Living Up to Online Advice: How Health Platforms Influence Physicians' Offline Practice

Online Appendix

Appendix A: Two-Stage Propensity Score Matching Approach

To address potential selection biases at both physician and patient levels, we employed a two-stage matching approach.

Stage 1: Physician-Level Matching

We adopted two propensity score matching (PSM) strategies matching the same treated group to different control groups. Among the 456 eligible physicians, a subset of 75 physicians who joined the online platform between 2008 and 2013 is designated as the treated group. In the first PSM strategy, the control group is defined as a comparable subgroup of physicians within the untreated group (those who did not join the online platform) that match the profile of the treated group. In the second PSM strategy, to further mitigate the potential bias arising from unobserved heterogeneity, we leverage the natural transition in adoption patterns observed in 2013 (as shown in Table A1). We use physicians who were initially non-adopters but subsequently joined the platform during the slower adoption phase of 2013-2015 (involving 23 physicians) as the control group. These later adopters are paired with earlier adopters—those who joined between 2008 and 2012 (totaling 69 physicians)—for look-ahead PSM (LAPSM) based on a set of observable characteristics to ensure comparability. The temporal distribution of online platform adoption by these physicians is meticulously documented in Table A1.

Table A1. Physician Entry From 2008 to 2015

Year	New Entry	Cumulative Entry
2008	4	4
2009	36	40
2010	8	48
2011	13	61
2012	8	69
2013	6	75
2014	10	85
2015	7	92

To conduct the physician-level PSM and LAPSM, we use a physician's observable inpatient-related characteristics in 2007, a year before any physician from Hospital A has joined the online platform. The matching variables include the following:

#_Patients: number of inpatient admissions to the physician in 2007.

Share_Surgical_Patient: the percentage of surgical patients among the inpatient admissions of 2007.

Share_Non_Emergency_Admissions: the percentage of admissions through non-emergency outpatient visit among the inpatient admissions of 2007.

Share_Female_Patient: the percentage of female patients among the inpatient admissions of 2007.

Share_Age_0_1: the percentage of infants (age between 0 and 1) among the inpatient admissions of 2007.

Share_Age_2_17: the percentage of children and teenagers (age between 1 and 18) among the inpatient admissions of 2007.

Share_Age_18_34: the percentage of young adults (age between 18 and 35) among the inpatient admissions of 2007.

Share_Age_35_64: the percentage of middle-age adults (age between 35 and 65) among the inpatient admissions of 2007.

Share_Age_65+: the percentage of retirees (age above 65) among the inpatient admissions of 2007.

Senior_Physician: = 1 if the physician has a "senior physician" title.

Vice_Senior_Physician: = 1 if the physician has a "vice senior physician" title.

Professor: = 1 if the physician also has an academic title.

One-to-one without replacement matching was performed using these matching variables. Implementing the first PSM strategy, we present the t-tests on the differences in the matching variables between the treated group and the untreated group both before and after PSM matching in Table A2

Similarly, the results of the matching t-tests based on the LA-PSM samples are presented in Table A3. The insignificant difference across all variables after matching indicates that the matched groups have no systematic difference with the treated group before they joined the online platform.

Table A2. Physician-level Propensity Score Matching Results

Matching Variables	Treated Group	Before PSM		After PSM	
		Untreated Group	Difference	Matched Group	Difference
#_Patients	135.547	101.514	-34.032*	132.507	-3.040
Share_Surgical_Patient	0.545	0.396	-0.148***	0.506	-0.039
Share_Non_Emergency_Admission	0.778	0.731	-0.047	0.791	0.013
Share_Female_Patient	0.444	0.430	-0.015	0.430	-0.014
Share_Age_0_1	0.005	0.016	0.011	0.001	-0.004
Share_Age_2_17	0.054	0.067	0.014	0.048	-0.005
Share_Age_18_34	0.210	0.171	-0.039	0.223	0.014
Share_Age_35_64	0.436	0.398	-0.038	0.426	-0.010
Share_Age_65+	0.296	0.348	0.052	0.301	0.006
Senior_Physician	0.867	0.551	-0.315***	0.880	0.013
Vice_Senior_Physician	0.133	0.420	0.287***	0.120	-0.013
Professor	0.453	0.202	-0.251***	0.427	-0.027
#_Physicians	75	381		75	

Note: *** p<0.01, ** p<0.05, * p<0.1

Table A3. Look-ahead Propensity Score Matching Results

Matching Variables	Control Group	Before LAPSM		After LAPSM	
		Treated Group	Difference	Matched Group	Difference
#_Patients	138.090	168.170	-30.080	166.640	-28.550
Share_Surgical_Patient	0.705	0.668	0.037	0.714	-0.009
Share_Non_Emergency_Admission	0.760	0.757	0.003	0.759	0.001
Share_Female_Patient	0.469	0.452	0.018	0.459	0.010
Share_Age_0_1	0.001	0.014	-0.012	0.008	-0.006
Share_Age_2_17	0.056	0.065	-0.009	0.056	0.000
Share_Age_18_34	0.250	0.225	0.026	0.270	-0.019
Share_Age_35_64	0.463	0.442	0.021	0.444	0.019
Share_Age_65+	0.229	0.258	-0.029	0.219	0.010
Senior_Physician	0.636	0.829	-0.193*	0.727	-0.091
Vice_Senior_Physician	0.364	0.171	0.193	0.273	0.091
Professor	0.273	0.366	-0.093	0.318	-0.045
#_Physicians	23	69		22	

Note: *** p<0.01, ** p<0.05, * p<0.1

Stage 2: Patient-Level Matching

Within our matched physician sample, we implement another PSM at the patient level to control for patient self-selection bias. Specifically, for each inpatient admission to physicians who joined the online platform, we identify one nearest neighbor from admissions to non-participating physicians within the same month, matching on patient demographic and socioeconomic features. This temporal matching approach helps control for seasonal variations and time trends while ensuring comparable patient populations between participating and non-participating physicians. The matching variables include the following patient characteristics:

Demographic Variables:

- *Gender*: Binary indicator (1 = female, 0 = male)
- *Age*: Patient's age in years at admission
- *Marital_Status*: Binary indicator (1 = married, 0 = otherwise)
- *Ethnic_Minority*: Binary indicator (1 = ethnic minority, 0 = Han ethnicity)

Socioeconomic Indicators:

- *High_Education*: Binary indicator (1 = employed in knowledge-intensive sectors including healthcare, education, or public administration, 0 = otherwise)
- *Rural_Residence*: Binary indicator (1 = residing in a rural area, 0 = urban area)

Insurance Coverage:

- *Insurance_Type*: Categorical variable indicating insurance coverage:
    - *No Coverage*: No insurance program
    - *Full_Coverage*: Full coverage program for civil servants
    - *Employee_Copay*: Urban employee insurance with co-payment
    - *Retiree_Copay*: Urban retiree insurance with co-payment

One-to-one without replacement matching was performed using these matching variables within each month, ensuring temporal comparability of matched pairs. This monthly matching approach helps control for seasonal variations and time trends in hospital admissions. Tables A4 and A5 present the balance diagnostics before and after matching for both the Physician-PSM and Physician-LAPSM samples, respectively. The successful balance achieved across both samples suggests our matching procedure effectively controlled for observable differences in patient characteristics.

Table A4. Patient-Level Propensity Score Matching Results Based on Physician-PSM Sample

Matching Variables	Before PSM			After PSM		
	Control Group	Treated Group	T-test	Control Group	Treated Group	T-test
Gender	0.492	0.514	-7.772***	0.502	0.504	-0.713
Age	49.909	48.014	17.477***	49.917	50.012	-0.881
Marital_Status	0.832	0.812	6.509***	0.834	0.839	-1.663*
Ethnic_Minority	0.007	0.006	5.226***	0.007	0.007	0.858
High_Education	0.026	0.016	12.543***	0.021	0.019	1.271
Rural_Residence	0.024	0.022	1.961**	0.024	0.023	0.492
No Coverage	0.276	0.322	-18.382***	0.279	0.283	-1.602
Full_Coverage	0.008	0.008	1.073	0.008	0.008	0.429
Employee_Copay	0.141	0.139	1.083	0.141	0.141	0.263
Retiree_Copay	0.496	0.466	11.087***	0.495	0.491	1.429
#_Admissions	64,452	70,364		62,759	64,410	

Note: *** p<0.01, ** p<0.05, * p<0.1

Table A5. Patient-Level Propensity Score Matching Results Based on Physician-LAPSM Sample

Matching Variables	Before PSM			After PSM		
	Control Group	Treated Group	T-test	Control Group	Treated Group	T-test
Gender	0.565	0.419	28.770***	0.421	0.418	0.536
Age	43.460	45.189	-8.728***	45.031	45.283	-1.147
Marital_Status	0.773	0.741	5.090***	0.750	0.745	0.729
Ethnic_Minority	0.005	0.007	-3.078***	0.007	0.007	-0.782
High_Education	0.013	0.024	-8.156***	0.021	0.024	-1.923*
Rural_Residence	0.023	0.023	-0.005	0.023	0.023	0.253
No Coverage	0.371	0.338	6.769***	0.347	0.339	1.498
Full_Coverage	0.002	0.004	-4.006***	0.003	0.004	-1.507
Employee_Copay	0.165	0.163	0.505	0.161	0.162	-0.443
Retiree_Copay	0.401	0.378	4.756***	0.398	0.379	1.626
#_Admissions	25,544	15,184		16,184	15,102	

Note: *** p<0.01, ** p<0.05, * p<0.1

Appendix B: Supplementary Tables

Table B1. First Stage Estimation Results for 2SLS

VARIABLES	(1)	(2)	(3)	(4)
	After_QA (Y= Dev_LOS)	*Ln_QA (Y= Dev_LOS)*	*After_QA (Y= Med_adj; UncoveredMedCost_Ratio)*	*Ln_QA (Y= Med_adj; UncoveredMedCost_Ratio)*
IV_ConsultGZ	0.003***	0.014***	0.003***	0.014***
	(0.001)	(0.004)	(0.001)	(0.004)
IV_RespGZ	-0.003***	-0.010**	-0.003***	-0.010**
	(0.001)	(0.004)	(0.001)	(0.004)
Num_LM_Discharge	0.000***	-0.001***	0.000***	-0.001***
	(0.000)	(0.000)	(0.000)	(0.000)

Ln_LM_Other_X[1]	0.003***	-0.046***	-0.001***	0.007***
	(0.001)	(0.005)	(0.000)	(0.002)
Post_Num_Patient	-0.000	-0.001***	-0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Ln_Post_Reply	-0.099***	0.793***	-0.099***	0.794***
	(0.001)	(0.007)	(0.001)	(0.007)
Ln_Post_Length	0.049***	-0.098***	0.049***	-0.098***
	(0.000)	(0.003)	(0.000)	(0.003)
Discharge_Pressure	0.000***	0.000	0.000***	-0.001**
	(0.000)	(0.000)	(0.000)	(0.000)
Ln_Tenure	0.154***	0.593***	0.154***	0.592***
	(0.000)	(0.001)	(0.000)	(0.001)
Female	0.000	0.001	0.000	0.001
	(0.001)	(0.003)	(0.001)	(0.003)
Age	-0.000***	-0.000	-0.000***	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Surgical	0.002**	-0.070***	0.002**	-0.068***
	(0.001)	(0.005)	(0.001)	(0.005)
Serious	-0.004**	-0.052***	-0.004**	-0.043***
	(0.002)	(0.012)	(0.002)	(0.012)
LOS			0.000	-0.002***
			(0.000)	(0.000)
Observations	127,166	127,166	127,166	127,166
R-squared	0.968	0.947	0.968	0.947

*** p<0.01, ** p<0.05, * p<0.1

Table B2. Supplementary Identification Tests II

	(1)	(2)	(3)	(4)	(5)	(6)
Dep. Var.	***Med_adj***	***UncoveredMed Cost_Ratio***	***Dev_LOS***	***Med_adj***	***UncoveredMed Cost_Ratio***	***Dev_LOS***
After_QA	0.171	0.015	0.087	0.171	0.015	0.087
	(0.145)	(0.015)	(0.050)	(0.145)	(0.015)	(0.050)
Physician_Treatment_Sim	-0.374***	-0.029***	-0.405***	-0.085***	-0.015***	-0.302***
	(0.084)	(0.011)	(0.012)	(0.036)	(0.007)	(0.032)
Other_Treatment_Sim				-0.013***	-0.007**	-0.017***
				(0.005)	(0.004)	(0.004)
Observations	127,166	127,166	127,166	127,166	127,166	127,166
# of Physicians	150	150	150	150	150	150
Adjusted R2	0.372	0.298	0.236	0.381	0.303	0.238
Within A-R2	0.109	0.023	0.115	0.113	0.027	0.116

The full set of control variables are included. *** p<0.01, ** p<0.05, * p<0.1

Table B3a. Patient Cost Exposure I-a

	(1)	(2)	(3)	(4)	(5)	(6)
Dep. Var.	***Med_adj***	***UncoveredMedCost_Ratio***	***Dev_LOS***	***Med_adj***	***UncoveredMedCost_Ratio***	***Dev_LOS***
After_QA	0.041	-0.004	0.015	-0.061	0.177	-0.529
	(0.218)	(0.004)	(0.058)	(0.900)	(0.141)	(0.324)
Ln_QA	-0.154***	-0.003***	-0.049***	**0.059**	**0.032**	**-0.020**
	(0.035)	(0.001)	(0.012)	(0.135)	(0.021)	(0.074)

[1] *Ln_LM_Other_X* denotes *Ln_LM_Other_LOS* if the DV is about medication cost, otherwise *Ln_LM_Other_Med* if the DV is *Dev_LOS*.

Insurance	*None*	*None*	*None*	*Full coverage*	*Full coverage*	*Full coverage*
Observations	37,018	37,018	37,018	857	857	857
# of Physicians	149	149	149	76	76	76
Adjusted R2	0.350	0.188	0.189	0.586	0.354	0.395
Within A-R2	0.140	0.057	0.065	0.108	0.113	0.266

The full set of control variables are included. *** p<0.01, ** p<0.05, * p<0.1

Table B3b. Patient Cost Exposure I-b

	(1)	(2)	(3)	(4)	(5)	(6)
Dep. Var.	***Med_adj***	***UncoveredMedCost_Ratio***	***Dev_LOS***	***Med_adj***	***UncoveredMedCost_Ratio***	***Dev_LOS***
After_QA	0.146	0.021	0.011	-0.029	0.012	0.067
	(0.216)	(0.023)	(0.090)	(0.133)	(0.023)	(0.041)
Ln_QA	-0.090**	-0.005***	-0.067***	-0.072***	-0.010**	-0.013***
	(0.037)	(0.002)	(0.020)	(0.024)	(0.005)	(0.005)
Insurance	*Employee Co-payment*	*Employee Co-payment*	*Employee Co-payment*	*Retiree Co-payment*	*Retiree Co-payment*	*Retiree Co-payment*
Observations	17,919	17,919	17,919	62,555	62,555	62,555
# of Physicians	148	148	148	150	149	150
Adjusted R2	0.401	0.311	0.236	0.442	0.187	0.267
Within A-R2	0.134	0.0229	0.140	0.128	0.0447	0.135

The full set of control variables are included. *** p<0.01, ** p<0.05, * p<0.1

Notes: This table presents boundary condition tests across patient insurance types to test whether cognitive consistency pressure varies with patient cost exposure. Columns (1)-(3) show results for uninsured patients paying full out-of-pocket costs. Columns (4)-(6) show results for fully insured patients (government employee insurance) with minimal/no direct costs. Table B4b shows results for employee co-payment (columns 1-3) and retiree co-payment insurance (columns 4-6). The full set of control variables are included. *** p<0.01, ** p<0.05, * p<0.1

Table B4 reports the robustness under several alternative specifications of our dependent variables, *Dev_LOS* and *Med_adj*. The dependent variable, *UncoveredMedCost_Ratio*, is excluded from the robustness check because it is used in its original form without a logarithmic transformation. We check whether the results stay the same (1) if normal value instead of logarithm value is used, (2) if the minimum value of the length of stay and the adjusted medication charge is adjusted to a higher value (i.e., 1/10 of the sample mean), and (3) if the maximum value of the length of stay and the adjusted medication charge is constrained to a lower value (i.e., sample mean + 3 * sample standard deviation). As reported in Table B4, all alternative specifications do not change the qualitative interpretation of our main results. Therefore, we are confident that our results are not driven by misspecification of our outcome variables or the extreme values.

Table B4. Robustness – Alternative Measures

	(1)	(2)	(3)	(4)	(5)	(6)
Dep. Var.	*Dev_LoS*	*Dev_LoS*	*Dev_LoS*	*Med_adj*	*Med_adj*	*Med_adj*
After_QA	0.278	0.062	0.063	-363.742	0.081	0.043
	(0.227)	(0.032)	(0.033)	(439.682)	(0.130)	(0.146)
Ln_QA	-0.343***	-0.018**	-0.021***	-266.548**	-0.081***	-0.102***
	(0.074)	(0.008)	(0.008)	(116.379)	(0.024)	(0.027)
Measurement Specification	N	Log + Min	Log + Max	N	Log + Min	Log + Max
Observations	127,166	127,166	127,166	127,166	127,166	127,166
# Physicians	150	150	150	150	150	150
Adjusted R^2	0.801	0.315	0.270	0.367	0.437	0.373
Within A-R^2	0.794	0.204	0.150	0.191	0.100	0.109

Cluster-robust standard errors in parentheses. Physician, year, month, day-of-week, DRG, department-month, and insurance group fixed effects included. Coefficients on control variables not reported. Alternative measurement specification for the dependent variable: nominal value (**N**); log-transformed value (**Log**); minimum value adjusted to 1/10 of sample mean (**Min**); maximum value adjusted to sample mean + 3*sample standard deviation (**Max**). *** p<0.01, ** p<0.05, * p<0.1

In Table B5, we consider robustness under alternative samples. Six alternative samples instead of the default one-to-one matched sample are compared: (1) one-to-three (with replacement) matched sample, (2) full sample, (3) online-only physician sample, (4) full sample (including late-entry physicians), (5) sample with the most popular department removed, (6) sample with non-recovery (i.e., death or transfer to other hospitals) discharges removed, and (7) using alternative matching method (Mahalanobis Distance Matching). Again, as shown in Table B5, our main results do not respond to such changes, which lends us further confidence that our results are not driven by sample construction.

Table B5. Robustness – Alternative Samples

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
				Panel A. Length of Stay			
Dep. Var.	*Dev_LoS*	*Dev_LoS*	*Dev_LoS*	*Dev_LoS*	*Dev_LoS*	*Dev_LoS*	*Dev_LoS*
After_QA	0.038	0.047	0.039	0.047	0.097**	0.048	0.051
	(0.039)	(0.038)	(0.040)	(0.038)	(0.043)	(0.038)	(0.038)
Ln_QA	-0.030***	-0.032***	-0.017***	-0.032***	-0.020***	-0.032***	-0.023**
	(0.009)	(0.010)	(0.008)	(0.010)	(0.009)	(0.010)	(0.010)
Adjusted R^2	0.254	0.237	0.256	0.237	0.258	0.232	0.246
Within A-R^2	0.143	0.115	0.142	0.115	0.129	0.112	0.134
				Panel B. Medication Charge			
Dep. Var.	*Med_adj*	*Med_adj*	*Med_adj*	*Med_adj*	*Med_adj*	*Med_adj*	*Med_adj*
After_QA	-0.065	0.043	0.039	0.043	0.093	0.037	0.024
	(0.159)	(0.146)	(0.150)	(0.146)	(0.141)	(0.146)	(0.149)
Ln_QA	-0.084***	-0.102***	-0.085***	-0.102***	-0.076***	-0.100***	-0.089***
	(0.028)	(0.027)	(0.027)	(0.027)	(0.025)	(0.026)	(0.028)

Adjusted R^2	0.366	0.373	0.416	0.373	0.429	0.369	0.422
Within A-R^2	0.126	0.109	0.137	0.109	0.123	0.106	0.125
Panel C. Uncovered Medication Cost Ratio (UMCR)							
Dep. Var.	*UMCR*	*UMCR*	*UMCR*	*UMCR*	*UMCR*	*UMCR*	*UMCR*
After_QA	0.003	0.008	0.018	0.008	0.025	0.007	0.007
	(0.015)	(0.015)	(0.014)	(0.015)	(0.017)	(0.015)	(0.015)
Ln_QA	-0.005***	-0.006**	-0.007**	-0.006**	-0.003***	-0.006**	-0.006**
	(0.002)	(0.003)	(0.003)	(0.003)	(0.001)	(0.003)	(0.003)
Adjusted R^2	0.300	0.297	0.300	0.297	0.285	0.296	0.304
Within A-R^2	0.0208	0.0238	0.0260	0.0238	0.0289	0.0234	0.0264
Observations	114,664	127,166	64,408	127,166	87,458	126,302	82,850
# of Physicians	144	150	75	150	118	150	90

Cluster-robust standard errors in parentheses. Physician, year, month, day-of-week, DRG, department-month, and insurance group fixed effects included. Coefficient on control variables not reported. One-to-three with replacement matching in Column (1); Full sample in Column (2); Online physician only sample in Column (3); Original sample including late entry physicians in Column (4); The most popular department removed in Column (5); Non-recovery (i.e., death or transfer to other hospitals) cases removed in Column (6). Mahalanobis Distance Matching result in Column (7)
*** p<0.01, ** p<0.05, * p<0.1

In Table B6, we perform falsification tests to refute the possibility that online physician information provision systematically alters the characteristics of offline patients, thereby affecting offline care delivery. This analysis is conducted using a linear probability model, as specified in Equation (B1).

$$Y_{ij} = \alpha_i + \beta_1 After\_QA_{ij} + \beta_2 Ln\_QA_{ij} + \boldsymbol{Z} + \varepsilon_{ij}, \tag{B1}$$

where Y_{ij} represents one of the patient characteristics (i.e., $Female_{ij}$, Age_{ij}, $Surgical_{ij}$, and $Serious_{ij}$) in a binary form, and no additional control variable are included. As shown in Table B6, none of the coefficients are significant. It implies that observable patient characteristics do not change with the online physician information provision. As such, the reduction of medication charge, uncovered medication cost ratio, and the converge of LOS towards GMLOS are not driven by the selection of patients with less severe complications or higher baseline knowledge level.

Table B6. Falsification Test – Patient Characteristics

	(1)	(2)	(3)	(4)
Dep. Var.	*Gender*	*Age*	*Surgical*	*Serious*
After_QA	0.001	-2.257	-0.006	-0.013
	(0.025)	(1.943)	(0.029)	(0.008)
Ln_QA	0.001	-0.075	-0.025	-0.004
	(0.004)	(0.198)	(0.017)	(0.003)
Observations	127,166	127,166	127,166	127,166

# of Physicians	150	150	150	150
Adjusted R2	0.194	0.464	0.577	0.0614
Within R2	9.99e-05	0.00146	0.00716	0.00398

Clustered-robust standard errors in parentheses. Physician, year, month, DRG, department-month, and insurance group fixed effects included. *** p<0.01, ** p<0.05, * p<0.1

To account for this recency effect, we implement two time-discounting strategies to adjust Ln_QA_{ij}. The first strategy applies a monthly discount rate of 95%,[2] reflecting that physicians' more recent public commitments are more psychologically salient than older ones in creating consistency pressure. The second strategy introduces a decremental discount for every set of ten Q&A interactions arranged in reverse chronological order, capturing how recent intensive engagement might heighten the salience of public position-taking.

Table B7. Robustness – Time Discounted Information Provision

	(1)	(2)	(3)	(4)	(5)	(6)
Dep. Var.	***Med_adj***	***Uncovered MedCost_Ratio***	***Dev_LOS***	***Med_adj***	***Uncovered MedCost_Ratio***	***Dev_LOS***
After_QA	0.133	0.013	0.076**	0.090	0.011	0.064
	(0.139)	(0.015)	(0.037)	(0.145)	(0.015)	(0.038)
Ln_discount_QA	-0.095***	-0.005***	-0.028***	-0.099***	-0.005***	-0.028***
	(0.024)	(0.002)	(0.009)	(0.026)	(0.002)	(0.010)
Discount type	*Discount by month*	*Discount by month*	*Discount by month*	*Discount in tens*	*Discount in tens*	*Discount in tens*
Observations	127,166	127,166	127,166	127,166	127,166	127,166
# of Physicians	150	150	150	150	150	150
Adjusted R2	0.373	0.297	0.237	0.373	0.297	0.237
Within A-R2	0.109	0.024	0.115	0.109	0.024	0.115

The full set of control variables are included. *** p<0.01, ** p<0.05, * p<0.1

Table B8a. 2SLS Estimation Results for Identification Tests I-a

	(1)	(2)	(3)	(4)	(5)	(6)
Dep. Var.	***Med_adj***	***UncoveredMedCost_Ratio***	***Dev_LOS***	***Med_adj***	***UncoveredMedCost_Ratio***	***Dev_LOS***
After_QA	0.933	0.368	0.528	0.627	0.606	0.787
	(0.587)	(0.505)	(0.336)	(0.524)	(0.646)	(0.497)
Ln_QA	-0.424**	-0.233***	-0.985***	-0.248***	-0.172***	-0.318***
	(0.163)	(0.091)	(0.472)	(0.066)	(0.079)	(0.145)
TP_ratio	-0.382***	-0.355***	-0.411***			
	(0.092)	(0.143)	(0.196)			
	-0.451***	-0.161***	-0.859***			

[2] We have explored various discounting rates, including 99%, and found that they yield comparable estimation outcomes.

Ln_QA × *TP_ratio*	(0.171)	(0.035)	(0.138)			
EBS_ratio				-0.310	-0.586	-0.637***
				(0.786)	(0.623)	(0.239)
Ln_QA ×				-0.201	-0.028	-0.907***
EBS_ratio				(0.203)	(0.224)	(0.409)
Observations	127,166	127,166	127,166	127,166	127,166	127,166
# of Physicians	150	150	150	150	150	150
Adjusted R2	0.525	0.449	0.388	0.485	0.308	0.454

The full set of control variables are included. *** p<0.01, ** p<0.05, * p<0.1

Table B8b. 2SLS Estimation Results for Identification Tests I-b

	(1)	(2)	(3)	(4)	(5)	(6)
Dep. Var.	***Med_adj***	***UncoveredMedCost_Ratio***	***Dev_LOS***	***Med_adj***	***UncoveredMedCost_Ratio***	***Dev_LOS***
After_QA	-0.868***	0.359	0.033	0.512	0.904*	0.026
	(0.740)	(0.355)	(0.044)	(0.504)	(0.475)	(0.036)
Ln_QA	-0.245***	-0.161**	-0.315***	-0.145***	-0.168**	-0.144***
	(0.048)	(0.027)	(0.119)	(0.019)	(0.072)	(0.029)
RS_ratio	-0.599***	-0.582***	-0.712*			
	(0.145)	(0.245)	(0.360)			
Ln_QA ×	-0.326**	-0.690***	-0.550			
RS_ratio	(0.124)	(0.108)	(0.515)			
Explicit				-0.775***	0.732***	-0.184***
				(0.082)	(0.068)	(0.086)
Ln_QA ×				-0.669***	-0.810**	-0.132***
Explicit				(0.115)	(0.347)	(0.047)
Observations	127,166	127,166	127,166	127,166	127,166	127,166
# of Physicians	150	150	150	150	150	150
Adjusted R2	0.538	0.358	0.244	0.568	0.484	0.370

The full set of control variables are included. *** p<0.01, ** p<0.05, * p<0.1

Table B9. 2SLS Estimation Results for Identification Tests II

	(1)	(2)	(3)	(4)	(5)	(6)
Dep. Var.	***Med_adj***	***UncoveredMedCost_Ratio***	***Dev_LOS***	***Med_adj***	***UncoveredMedCost_Ratio***	***Dev_LOS***
After_QA	-0.379	-0.138	0.805	0.675	0.335	0.784
	(0.946)	(0.515)	(0.536)	(0.678)	(0.427)	(0.471)
Ln_DiseaseSpecific_QA	-0.525***	-0.350**	-0.778**			
	(0.096)	(0.141)	(0.161)			
Ln_No_DiseaseSpecific_QA				0.624	0.653	0.375
				(0.634)	(0.472)	(0.319)
Observations	127,166	127,166	127,166	127,166	127,166	127,166
# of Physicians	150	150	150	150	150	150
Adjusted R2	0.329	0.475	0.459	0.439	0.499	0.465

The full set of control variables are included. *** p<0.01, ** p<0.05, * p<0.1

Table B10a. 2SLS Estimation Results for Patient Cost Exposure I-a

	(1)	(2)	(3)	(4)	(5)	(6)
Dep. Var.	***Med_adj***	***UncoveredMedCost_Ratio***	***Dev_LOS***	***Med_adj***	***UncoveredMedCost_Ratio***	***Dev_LOS***

After_QA	0.161	0.075	0.290	0.418	0.659	-0.038
	(0.333)	(0.146)	(0.970)	(0.545)	(0.746)	(0.086)
Ln_QA	-0.369***	-0.533***	-0.478***	**0.664**	**0.016**	**0.603**
	(0.089)	(0.261)	(0.237)	(0.569)	(0.048)	(0.427)
Insurance	*None*	*None*	*None*	*Full coverage*	*Full coverage*	*Full coverage*
Observations	37,018	37,018	37,018	857	857	857
# of Physicians	149	149	149	76	76	76
Adjusted R2	0.429	0.407	0.489	0.269	0.203	0.307

The full set of control variables are included. *** p<0.01, ** p<0.05, * p<0.1

Table B10b. 2SLS Estimation Results for Patient Cost Exposure I-b

	(1)	(2)	(3)	(4)	(5)	(6)
Dep. Var.	***Med_adj***	***UncoveredMedCost_Ratio***	***Dev_LOS***	***Med_adj***	***UncoveredMedCost_Ratio***	***Dev_LOS***
After_QA	0.220	0.171	0.832	0.146	0.021	0.011
	(0.307)	(0.299)	(0.647)	(0.216)	(0.023)	(0.090)
Ln_QA	-0.294***	-0.125***	-0.210***	-0.213***	-0.161***	-0.307***
	(0.137)	(0.057)	(0.042)	(0.078)	(0.045)	(0.050)
Insurance	*Employee Co-payment*	*Employee Co-payment*	*Employee Co-payment*	*Retiree Co-payment*	*Retiree Co-payment*	*Retiree Co-payment*
Observations	17,919	17,919	17,919	62,555	62,555	62,555
# of Physicians	148	148	148	150	149	150
Adjusted R2	0.334	0.305	0.223	0.462	0.316	0.285

The full set of control variables are included. *** p<0.01, ** p<0.05, * p<0.1

Table B11. 2SLS Estimation Results for Patient Visibility and Audience Convergence

	(1)	(2)	(3)	(4)	(5)	(6)
Dep. Var.	***Med_adj***	***UncoveredMedCost_Ratio***	***Dev_LOS***	***Med***	***UncoveredMedCost_Ratio***	***Dev_LOS***
After_QA	0.595	0.438	0.780	0.995	0.491	6.426***
	(0.338)	(0.466)	(0.725)	(0.654)	(0.416)	(1.623)
Ln_QA	-0.813***	-0.153**	-0.113***	-0.465***	-0.137***	-0.597***
	(0.372)	(0.076)	(0.024)	(0.171)	(0.063)	(0.204)
Web_Access	0.640***	0.118***	0.067***			
	(0.202)	(0.041)	(0.015)			
Ln_QA × *Web_Access*	-0.307***	-0.053***	-0.013***			
	(0.109)	(0.018)	(0.006)			
Conversion_Rate				6.853***	1.847***	0.853
				(2.546)	(0.728)	(0.546)
Ln_QA × *Conversion_Rate*				-4.597***	-2.049***	0.305
				(2.104)	(1.188)	(0.315)
Observations	127,166	127,166	127,166	127,166	127,166	127,166
# of Physicians	150	150	150	150	150	150
Adjusted R2	0.461	0.487	0.490	0.446	0.685	0.480

The full set of control variables are included. *** p<0.01, ** p<0.05, * p<0.1

Table B12. 2SLS Estimation Results for Disease Chronicity Heterogeneity

	(1)	(2)	(3)	(4)	(5)	(6)
Dep. Var.	***Med_adj***	***UncoveredMe***	***Dev_LOS***	***Med_adj***	***UncoveredMe***	***Dev_LOS***

	dCost_Ratio			*dCost_Ratio*		
After_QA	6.444	2.510	4.324	2.634	1.928	1.708
	(8.560)	(3.077)	(4.076)	(2.018)	(0.941)	(2.492)
Ln_QA	-3.136	-1.550	2.223	-1.143***	-0.374***	-1.262**
	(3.915)	(1.467)	(2.046)	(0.516)	(0.114)	(0.568)
Chronic disease Y/N?	Yes	Yes	Yes	No	No	No
Observations	22,510	22,510	22,510	104,500	104,500	104,500
# of Physicians	139	139	139	149	149	149
Adjusted R2	0.683	0.579	0.767	0.649	0.540	0.737

The full set of control variables are included. *** p<0.01, ** p<0.05, * p<0.1

Table B13. 2SLS Estimation Results for Robustness Check I

	(1)	(2)	(3)
Dep. Var.	***Med_adj***	***UncoveredMedCost_Ratio***	***Dev_LOS***
After_QA	0.736	0.418	0.257
	(0.810)	(0.396)	(0.537)
Ln_QA	-0.187***	-0.117***	-0.058***
	(0.068)	(0.054)	(0.025)
Ln_Other_QA	-0.087***	-0.003***	-0.010***
	(0.013)	(0.001)	(0.002)
Observations	127,166	127,166	127,166
# of Physicians	150	150	150
Adjusted R2	0.384	0.370	0.429

The full set of control variables are included. *** p<0.01, ** p<0.05, * p<0.1

Table B14. 2SLS Estimation Results for Robustness Check II

	(1)	(2)	(3)	(4)	(5)	(6)
Dep. Var.	***Med_adj***	***UncoveredMedCost_Ratio***	***Dev_LOS***	***Med_adj***	***UncoveredMedCost_Ratio***	***Dev_LOS***
After_QA	1.884	-0.348	-1.039	0.136	0.984	0.994
	(1.761)	(0.497)	(1.050)	(0.075)	(0.491)	(0.812)
Ln_QA	-0.445***	-0.219***	-0.924***	-0.429***	-0.164***	-0.034***
	(0.194)	(0.084)	(0.290)	(0.177)	(0.065)	(0.309)
Rcom_Rate	0.265***	0.694***	0.749***			
	(0.108)	(0.274)	(0.282)			
Ln_QA × *Rcom_Rate*	-1.341***	-0.691**	-1.556***			
	(0.577)	(0.277)	(0.584)			
Market_Share				0.211	0.892	0.756
				(0.191)	(0.646)	(0.496)
Ln_QA × *Market_Share*				0.619***	0.757***	0.168***
				(0.237)	(0.363)	(0.065)
Observations	127,166	127,166	127,166	127,166	127,166	127,166
# of Physicians	150	150	150	150	150	150
Adjusted R2	0.361	0.310	0.446	0.357	0.296	0.482

The full set of control variables are included. *** p<0.01, ** p<0.05, * p<0.1

Appendix C: Construction of DRG_Severity_Score Variable

This appendix provides a comprehensive description of how we construct *DRG_severity_score*, a standardized, patient-level measure of disease severity and resource intensity based on the Chinese Hospital Diagnosis-Related Group (CHS-DRG) Version 2.0 classification and its official relative weights.

C.1. Conceptual Framework of DRG System

The Diagnosis-Related Groups (DRG) system is a patient classification system that groups hospital cases into clinically coherent categories with similar resource consumption patterns. Originally developed in the United States in the late 1960s, DRG has become a widely adopted tool for healthcare payment systems, resource allocation, and quality management globally.

The core principle of DRG classification rests on two fundamental criteria:

- **Clinical similarity**: Patients within a DRG should have similar clinical conditions and treatment pathways
- **Resource homogeneity**: Patients within a DRG should consume similar amounts of healthcare resources

In China's context, the National Healthcare Security Administration (NHSA) developed the Chinese Hospital DRG (CHS-DRG) system, with the most recent version being DRG Version 2.0 (released in July 2024 by the Beijing Municipal Medical Insurance Bureau's DRG Technical Guidance Group). This version contains 26 Major Diagnostic Categories (MDCs), 409 Adjacent Diagnosis Related Groups (ADRGs), and 760 DRGs, that can be further subdivided based on complication and comorbidity levels.

C.2. Data Infrastructure and Coding Systems

C.2.1 Medical Insurance Fund Settlement List

Our DRG grouping relies on data from the Medical Insurance Fund Settlement List (医疗保障基金结算清单), which is the standardized data collection format mandated by China's National Healthcare Security Administration. This settlement list supersedes the traditional hospital discharge summary and integrates:

- Clinical information from hospital medical records
- Insurance enrollment and eligibility data
- Detailed billing and cost information
- Healthcare service utilization data

The settlement list was formally implemented nationwide following the issuance of the "Medical Insurance Fund Settlement List Filling Specifications" (Document No. 34, 2021) by the National Healthcare Security Administration.

C.2.2 Disease Classification Coding (ICD-10)

Disease diagnoses are coded using the National Healthcare Security Administration's version of ICD-10 (International Classification of Diseases, 10th Revision): "Medical Insurance Disease Classification and Coding" (医疗保障疾病分类与代码).

Key features of this coding system:

- Based on GB/T 14396-2016 (Chinese National Standard for ICD-10)
- Version 2.0 updated in January 2021
- Contains 2,048 disease categories, 10,171 subcategories, and 33,304 specific diagnosis codes
- Designed for high inclusiveness and compatibility with existing regional coding systems
- Standardized for use across all medical insurance settlement systems in China

The coding structure follows a hierarchical format:

- Chapter level (e.g., Chapter I: Certain infectious and parasitic diseases)

- Block level (groups of related conditions)
- Category level (3-character codes, e.g., I21 for acute myocardial infarction)
- Subcategory level (4-character codes with decimal point)
- Detailed codes (5-6 character codes for specific clinical presentations)

C.2.3 Procedure Classification Coding (ICD-9-CM-3)

Surgical and medical procedures are coded using the National Healthcare Security Administration's version of ICD-9-CM-3 (International Classification of Diseases, 9th Revision, Clinical Modification, Volume 3): "Medical Insurance Surgical and Procedure Classification and Coding" (医疗保障手术及操作分类与代码).

Key features:

- Based on the International ICD-9-CM-3 standard
- Version 2.0 updated in January 2021
- Contains 890 procedure subcategories, 3,666 detailed categories, and 13,686 specific procedure codes
- Covers surgical procedures and non-surgical operations
- Maintains compatibility across different regional coding practices

The coding structure includes:

- Major procedure categories (2-digit codes, 00-99)
- Procedure subcategories (3-digit codes)
- Detailed procedure specifications (4-digit codes with decimal point)

C.3. DRG Grouping Methodology

The grouping process follows a hierarchical classification logic, proceeding through three levels:

C.3.1 Level 1: Major Diagnostic Categories (MDC)

At the highest level, cases are classified into Major Diagnostic Categories based primarily on anatomical organ systems or etiology. The CHS-DRG system includes 26 MDCs organized by body system:

- MDC A: Organ transplantation and special procedures
- MDC B: Diseases and disorders of the nervous system
- MDC C: Diseases and disorders of the eye
- MDC D: Diseases and disorders of the ear, nose, mouth, and throat
- MDC E: Diseases and disorders of the respiratory system
- MDC F: Diseases and disorders of the circulatory system
- MDC G: Diseases and disorders of the digestive system
- MDC H: Diseases and disorders of the hepatobiliary system and pancreas
- And 18 additional MDCs covering all body systems and special conditions

Assignment to MDC is primarily determined by the principal diagnosis code, with certain procedures (e.g., organ transplantation) taking precedence regardless of diagnosis.

C.3.2 Level 2: Adjacent Diagnosis Related Groups (ADRG)

Within each MDC, cases are further classified into Adjacent Diagnosis Related Groups. The DRG 2.0 scheme contains 409 ADRGs total, distributed as follows:

- 182 surgical ADRGs (procedures performed in operating room)
- 37 non-operating room procedure ADRGs
- 190 medical ADRGs (no significant procedures)

ADRG classification logic:

- **Step 1:** Identify if a surgical procedure was performed
- **Step 2:** Match principal diagnosis and procedures to ADRG definition
- **Step 3:** Apply special rules for specific conditions

C.3.3 Level 3: Diagnosis Related Groups (DRG) - Final Assignment

Each ADRG is subdivided into multiple DRGs based on complication/comorbidity (CC) levels and other factors that affect resource consumption.

Classification factors:

a) Complication and Comorbidity (CC) Level:

- Level 1 (MCC): Major complications or comorbidities that significantly increase resource use
- Level 2 (CC): Moderate complications or comorbidities with some resource impact
- Level 3 (No CC): No significant complications or comorbidities

b) Age groups (for certain conditions):

- Neonates (age 0-28 days)
- Infants and children (29 days - 17 years)
- Adults (18+ years)
- Elderly (65+ years, 80+ years for some conditions)

c) Birth weight (for neonates):

- Extremely low birth weight (<1000g)
- Very low birth weight (1000-1499g)
- Low birth weight (1500-2499g)
- Normal birth weight (≥2500g)

d) Severity markers:

- Use of mechanical ventilation
- ICU admission
- Length of stay categories
- Death during hospitalization

e) Surgical complexity (for surgical cases):

- Major procedures
- Moderate procedures
- Minor procedures

Special group: DRG 0000

Cases that cannot be properly classified due to coding errors, invalid codes, or missing critical information are assigned to DRG 0000, which serves as an "ungroupable" category and typically triggers data quality review.

C.4. Linkage Process: From Patient Data to DRG Code

C.4.1 Data Extraction

For each hospitalization episode in our dataset, we extract the following variables from the Medical Insurance Fund Settlement List:

From diagnosis fields:

- **Principal diagnosis (主要诊断):** the condition established to be chiefly responsible for the admission
- **Secondary diagnoses (其他诊断):** up to 20 additional diagnosis codes including comorbidities, complications, and other conditions

From procedure fields:

- **Principal procedure (主要手术及操作):** the most significant procedure performed
- **Other procedures (其他手术及操作):** up to 20 additional procedure codes

From demographic and clinical fields:

- Age, Sex, Birth weight (for neonates), Length of stay, Admission source, Discharge disposition, Total hospitalization costs

C.4.2 Grouping Algorithm

We implement the CHS-DRG grouping algorithm using a multi-step classification process:

STEP 1: Check for Pre-MDC conditions

If patient has organ transplantation, ECMO, total artificial heart, or ventilator support ≥96 hours, assign to corresponding Pre-MDC ADRG and return.

STEP 2: Determine primary MDC

Map principal diagnosis to MDC using ICD-10 chapter/block classification. If operating room procedure exists, check if procedure determines MDC.

STEP 3: Identify ADRG

Within assigned MDC:

- ◦ If patient has operating room procedure: Match (principal_diagnosis, principal_procedure) to surgical ADRG table
- ◦ Else if patient has non-OR procedure: Match procedure to non-OR procedure ADRG table
- ◦ Else: Match principal_diagnosis to medical ADRG table

STEP 4: Determine CC level

Initialize CC_level = 0. For each secondary diagnosis, check if it is in MCC or CC list and relevant to principal diagnosis. Apply exclusion logic to remove non-significant conditions.

STEP 5: Apply severity factors

Identify applicable severity dimensions: age categories, birth weight categories, procedure complexity level, death flag.

STEP 6: Final DRG assignment

Map (ADRG_code, CC_level, severity_factors) to final DRG_code using DRG decision tree. If mapping fails, assign DRG_code = 0000.

C.4.3 Implementation Details

In our study, we implement this algorithm using Python with the following specifications:

Data preprocessing:

- Validate all ICD-10 diagnosis codes against the NHSA code table
- Validate all ICD-9-CM-3 procedure codes against the NHSA code table
- Clean and standardize code formats

- Handle missing values: cases with missing principal diagnosis → DRG 0000[3]

Grouping engine:

- Load official DRG grouping tables published by NHSA
- Load CC/MCC exclusion lists
- Process cases sequentially through the grouping algorithm
- Record group assignment at each level (MDC, ADRG, DRG)
- Flag ungroupable cases for manual review

Quality checks:

- Verify >95% of cases successfully group (not DRG 0000)
- Check distribution across MDCs matches expected epidemiological patterns
- Validate that case mix matches institutional case mix profiles
- Review high-cost outliers within each DRG

C.5. Relative Weight Assignment

C.5.1 Concept of Relative Weight

The relative weight (RW) is a numerical value assigned to each DRG that represents the average resource intensity required to treat cases in that group, relative to the overall average hospitalization. By convention:

- RW = 1.0 represents the average resource consumption across all hospitalizations
- RW > 1.0 indicates above-average resource consumption (more complex/costly)
- RW < 1.0 indicates below-average resource consumption (less complex/costly)

For example:

- Kidney transplant (AE19): RW = 11.5621 (consumes 11.5621 times average resources)
- Allogeneic hematopoietic stem cell transplant (AG19): RW = 21.1225

[3] DRG 0000 cases do not receive an RW value. These cases are flagged, and depending on model requirements: excluded from primary analysis, or addressed via sensitivity analyses (e.g., imputation or separate category).

- Autologous hematopoietic stem cell transplant (AG29): RW = 7.1748
- Intracranial vascular procedure with hemorrhage (BC19): RW = 7.1143
- Respiratory infection without severe complications (ES35): RW = 0.3807
- Normal vaginal delivery (OR19): RW = 0.3602

C.5.2 Methodology for Weight Calculation

The NHSA calculates DRG relative weights using aggregated cost data from participating hospitals. The general methodology follows these steps:

Step 1: Data collection

Gather hospitalization cost data from representative sample of hospitals nationwide including total costs, case volume per DRG, geographic region, hospital level, and both public and qualified private hospitals.

Step 2: Cost standardization

Adjust raw costs for regional price differences, hospital teaching status and complexity, urban vs. rural location, and outlier costs.

Step 3: Calculate DRG average cost

For each DRG, compute the mean standardized cost across all cases in that group.

Step 4: Calculate overall average cost

Compute the mean cost across all hospitalizations.

Step 5: Calculate relative weight

For each DRG: RW = DRG_Average_Cost / Overall_Average_Cost

Step 6: Validation and calibration

Ensure weights are stable across regions, clinically coherent, accurately predict costs, and adjust for biases.

Step 7: Annual updates

Recalculate weights annually using recent data, incorporate new procedures and diagnoses, adjust for practice changes, and implement smoothing.

C.5.3 Published Relative Weights

The official DRG relative weights are published by the NHSA and updated annually. For DRG 2.0 (2024 edition), hospitals and researchers can access the weight table through:

- National Healthcare Security Administration's official website
- Regional medical insurance bureaus
- DRG grouping software platforms approved by NHSA

Weight distribution characteristics in DRG 2.0:

- Range: 0.0456 (e.g., TU19, Childhood Disorders of Psychological Development) to 21.1225 (AG19, Allogeneic Hematopoietic Stem Cell Transplantation)
- Median: 0.8001
- Mean: 1.00 (by definition)
- Standard deviation: 2.25
- Interquartile range: 0.5567 - 1.29315

C.6. Construction of DRG_Severity_Score Variable

C.6.1 Definition

For each patient encounter i in our dataset, we define:

$$DRG\_severity\_score_i = RW_k$$

where k is the DRG group to which patient i is assigned, and RW_k is the official relative weight for that DRG.

C.6.2 Interpretation

The DRG_severity_score serves as a composite measure that simultaneously captures:

- **Disease severity:** More severe conditions typically require more intensive treatment
- **Clinical complexity:** Presence of complications and comorbidities
- **Resource intensity:** Expected resource consumption for treatment
- **Case mix:** Allows comparison of patient populations with different disease profiles

Appendix D: Measurement of Professional Commitment Dimensions

We operationalize physicians' professional commitments using Large Language Model (LLM)-based classification of their online Q&A responses. For each physician *i* and patient *j*, we analyze all Q&A responses posted by physician *i* prior to patient *j*'s admission date, creating four moderating variables that capture cumulative public professional stance.

D1. LLM-Based Classification Methodology

We employ a prompt-engineered LLM to classify each Q&A response across four dimensions. The classification prompt instructs the LLM to act as a rigorous medical information research assistant, analyzing only the physician's response content (excluding patient questions). The LLM evaluates each response based on explicit definitions and scoring criteria, with examples provided to ensure consistent interpretation.

D2. Professional Commitment Dimensions

D2.1. Treatment Philosophy (TP_ratio)

Definition:

- Binary indicator (0/1) capturing whether a physician's response emphasizes patient-centered care values, including:

  - Shared decision-making and respect for patient autonomy
  - Balancing clinical benefits against potential harms (beneficence/nonmaleficence)
  - Consideration of patients' personal circumstances (psychological, economic, social factors)
  - Emphasis on quality of life and treatment burden

Scoring:

A response receives TP=1 if it explicitly discusses patient preferences, collaborative decision-making, or individualized care considerations; otherwise TP=0.

Calculation:

TP_ratio = (Number of physician i's responses with TP=1) / (Total responses by physician i prior to patient j's admission)

Examples from Online Forum:

Example TP-1:

- **Original Chinese:** "病人选择可根据自己的病情、心理承受、经济等自身因素考虑选择治疗方案。"
- **English Translation:** "Patients can choose treatment plans based on their own condition, psychological tolerance, economic factors, and other personal circumstances."
- **Classification:** TP=1 (explicitly empowers patient choice and considers multiple personal dimensions)

Example TP-2:

- **Original Chinese:** "我已经权衡了所有情况才会给你推荐治疗方案，但最终还是由您自己决定。"
- **English Translation:** "I have weighed all circumstances before recommending a treatment plan to you, but ultimately the decision is yours."
- **Classification:** TP=1 (respects patient autonomy and shared decision-making)

Example TP-3:

- **Original Chinese:** "我觉得首先需做一下全面的检查，评估一下再根据经济情况制定治疗方案！"
- **English Translation:** "I think we should first conduct a comprehensive examination and evaluation, then formulate a treatment plan based on your economic situation."
- **Classification:** TP=1 (considers patient's economic circumstances in treatment planning)

D2.2. Evidence-Based Standards (EBS_ratio)

Definition:

- Binary indicator (0/1) capturing whether a physician's response demonstrates commitment to scientific knowledge and professional competence through:
  - Explicit citation of clinical practice guidelines
  - Reference to specific research evidence (e.g., randomized controlled trials)
  - Mention of professional society recommendations

Scoring:

A response receives EBS=1 if it explicitly references external evidence sources or standards (e.g., "According to the 2009 American Thyroid Association guidelines..."); otherwise EBS=0. General medical knowledge without citation receives EBS=0.

Calculation:

EBS_ratio = (Number of physician i's responses with EBS=1) / (Total responses by physician i prior to patient j's admission)

Examples from Online Forum:

Example EBS-1:

- **Original Chinese:** "根据目前最新的欧美国家诊治甲状腺癌的指南和规范，您的情况应该进行碘-131 治疗..."
- **English Translation:** "According to the latest European and American guidelines for treating thyroid cancer, your situation should receive iodine-131 treatment..."
- **Classification:** EBS=1 (explicitly cites international clinical guidelines)

Example EBS-2:

- **Original Chinese:** "桥本甲状腺炎合并癌症发生率达 40%以上，故建议桥本甲状腺炎合并甲状腺结节者应积极外科治疗。"
- **English Translation:** "The incidence of cancer combined with Hashimoto's thyroiditis reaches over 40%, therefore it is recommended that patients with Hashimoto's thyroiditis combined with thyroid nodules should actively pursue surgical treatment."

- **Classification:** EBS=1 (cites specific statistical evidence to support clinical recommendation)

Example EBS-3:

- **Original Chinese:** "大颗粒淋巴细胞白血病...FAB 协作组把其归为慢性 T 淋巴细胞白血病。...REAL 分类将 LGLL 分为 T-LGLL 和 NK-LGLL。"
- **English Translation:** "Large granular lymphocytic leukemia...The FAB Cooperative Group classifies it as chronic T-cell lymphocytic leukemia...The REAL classification divides LGLL into T-LGLL and NK-LGLL."
- **Classification:** EBS=1 (references multiple professional classification systems)

D2.3. Resource Stewardship (RS_ratio)

Definition:

- Binary indicator (0/1) capturing whether a physician's response reflects commitment to social justice through responsible resource use, including:
  - Discussion of cost-effectiveness or financial burden
  - Recommendation of more affordable alternatives
  - Advice to avoid unnecessary tests or procedures
  - Consideration of resource efficiency

Scoring:

A response receives RS=1 if it discusses costs, value, or resource efficiency considerations; otherwise RS=0.

Calculation:

RS_ratio = (Number of physician i's responses with RS=1) / (Total responses by physician i prior to patient j's admission)

Examples from Online Forum:

Example RS-1:

- **Original Chinese:** "费用约 4~5 万元。"
- **English Translation:** "The cost is approximately 40,000-50,000 yuan."
- **Classification:** RS=1 (provides explicit cost information)

Example RS-2:

- **Original Chinese:** "手术费用 8，9 千元，没有自费项目，先来我门诊，周四下午给你加个号。手术需要排队的，要等几个月。"
- **English Translation:** "Surgery costs 8,000-9,000 yuan, with no out-of-pocket items. Come to my clinic first, I'll add you to Thursday afternoon appointments. Surgery requires queuing and will take a few months."
- **Classification:** RS=1 (detailed cost breakdown and clarifies no additional self-pay items)

Example RS-3:

- **Original Chinese:** "你好! 根据情况，畸形程度，可以考虑手术矫正。风险有截骨处愈合延迟，不愈合；矫枉过正，或矫正不足，矫正后双侧不完全一致等。疤痕不大，约 10cm 长，是否会疤痕增生要看个人的皮肤特质。费用约 4~5 万元。"
- **English Translation:** "Hello! Depending on the situation and degree of deformity, surgical correction can be considered. Risks include delayed osteotomy healing, non-union; overcorrection or undercorrection, bilateral asymmetry after correction, etc. The scar is not large, about 10cm long; whether there will be scar hyperplasia depends on individual skin characteristics. Cost is approximately 40,000-50,000 yuan."
- **Classification:** RS=1 (provides comprehensive cost information along with risk-benefit assessment)

D2.4. Explicitness (explicit)

Definition:

Continuous measure (0-1 scale) capturing the clarity, specificity, and unambiguousness of a physician's position-taking. According to commitment-consistency theory, explicit, definitive public statements create stronger behavioral consistency pressure than vague or qualified statements.

Scoring:

The LLM assigns each response a score from 1-10:

- 1-3 (Low explicitness): Vague, general, or noncommittal statements with possibility language (e.g., "could consider," "suggest consultation before deciding")
- 4-7 (Medium explicitness): Specific recommendations with qualifications or interpretive flexibility (e.g., "typically requires treatment, but case-dependent")
- 8-10 (High explicitness): Direct, unambiguous, actionable recommendations representing clear clinical positions (e.g., "definitively indicated," "costs approximately 40,000 yuan")

Calculation:

explicit = Average[(Score_k - 1)/9] across all responses k by physician i prior to patient j's admission, normalized to 0-1 scale.

Examples from Online Forum:

Example Explicit-1 (High Explicitness - Score 9/10):

- **Original Chinese:** "骨头基本上已经长好。建议去掉拐杖锻炼走路，骨头不会断掉，放心。"
- **English Translation:** "The bone has basically healed. I recommend removing the crutches and practicing walking; the bone will not break, don't worry."
- **Classification:** Explicitness = 9/10 (direct, unambiguous recommendation with reassuring definitive statement)

Example Explicit-2 (High Explicitness - Score 10/10):

- **Original Chinese:** "甲状腺癌是所有癌中预后最好的一种，乳头状癌又是甲状腺癌中预后最好的一种，并且是微小癌，只有一个毫米，手术后和正常人一样，不需担心，口服一粒半即可。"

- **English Translation:** "Thyroid cancer has the best prognosis among all cancers, papillary carcinoma has the best prognosis among thyroid cancers, and yours is a microcarcinoma, only one millimeter. After surgery, you will be the same as a normal person. No need to worry, just take one and a half tablets orally."
- **Classification:** Explicitness = 10/10 (extremely definitive statements about prognosis with specific medication dosage)

Example Explicit-3 (Medium Explicitness - Score 6/10):

- **Original Chinese:** "如果症状很重影响了正常生活，保守治疗不能缓解，身体条件又能承受，可以考虑手术。"
- **English Translation:** "If symptoms are severe and affect normal life, conservative treatment cannot provide relief, and the physical condition can tolerate it, surgery can be considered."
- **Classification:** Explicitness = 6/10 (conditional recommendation with multiple qualifications; uses "can be considered" rather than definitive language)

D3. Annotated Classification Examples

Table D1 presents three annotated examples from the LLM classification prompt, demonstrating how the algorithm applies scoring criteria to actual physician responses. These examples were selected to represent diverse combinations of professional dimensions and explicitness levels, providing the LLM with clear guidance for consistent classification.

Table D1. Annotated Examples for LLM Classification[4]

Example	Physician Response Text (English Translation)	TP	EBS	RS	Explicit
1	1. You may be admitted to our department for inpatient treatment using bone cement injection therapy; 90% of patients can achieve fracture fixation and pain relief. 2. Hospitalization lasts approximately one week, with a cost of around 40,000 yuan. 3. The material costs for this procedure are out-of-pocket	0	0	1	9

[4] Note on data privacy: The physician response texts displayed in Table D1 are English translations of the original Chinese posts. To avoid the data leakage of the hospital studied in this paper, the original Chinese post content authored by the physicians of the hospital is not displayed here but available with the authors upon request.

	expenses in Shanghai; please refer to your local medical insurance policy for coverage.				
2	Patients may choose their treatment plan based on their own condition, psychological tolerance, and economic factors.	1	0	1	6
3	It can be confirmed that, based on your description, there are clear indications for iodine-131 ablation (referring to the relevant guidelines from the 2009 American Thyroid Association).	0	1	0	10

Notes:

Example 1 demonstrates high resource stewardship (RS=1) through explicit cost discussion ("approximately 40,000 yuan for one week") and high explicitness (9/10) with specific, actionable details. It scores 0 on treatment philosophy and evidence-based standards as it focuses on procedural information without discussing patient preferences or citing guidelines.

Example 2 illustrates treatment philosophy (TP=1) by explicitly empowering patient choice based on "their own condition, psychological tolerance, and economic factors," while also showing resource stewardship (RS=1) through economic consideration. The moderate explicitness (6/10) reflects the general guidance without specific recommendations.

Example 3 exemplifies evidence-based standards (EBS=1) through explicit guideline citation ("referring to the 2009 American Thyroid Association guidelines") and maximum explicitness (10/10) with definitive language ("can be certain... has clear indications").

D4. Implementation and Validation

We implemented this classification system using Claude 3.5 Sonnet API with temperature=0 to ensure deterministic outputs. Each physician response was classified independently, with the LLM outputting structured JSON containing the four dimension scores. To validate classification reliability, we manually annotated a random sample of 200 responses and compared them with LLM outputs, achieving substantial agreement (Cohen's kappa > 0.75 for binary dimensions, Pearson correlation > 0.82 for explicitness).

Content analysis of physicians' online responses confirms that they publicly emphasize patient welfare and evidence-based care responding to patient concerns about treatment approaches, clinical reasoning, and medical expenses. Across 71,711 treatment-relevant online consultations, physicians' responses devoted on average 26.9% to treatment approaches discussions with patient-centered care, 14.3% to cost-conscious recommendations and resource considerations, and 18.4% to clinical guidelines and scientific reasoning. These three professional dimensions collectively comprised 42.8% of response content, establishing the empirical basis for public commitment in our theoretical framework. Examining the cost-conscious dimension more closely, we find that 53.4% of physicians' discussions on medication costs explicitly expressed propensity toward cost reduction through mechanisms such as generic drug recommendations, therapeutic substitutions, or value-based treatment alternatives, while 21.8% acknowledged cost constraints without specific reduction strategies. These public articulations of cost-consciousness and resource stewardship establish what we term a cognitive anchor—a publicly declared standard that generates consistency pressure on subsequent clinical behavior.

Appendix E: Knowledge Graph-Based Disease Recognition Process

In order to accurately capture the relevance of online physician information provision to the chief complaints of patients in offline encounters, we have implemented a sophisticated information extraction approach based on medical knowledge graph to ensure precise matching between online consultations and offline diagnosed diseases. Below is a detailed explanation of the disease recognition process based on online conversation content.

E1. Mention Identification

The Mention Identification phase combines two sophisticated methodologies: dictionary-based matching and a BERT_CRF (Souza et al. 2019) (Bidirectional Encoder Representations from Transformers with Conditional Random Fields) model for Named Entity Recognition (NER). This dual approach is designed to harness the strengths of both traditional and advanced AI techniques in recognizing medical entities.

- Dictionary-based Matching: This involves the creation of a comprehensive medical lexicon that includes a wide array of disease names, symptoms, and related medical terminology. The dictionary is meticulously curated from authoritative medical literature, databases, and ontologies to ensure it covers both common and rare conditions. During the mention identification process, the text from online consultations is scanned for exact matches within this dictionary. This method is particularly effective for identifying well-defined and widely recognized medical terms.
- BERT_CRF-Based NER: The incorporation of a BERT_CRF model for NER represents the application of state-of-the-art AI in understanding the context and nuances of natural language used in online conversations. Unlike the static approach of dictionary-based matching, the BERT_CRF model dynamically interprets the text by considering the surrounding words, linguistic patterns, and the overall context. This allows for the identification of disease mentions that may be phrased in less conventional terms or embedded in complex sentence structures. The CRF layer enhances the model's ability to predict the correct labels for each word in a sentence by considering the context provided by adjacent words, thus improving the precision of entity recognition.

When both the dictionary-based matching and BERT_CRF model identify disease mentions within the same conversation snippet, a comparative evaluation is conducted. The preference is given to the longer of the two identified mentions, under the assumption that the more detailed term likely provides a higher level of specificity and relevance to the patient's condition. This strategy is rooted in the understanding that longer, more descriptive mentions often convey a more precise indication of the disease entity, thereby enhancing the accuracy of the disease recognition process.

To ensure the reliability and effectiveness of the Mention Identification phase, extensive validation exercises are conducted. We manually annotated a random sample of 1,000 online conversations as our gold standard validation set. The dual approach achieved robust performance with:

- Dictionary-based matching: Precision = 0.92, Recall = 0.87, F1 = 0.89
- BERT_CRF model: Precision = 0.95, Recall = 0.93, F1 = 0.94
- Combined approach: Precision = 0.96, Recall = 0.94, F1 = 0.95

These validation results demonstrate that our combined approach outperforms either method alone, particularly in handling complex medical terminology and contextual variations.

E2. Entity Linking

The Entity Linking phase involves sophisticated techniques to accurately map the identified disease mentions from online conversations to specific entities within a structured medical domain dictionary. In this phase, we employ Sentence-BERT (Reimers and Gurevych 2019), a modification of the pre-trained BERT model optimized for generating sentence embeddings. This approach allows for the capture of semantic similarities between the disease mentions identified in the previous step and the entities listed in the medical domain dictionary. Sentence-BERT enhances the traditional BERT model by producing embeddings that effectively represent the meaning of entire sentences, rather than just individual words or phrases. This is particularly important in the medical domain, where the context and nuance can significantly alter the meaning and relevance of terms.

For each disease mention, the Sentence-BERT model generates embeddings and compares them with embeddings of entities in the medical domain dictionary. We calculate the semantic similarity between the mention and each entity, ranking entities based on their similarity scores. The top 20 entities with the highest similarity scores are selected as candidate entities for further evaluation. This selection process ensures that we consider a broad set of potentially relevant entities while maintaining a focus on those most likely to match the mention semantically.

Once the candidate entities are selected, we refine the list by examining the overlap between the words in the disease mention and the words in each candidate entity. A candidate is considered a potential match if the number of overlapping words constitutes at least half the length of the mention. This criterion is based on the premise that a higher degree of word overlap between the mention and an entity indicates a greater likelihood of relevance. This step significantly reduces the likelihood of incorrect linkages by ensuring that the linked entity closely matches the context and specificity of the mention.

The Entity Linking phase incorporates a rigorous validation process. From our disease mention dataset, we randomly selected 500 mention-entity pairs for manual review by medical professionals. The validation results show:

- Overall accuracy: 93.2%
- Top-1 matching accuracy: 89.5%
- Top-3 matching accuracy: 95.8%
- Error analysis:
    - Wrong entity matching: 4.3%
    - Ambiguous cases: 2.5%
    - Missing appropriate entity: 3.2%

These results confirm the robustness of our entity linking approach in accurately mapping disease mentions to standardized medical entities.

E3. Knowledge Graph Search

The Knowledge Graph Search step leverages the extensive network of medical knowledge encapsulated in the knowledge graph to find direct linkages between the recognized entities and disease type entities. Our knowledge graph[5] is an expansive and intricately structured database containing approximately 44,000 entities that cover a wide spectrum of medical knowledge. These entities encompass various types of medical concepts such as symptoms, diseases, drugs, etc. (See Table E1 for

[5] https://github.com/pen-ho/medical_knowledge_graph_app-master

details). Additionally, the graph includes a multitude of relationships among these entities, such as symptom-disease associations, common treatment protocols, drug-disease interactions, and disease-disease relationships (See Table E2 for details). The graph's design facilitates a multi-dimensional exploration of medical concepts, allowing for a nuanced understanding of the connections and interdependencies within the medical domain.

The core objective of the Knowledge Graph Search step is to identify the specific disease types that are associated with the entities recognized in the online dialogues. Utilizing advanced graph search algorithms, we query the knowledge graph for direct linkages between the previously linked entities (from the Entity Linking phase) and disease type entities within the graph. This search is not merely a lookup; it is a sophisticated process that considers the context, strength, and nature of the connections between entities to deduce the most relevant disease types. The outcome of the graph search is a set of disease entity nodes that are directly linked to the recognized entities. These nodes represent the diseases that are most closely associated with the specific online conversations under analysis.

Table E1. Entity types defined in medical knowledge graph

Entity type	Number of entities	Examples
Check (诊断检查项目)	3,353	支气管造影: Bronchography; 关节镜检查: Arthroscopy
Department (医疗科目)	54	整形美容科: Plastic Surgery Department; 烧伤科: Burn Department
Disease (疾病)	8,807	血栓闭塞性脉管炎: Thromboangiitis Obliterans; 胸降主动脉动脉瘤: Thoracic Descending Aortic Aneurysm
Drug (药品)	3,828	京万红痔疮膏: Jing Wan Hong Hemorrhoid Ointment; 布林佐胺滴眼液: Brinzolamide Eye Drops
Food (食物)	4,870	番茄冲菜牛肉丸汤: Tomato with Bok Choy and Beef Meatball Soup; 竹笋炖羊肉: Bamboo Shoots Stewed with Lamb
Producer (药品大类)	17,201	通药制药青霉素V钾片: Tongyao Pharmaceutical Penicillin V Potassium Tablets; 青阳醋酸地塞米松片: Qingyang Acetate Dexamethasone Tablets
Symptom (疾病症状)	4,377	乳腺组织肥厚: Mammary Gland Hyperplasia; 脑实质深部出血: Deep Cerebral Parenchymal

		Hemorrhage
Total	44,111	

Table E2. Entity Relation Types Defined in Medical Knowledge Graph

Entity Relation type	Number of relations	Examples
belongs_to (属于)	8,844	<妇科,属于,妇产科> <Gynecology, belongs_to, Obstetrics and Gynecology>
common_drug（疾病常用药品)	14,649	<阳强,常用,甲磺酸酚妥拉明分散片> <erectile dysfunction, common_drug, Tamsulosin Hydrochloride Dispersible Tablets>
do_eat (疾病宜吃食物)	22,238	<胸椎骨折,宜吃,黑鱼> < thoracic vertebral fracture, do_eat, snakehead fish >
drugs_of（药品在售药品)	17,315	<青霉素V钾片,在售,通药制药青霉素V钾片> < Penicillin V Potassium Tablets, drugs_of, Tongyao Pharmaceutical Penicillin V Potassium Tablets >
need_check（疾病所需检查)	39,422	<单侧肺气肿,所需检查,支气管造影> <unilateral emphysema, need_check, Bronchography>
no_eat (疾病忌吃食物)	22,247	<唇病,忌吃,杏仁> < lip disease, no_eat, almonds>
recommend_drug（疾病推荐药品)	59,467	<混合痔,推荐用药,京万红痔疮膏> <mixed hemorrhoids, recommend_drug, Jing Wan Hong Hemorrhoid Ointment>
recommend_eat (疾病推荐食谱)	40,221	<鞘膜积液,推荐食谱,番茄冲菜牛肉丸汤> <effusion in the tendon sheath, recommend_eat, Tomato with Bok Choy and Beef Meatball Soup >
has_symptom（疾病症状)	99,492	<早期乳腺癌,疾病症状,乳腺组织肥厚> <Early-stage breast cancer, symptom, Mammary Gland Hyperplasia>
acompany_with (疾病并发疾病)	12,029	<下肢交通静脉瓣膜关闭不全,并发疾病,血栓闭塞性脉管炎> <incompetence of the lower limb venous valves, acompany_with, Thromboangiitis Obliterans>
Total	294,149	

E4. Diagnosed Disease Matching

The Diagnosed Disease Matching phase employs sophisticated natural language processing (NLP) and machine learning techniques to ensure that the insights derived from online patient-physician interactions are effectively applied to the treatment and understanding of diseases in offline inpatient settings. To facilitate the matching of diagnosed disease names from offline inpatient admissions with the identified disease entities from online consultations, we employ the Word2Vec model. Word2Vec is a powerful tool in the field of NLP that generates vector representations of words in a high-dimensional space. By training on a large corpus of medical texts, Word2Vec captures the semantic and syntactic nuances of disease names and medical terminology, allowing for a nuanced understanding of disease concepts that goes beyond simple keyword matching.

The process utilizes cosine similarity to match the vector representations of diagnosed disease names, coded in Diagnosis-related Groups (DRG), with those of disease entities identified from online conversations. The average cosine similarity between the matched diagnosed disease names and identified disease entities is 0.8466, spanning values from 0.775 to 1.0. DRG codes, which categorize hospital cases into groups to facilitate payment and analysis, serve as a structured framework for this matching. By associating diagnosed diseases with specific DRG codes, and comparing these to the disease entities identified online, the matching process accounts for the broad categorizations of diseases as well as their specific manifestations.

We then calculate the cumulative number of online replies by physicians to queries regarding diseases associated with specific DRG codes. This quantification is achieved by aggregating the replies related to diseases within the same DRG categories, providing a metric of physician engagement and expertise in relation to specific diseases. This variable, replacing the original independent variable (*Ln_QA*), provides a more accurate reflection of the extent to which information asymmetry—the gap between what patients know and what physicians know—is being reduced regarding the specific diagnosed disease.

E5. Text Similarity of Suggested Treatment

To assess the clinical relevance and quality of physicians' online treatment recommendations, we implement a comprehensive text similarity analysis that compares these recommendations with standard clinical guidelines stored in our medical knowledge graph. This analysis follows a multi-step process:

First, we extract treatment-specific content from physicians' online responses using the same BERT_CRF model described in the Mention Identification phase. The model is specifically trained to identify text segments that contain treatment recommendations, medical procedures, and therapeutic suggestions. These extracted segments form the basis for our treatment similarity analysis.

Second, for each identified disease entity from the Knowledge Graph Search phase, we retrieve the corresponding standard treatment protocols from our knowledge graph. These protocols represent evidence-based clinical guidelines and established treatment practices for each disease type. The knowledge graph contains structured information about recommended treatments, including first-line therapies, alternative approaches, and condition-specific interventions.

Third, we employ Sentence-BERT to generate vector representations of both the extracted physician treatment recommendations and the standard treatment protocols from the knowledge graph. This embedding process captures the semantic meaning of the treatment descriptions, allowing for nuanced comparison beyond simple keyword matching.

We then calculate two key metrics:

1. *Physician_Treatment_Sim*$_{ij}$: For each physician *i*, we compute the cosine similarity between their online treatment recommendations and the knowledge graph's standard treatments for diseases matching inpatient *j*'s diagnosis. This is calculated as the average similarity score across all relevant treatment recommendations made by the physician prior to inpatient *j*'s admission.
2. *Other_Treatment_Sim*$_{ij}$: Similarly, we calculate the average cosine similarity between treatment recommendations made by other Hospital A physicians (excluding physician *i*) and the knowledge graph's standard treatments for diseases matching inpatient *j*'s diagnosis. This metric serves as a benchmark for comparing individual physician's alignment with clinical guidelines against their peers.

The similarity scores range from 0 to 1, where higher scores indicate greater alignment with standard clinical guidelines. By comparing these two metrics, we can assess not only how closely a physician's recommendations align with established clinical practices but also how their alignment compares to their peers at the same hospital.

References

Johnston AC, Worrell JL, Di Gangi PM, Wasko M. (2013) Online Health Communities: An Assessment of the Influence of Participation on Patient Empowerment Outcomes. *Information Technology & People*. (26:2):213–235.

Kim Y. (2014) Convolutional Neural Networks for Sentence Classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1746–1751.

Reimers N, Gurevych I. (2019) Sentence-bert: Sentence Embeddings Using Siamese Bert-Networks. arXiv preprint arXiv:1908.10084.

Serrano CI, Karahanna E (2016) The Compensatory Interaction between User Capabiities and Technology Capabilities in Influencing Task Performance: An Empirical Assessment in Telemedicine Consultations. *MIS Quarterly* 40(3):597-621.

Souza F, Nogueira R, Lotufo R (2019) Portuguese Named Entity Recognition Using BERT-CRF. arXiv preprint arXiv:1909.10649.

Appendix F: Text Classification

To construct our measure of treatment-specific online information provision (*DiseaseSpecific_QA*$_{ij}$), we need to identify Q&As that contain specific medical treatment recommendations. Following Haas-Wilson (2001), who argues that reliable treatment information is key to creating better-informed patients, we deploy a supervised neural-network classification model to identify whether a physician reply contains recommendations on specific medical treatments. An iterative labeling and training process is used to ensure the performance of the classification model. The output of text classification is a binary indicator of the desired type. Our model achieves reasonably good performance with training/testing F1 score as 0.988/0.914. The detailed text labeling and training steps as well as the classification results are reported here.

F1. Manual Labeling Process

1. Sample Set Construction

To ensure the quality of the manual labels while balancing the cost, we limited the initial scope of the sample selection process to all consultation texts coming from two representative general hospitals in Shanghai city. Below are the four steps we took to construct our sample dataset for manual labelling.

- **Step 1**: From the initial 210,148 physician-side consultation text records,[6] we extracted the unique ID for all physicians in these records. We also extracted the year dummy.
- **Step 2**: We randomly selected 20% of all unique consultations (based on consultation IDs) that belonged to a physician in a particular year, which yielded 24,446 unique consultation IDs.
- **Step 3**: We put all the text consultation records that belonged to the 24,446 unique consultation IDs into our labeling sample set.
- **Step 4**: We randomly divided all the selected text records into blocks. Each block contained 2,000 text records.

In total, we constructed a sample set of more than 40,000 physician-side text records. The block segmentation was designed to facilitate the manual labeling process later.

[6] Note that the unique identifier for the text records is the consultation ID generated by the platform for each physician-patient text consultation encounter plus the floor number of the text record within all the text records belonging to the consultation ID.

2. Manual Labeling Process

The initial round of manual labeling involved two coders (undergraduates in a Chinese university) spanning several months. Before the labeling, each coder was briefed on the objective of the labeling and the definition of the required label. More specifically, we defined our label – treatment recommendation as follows:

> *"Physician recommends surgery, Chinese medicine, cancer treatment (e.g. chemotherapy, radiology, targeted drug), or physical therapy to the patient."*

To ensure the independence of the coders, we used the block segmentation of the sample data. We gave each coder only one block to label at a time. And the sequences of blocks given to each coder were different from each other so that each coder would not label the same block of sample simultaneously.

After both coders completed the labelling, we compared the consistency of their labeling and kept only the consistent ones, which yielded us a labeled sample set of 27,142 text records.

F2. Machine Classification

We developed and trained a variant of the convolutional neural networks (CNN) model for sentence classification as developed by Kim (2014). The environment used to build and train the classification model is Python v3.6.1 along with PyTorch v1.0. Sentence segmentation and word vectorization were conducted using an open-source Python library developed for Chinese text – Jieba v0.39.

We used the hold-out validation method[7] to evaluate the performance of our model. Before training, we split the manually labeled 27,142 text records randomly into the training set (70%) and testing set (30%). As a result, the training set contained 18,999 text records and the testing set contained 8,143 text records. Table F1 reports the text classification performance. We used accuracy and an F_1 score to evaluate the classification performance due to the imbalance distribution of positive and negative label (see Table F2). An F_1 score is the harmonic average of the precision and recall score, which is widely considered a more informative performance measure.

[7] We did not use the k-fold cross validation method here because our primary objective was not to improve on the classification model. Hence, balancing against the computational complexity of conducting k-fold cross validation, we contend that a simple hold-out validation is sufficient for us to evaluate the out-of-sample performance of the trained model.

Table F1. Text Classification Performance

Text Class	Training Accuracy	Training F_1 Score	Testing Accuracy	Testing F_1 Score
Treatment Recommendation	99.65%	98.83%	97.56%	91.42%

Table F2. Distribution of Machine-Generated Labels

Text Class	Frequency	
	Manual Label	Machine Label
Treatment Recommendation	7.94%	7.34%
# of text records	27,142	114,149

As shown in Table F1, the F_1 score of the binary classification model on the hold-out sample is above 90%. Table F2 shows that the distribution of our text class in the manually labelled dataset and the machine-generated full dataset is highly similar, suggesting a good out-of-sample performance as well. We subsequently deployed the trained models to the full text sample to classify the replies by physicians that contain specific treatment recommendation.

Appendix G: Leveraging LLM for Address Inference and Workplace Classification

For our analysis of how online information provision affects offline healthcare delivery across different patient populations, we need to identify each patient's residential location (for calculating local Internet penetration rates) and workplace sector (for inferring socioeconomic characteristics). In our dataset, we have two relevant but partially masked fields for each inpatient admission:

- Contact addresses with masked details for privacy protection (e.g., "Dalian Shahe**" where asterisks represent masked characters)
- Self-disclosed workplace information with varying levels of detail

To systematically extract structured information from these data, we employ ChatGLM[8], an autoregressive large language model (LLM) based on the Transformer architecture. The LLM-based extraction process focuses on two main tasks:

(1) **Administrative Zone Inference**: From the partially masked contact addresses, we use ChatGLM to infer the complete administrative location information including province, city, and district. Table G1 demonstrates how the model handles various types of masked addresses, from complete city information to partial street names.

(2) **Workplace Sector Classification**: From the self-disclosed workplace information, we utilize ChatGLM to categorize workplaces into following sectors:

Public Sector

- Healthcare Institutions (Hospitals, Clinics)
- Educational Institutions
    - Universities & Research Institutes
    - Primary/Secondary Schools & Kindergartens
- Government Agencies
- Other Public Institutions

State-owned Sector

- State-owned Enterprises

[8] https://chatglm.cn/

Private Sector

- Private Enterprises

Other

- Other Organizations

Table G2 shows examples of how the model classifies different workplace descriptions into these standardized categories. This systematic approach enables us to derive crucial variables for our analysis, such as *Web_Access* (based on administrative zones) and *High_Education* (based on workplace sectors), while maintaining data privacy.

Table G1. ChatGLM Address Information Recognition Examples

Step	Content
Prompt	Please identify the province, city, and district information from the given address. Note: If the address includes clear local information such as street or village but lacks clear province and city information, prioritize reasoning it as Shanghai. Return format: {‘province’: ‘province’, ‘city’: ‘city’, ‘district’: ‘district’}
Input Example 1	Dalian Shahe** (大连沙河**)
Output Example 1	Based on the provided address "Dalian Shahe**", we can make the following inferences: 1. Province: Dalian is a city in Liaoning Province, so the province should be "Liaoning". 2. City: The address clearly mentions "Dalian", so the city is "Dalian". 3. District: The address mentions "Shahe", and there is a district in Dalian called "Shahekou District". Although the address is incomplete, "Shahe" is likely an abbreviation for "Shahekou District". Therefore, the province, city, and district are: {‘province’: ‘Liaoning’, ‘city’: ‘Dalian’, ‘district’: ‘Shahekou District’}
Input Example 2	Caoxi Road Zha** (漕溪路闸**)
Output Example 2	Based on the provided address "Caoxi Road Zha**", the full address is not provided, so it is difficult to directly determine the specific province, city, and district. However, "Caoxi Road" is closely associated with Shanghai, as there is a well-known Caoxi Road in Shanghai. Additionally, "Zha" might refer to the former Zhabei District of Shanghai, which has since merged with Jing’an District. Based on this information and the inherent pattern of address information, we infer that the address might be located in the following region: {‘province’: ‘Shanghai’, ‘city’: ‘Shanghai’, ‘district’: ‘Jing’an District’}
Input Example 3	Guilin Road 3** (桂林路 3**)
Output Example 3	Based on the provided address "Guilin Road 3**", but with no clear province or city information, and following the reasoning rule that if the address includes

	clear local information (such as street or village) but lacks clear province and city information, prioritize reasoning it as Shanghai. In Shanghai, Guilin Road is a well-known street in Xuhui District. Therefore, we can infer that the address is located in Xuhui District, Shanghai. Return format: {'province': 'Shanghai', 'city': 'Shanghai', 'district': 'Xuhui District'}

Table G2. ChatGLM Unit Information Recognition Examples

Step	**Content**
Prompt	Please determine the category of the unit name based on the following information. The unit categories include: Primary and Secondary Schools & Kindergartens, Hospitals, National Government Agencies, State-Owned Enterprises, Private Enterprises, Universities & Research Institutes, Other Public Institutions, Other Units. Return format: {'unit_type': 'unit type'}
Input Example 1	Shanghai Medical Insurance Bureau (上海市医保局)
Output Example 1	{'unit_type': 'National Government Agency'}
Input Example 2	East China University of Science and Technology (华东理工大学)
Output Example 2	{'unit_type': 'University & Research Institute'}
Input Example 3	Shanghai Film Studio (上海电影制片厂)
Output Example 3	{'unit_type': 'State-Owned Enterprise'}