHRI. Sampling method and representativeness of Longitudinal Health Insurance Database (LHID). National Health Research Institutes; 2016. http://nhird.nhri.org.tw/date_cohort.html.
[7] Tsan YT, Lee CH, Wang JD, Chen PC. Statins and the risk of hepatocellular carcinoma in patients with hepatitis B virus infection. J Clin Oncol 2012;30:623–30.
[8] Yang YH, Chen WC, Tsan YT, Chen MJ, Shih WT, Tsai YH, et al. Statin use and the risk of cirrhosis development in patients with hepatitis C virus infection. J Hepatol 2015;63:1111–7.
[9] CDC. International Classification of Diseases, Ninth Revision (ICD-9): Centers for Disease Control and Prevention. 1978. http://www.cdc.gov/nchs/icd/icd9.htm.
[10] Chiu HF, Ho SC, Chen CC, Yang CY. Statin use and the risk of liver cancer: a population-based case-control study. Am J Gastroenterol 2011;106:894–8.
[11] Lai MN, Wang SM, Chen PC, Chen YY, Wang JD. Population-based case-control study of Chinese herbal products containing aristolochic acid and urinary tract cancer risk. J Natl Cancer Inst 2010;102:179–86.
[12] Wu CY, Kuo KN, Wu MS, Chen YJ, Wang CB, Lin JT. Early helicobacter pylori eradication decreases risk of gastric cancer in patients with peptic ulcer disease. Gastroenterology 2009;137:1641–8.
[13] Chen CY, Liu CY, Su WC, Huang SL, Lin KM. Factors associated with the diagnosis of neurodevelopmental disorders: a population-based longitudinal study. Pediatrics 2007;119:e435–43.
[14] Liu CY, Hung YT, Chuang YL, Chen YJ, Weng WS, Liu JS, et al. Incorporating development stratification of Taiwan townships into sampling design of large scale health interview survey. J Health Manag 2006;14:1–22.
[15] Richard A, Deyo DCC, Ciol Marcia A. Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. J Clin Epidemiol 1992;45:613–9.
[16] MHW. 2011 statistics of causes of death in Taiwan: Ministry of Health and Welfare of Taiwan. 2011. http://www.mohw.gov.tw/EN/Ministry/Statistic.aspx?f_list_no=474&fod_list_no=3485.
[17] Quan H, Sundararajan V, Halfon P, Fong A, Burnand B, Luthi J-C, et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. Med Care 2005;43:1130–9.
[18] Chen YC, Yeh HY, Wu JC, Haschler I, Chen TJ, Wetter T. Taiwan's National Health Insurance Research Database: administrative health care database as study object in bibliometrics. Scientometrics 2011;86:365–80.
[19] Hsu YC. Analyzing Taiwan's National Health Insurance Research Database to explicate the allocation of health-care resources. Adv Dig Med 2015;2:41–3.
[20] Chen MF, Yang YH, Lai CH, Chen PC, Chen WC. Outcome of patients with esophageal cancer: a nationwide analysis. Ann Surg Oncol 2013;20:3023–30.
[21] Wu CC, Hsu TW, Chang CM, Yu CH, Lee CC. Age-adjusted Charlson comorbidity index scores as predictor of survival in colorectal cancer patients who underwent surgical resection and chemoradiation. Med (Baltimore) 2015;94:e431.