LDL-콜레스테롤 추정식 개발

1. 배 경

- 심근경색이나 뇌졸중 같은 심뇌혈관질환(cardio-cerebrovascular disease)은 우리나라 사람의 대표적인 사망원인 질환이다 (우리나라 사람의 사망원인 질환 1위는 암이며, 심뇌혈관질환은 2위이다).
- LDL-콜레스테롤(low-density lipoprotein cholesterol)은 심뇌혈관질환의 주된 위험인자로서. 그 정확한 측정값을 임상(질환의 예방과 관리)에 적용하는 것이 중요하다.
- 하지만 전 세계적으로 건강검진과 같은 대규모 LDL-콜레스롤 측정에서, 비용을 절감하기 위해 직접 측정이 아닌 Friedewald 공식에 의한 계산법이 널리 이용되고 있다.
- 우리나라 국가건강검진에서도 중성지방(triglyceride) 수준이 400 mg/dL 이상인 사람들에 대해서는 LDL-콜레스테롤을 직접 측정하지만, 중성지방 400 mg/dL 미만인 사람들에 대해서 Friedewald 공식에 의한 계산값을 사용하도록 되어있다. 2015년 국가건강검진에서 중성지방을 측정한 14,024,331명 중 중성지방 농도가 400 mg/dL 이상인 사람은 281,059명으로 2%에 불과하였다. 이는 건강검진 수검자의 98%에 대하여 LDL-콜레스테롤 실측값이 아닌 계산값이 사용되었다는 것을 의미한다.
- 하지만 Friedewald 공식에 의한 계산값은 중성지방 농도 400 mg/dL 미만에서도, 그 정확도가 문제될 수 있음이 지적되어왔다. 중성지방 수준이 높을수록 LDL-콜레스테롤을 과소 추정하여 심뇌혈관질환 위험도를 낮게 평가하는 경향이 있다. 특히 중성지방 수준이 높고 LDL-콜레스테롤 계산값이 낮은 경우에서의 과소 평가 문제점이 지적되어 왔다.

2. 목 표

- LDL-콜레스테롤 추정식으로서 Friedewald 공식의 정확도를 평가하고(중성지방 < 400 mg/dL), 이의 문제점 개선하여 정확도를 높이는 추정식을 개발하는 것이다.

3. 기존 연구

(1) Friedewald formula

- 총 콜레스테롤(total cholesterol, **TC**)은 다음 세 가지로 구성된다: ① HDL-콜레스테롤(high-density lipoprotein cholestrol, **HDL-C**), ② LDL-콜레스테롤(low-density lipoprotein cholesterol, **LDL-C**), 그리고 ③ VLDL-콜레스테롤(very low-density lipoprotein cholesterol, **VLDL-C**). 즉

  TC = HDL-C + LDL-C + VLDL-C

- Friedewald 등은 1972년 논문에서 중성지방 400 mg/dL 미만에서 VLDL-콜레스테롤(**VLDL-C**)은 중성지방(triglyceride, **TG**)의 1/5 정도로 추정될 수 있음을 발견하였다. 즉 VLDL-콜레스테롤의 추정값 "**$VLDL\text{-}C_F = TG / 5$**"라는 것이다. 따라서 총 콜레스테롤(**TC**),

HDL-콜레스테롤(**HDL-C**), 중성지방(**TG**)의 실측값을 알고 있는 경우, LDL-콜레스테롤의 추정값 **LDL-C$_F$**는 다음과 같다.

LDL-C$_F$ = TC − HDL-C − (TG/5)

- 이 Friedewald 공식에서 총 콜레스테롤에서 HDL-콜레스테롤을 뺀 "**TC − HDL-C**"을 Non-HDL-콜레스테롤(**Non-HDL-C**)라고 하면,

LDL-C$_F$ = Non-HDL-C − (TG/5)

(2) Marin formula

- Matin 등(2013)은 Friedewald 공식의 대안으로서, LDL-콜레스테롤 추정값의 정확도를 개선하기 위한 새로운 방법을 제안하였다.
- Friedewald 공식에서 VLDL-콜레스테롤의 추정값을 "**TG / 5**"로 사용하는 반면, Martin은 아래 표와 같이 중성지방(Triglyceride)과 Non-HDL-C 수준에 따라 "**TG / 5**"의 5 대신에 다양한 값을 사용할 것을 제시하였다.

Triglyceride Levels, mg/dL[a]	Non-HDL-C, mg/dL					
	<100	100-129	130-159	160-189	190-219	≥220
7-49	3.5	3.4	3.3	3.3	3.2	3.1
50-56	4.0	3.9	3.7	3.6	3.6	3.4
57-61	4.3	4.1	4.0	3.9	3.8	3.6
62-66	4.5	4.3	4.1	4.0	3.9	3.9
67-71	4.7	4.4	4.3	4.2	4.1	3.9
72-75	4.8	4.6	4.4	4.2	4.2	4.1
76-79	4.9	4.6	4.5	4.3	4.3	4.2
80-83	5.0	4.8	4.6	4.4	4.3	4.2
84-87	5.1	4.8	4.6	4.5	4.4	4.3
88-92	5.2	4.9	4.7	4.6	4.4	4.3
93-96	5.3	5.0	4.8	4.7	4.5	4.4
97-100	5.4	5.1	4.8	4.7	4.5	4.3
101-105	5.5	5.2	5.0	4.7	4.6	4.5
106-110	5.6	5.3	5.0	4.8	4.6	4.5
111-115	5.7	5.4	5.1	4.9	4.7	4.5
116-120	5.8	5.5	5.2	5.0	4.8	4.6
121-126	6.0	5.5	5.3	5.0	4.8	4.6
127-132	6.1	5.7	5.3	5.1	4.9	4.7
133-138	6.2	5.8	5.4	5.2	5.0	4.7
139-146	6.3	5.9	5.6	5.3	5.0	4.8
147-154	6.5	6.0	5.7	5.4	5.1	4.8
155-163	6.7	6.2	5.8	5.4	5.2	4.9
164-173	6.8	6.3	5.9	5.5	5.3	5.0
174-185	7.0	6.5	6.0	5.7	5.4	5.1
186-201	7.3	6.7	6.2	5.8	5.5	5.2
202-220	7.6	6.9	6.4	6.0	5.6	5.3
221-247	8.0	7.2	6.6	6.2	5.9	5.4
248-292	8.5	7.6	7.0	6.5	6.1	5.6
293-399	9.5	8.3	7.5	7.0	6.5	5.9
400-13975	11.9	10.0	8.8	8.1	7.5	6.7

• Friedewald 공식과 Martin 방법은 모두 총 콜레스테롤(**TC**), HDL-콜레스테롤(**HDL-C**), 중성지방(**TG**)의 실측값을 이용하여 LDL-콜레스테롤을 추정한다. 하지만 Friedewald 공식은 중성지방에 대한 VLDL-콜레스테롤(**TG:VLDL-C**)의 비율로 고정된 **5**를 사용하는 반면, Martin 방법은 위의 표와 같이 중성지방(Triglyceride)과 Non-HDL-C 수준에 따라 분류된 하위 집단에 최적의 **TG:VLDL-C 비율**을 적용하고자 한다. 이를 위해 각 하위집단의 **TG:VLDL-C 비율**의 **중앙값(median)**을 도출하여 사용하였다. 이 각 하위집단의 중앙값을 **AF**(adjustable factor), 그리고 Martin 방법의 LDL-콜레스테롤의 추정값을 $\mathbf{LDL\text{-}C_M}$라고 하면,

$$\mathbf{LDL\text{-}C_M = Non\text{-}HDL\text{-}C - (TG/AF_{ij})}$$

• Martin 등(2013)은 Friedewald 공식에 비해 자신들의 방법이 LDL-콜레스테롤의 추정값의 정확도를 개선하였다고 보고하였다.

(3) Simpson formula

• Simpson 등(2020)은 Friedewald 공식의 대안으로 아래와 같은 회귀식 모델을 제시하였다. Simpson 방법의 LDL-콜레스테롤의 추정값을 $\mathbf{LDL\text{-}C_S}$라고 하면,

$$\mathbf{LDL\text{-}C_S} = \frac{\mathbf{TC}}{0.948} - \frac{\mathbf{HDL\text{-}C}}{0.971} - \left[\frac{\mathbf{TG}}{8.56} + \frac{\mathbf{TG \times Non\text{-}HDL\text{-}C}}{2140} - \frac{\mathbf{TG^2}}{16100}\right] - 9.44$$

• Simpson 방법은 중성지방 농도 400 mg/dL 이상인 집단의 LDL-콜레스테롤 추정값을 개선하고자 하는 것이었지만, 중성지방 농도 400 mg/dL 미만에서도 회귀식 모델이 추정식으로 사용될 수 있음을 제시한다.

(4) AI-based Model

• 최근에는 인공지능 알고리즘을 사용하여 LDL-콜레스테롤 추정값의 정확도를 높이려는 연구들이 있다. 『Deep neural network for estimating low density lipoprotein cholesterol, Taesic Lee 외 3명』

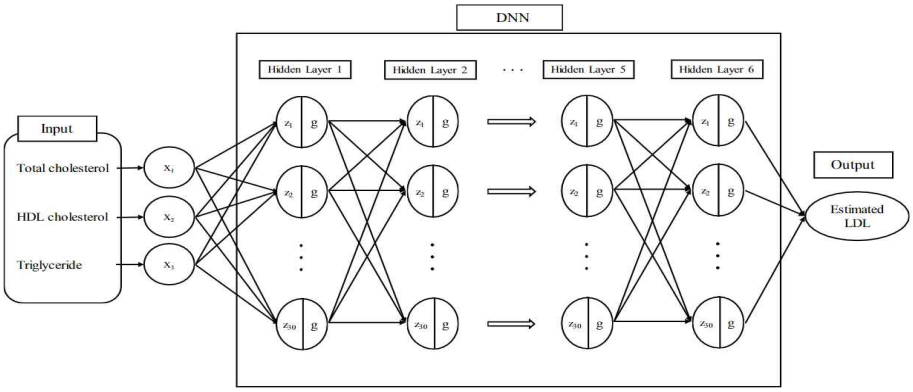

Fig. 1. DNN model for estimating LDL-C.
Input takes three values: total cholesterol, HDL−C, and triglyceride, and the output is the LDL-C. The DNN model consists of six hidden layers, and each hidden layer has 30 nodes. Each node's response is "z" based on activation function "g," where $z = bias + \Sigma_{i=1}{}^{n} X_i w_i$, and $g(z) = \max(0, z)$ (Rectified Linear Unit).

4. 데이터와 변수

(1) Data Set

- 여러분이 본 과제를 수행하기 위한 데이터로서 다음 두 가지가 제공된다.
  ① **Data Set 1 [Derivation Data Set]** 새로운 추정식을 개발하기 위해 사용되는 데이터
  ② **Data Set 2 [Validation Data Set]** 개발된 새로운 추정식의 타당도(LDL-콜레스테롤 추정값의 정확도 개선 등)를 평가하기 위해 사용되는 데이터
- **Data Set 2 [Validation Data Set]** 는 우리나라 국민을 대표할 수 있도록 표본이 추출되어, 타당성 분석에 적절하다 (n = 5,483명)
- **Data Set 1 [Derivation Data Set]** 는 **Data Set 2 [Validation Data Set]**과 비교하여, 중성지방 200mg/dL~ 399mg/dL인 사람의 비중이 높다 (n = 11,564명)

(2) Variables

- 두 데이터 셋은 다음과 같은 변수로 구성된다.
  ① **ID**: 각 개인의 고유 번호
  ② **Sex**: 각 개인의 성별 구분(남자 = 0, 여자 = 1)
  ③ **Age**: 각 개인의 연령(years)
  ④ **TC**: 각 개인의 총 콜레스테롤(total cholesterol) 실측값
  ⑤ **HDL-C**: 각 개인의 HDL-콜레스테롤(high-density lipoprotein cholestro) 실측값
  ⑥ **LDL-C**: 각 개인의 LDL-콜레스테롤(low-density lipoprotein cholesterol) 실측값
  ⑦ **TG**: 각 개인의 중성지방(triglyceride) 실측값
  ⑧ **Non-HDL-C**: 각 개인의 Non-HDL-콜레스테롤 계산값, **TC − HDL-C**
  ⑨ **VLDL-C**: 각 개인의 VLDL-콜레스테롤 계산값, **Non-HDL-C − LDL-C**
  ⑩ **TG_to_VLDL-C**: 각 개인의 TG:VLDL-C 비율 계산값, **TG / VLDL-C**

5. 새로운 추정식 개발과 성과 평가

(1) 새로운 추정식의 개발

- 먼저 **Data Set 1 [Derivation Data Set]**을 이용하여 Friedewald 공식, Martin 공식, Sampson 공식의 LDL-콜레스테롤 추정값을 도출한다.
- 아래 “성과 평가”를 참조하여 세 가지 공식(Friedewald 공식, Martin 공식, Sampson 공식)의 타당성을 평가한다.
- 이러한 타당성 평가를 기초로 새로운 추정식을 개발한다.

(2) 성과 평가

- 여러분이 개발한 새로운 추정식의 성과는 다음과 같은 두 가지 측면에서 평가된다.

① LDL-콜레스테롤 수준의 분류 일치도

- 임상적으로 LDL-콜레스테롤 수준은 다음 6가지로 분류되어 관리된다:

(1) 70 mg/dL 미만, (2) 70-99 mg/dL, (3) 100-129 mg/dL, (4) 130-159 mg/dL, (5) 160-189 mg/dL, 그리고 (6) 190 mg/dL 이상

- 따라서 여러분이 개발한 "LDL-콜레스테롤 추정값"과 "LDL-콜레스테롤 실측값(**LDL-C**)" 수준을 상기 6가지 집단으로 분류하고, 이의 분류 일치도를 계산한다.
- <Table 1>은 Friedewald 공식의 "LDL-콜레스테롤 추정값"과 "LDL-콜레스테롤 실측값(**LDL-C**)"의 분류 일치도를 분석한 결과이다.
- <Table 2>는 중성지방 수준에 따라, Friedewald 공식의 "LDL-콜레스테롤 추정값"과 "LDL-콜레스테롤 실측값(**LDL-C**)"의 분류 일치도를 분석한 결과이다. 중성지방 수준이 높은 집단일수록 분류 일치도가 낮아지는 것을 알 수 있다.
- <Table 2>에서 Friedewald 공식의 "LDL-콜레스테롤 추정값"과 "LDL-콜레스테롤 실측값(**LDL-C**)"의 분류 일치도는 80.2% 수준임을 알 수 있다.
- **Data Set 2 [Validation Data Set]**을 이용하여 세 가지 공식(Friedewald 공식, Martin 공식, Sampson 공식)의 "LDL-콜레스테롤 추정값"과 "LDL-콜레스테롤 실측값(**LDL-C**)"의 분류 일치도를 각각 계산하여 비교한다.
- **Data Set 2 [Validation Data Set]**을 이용하여, 여러분이 개발한 새로운 추정식의 분류 일치도를 계산하고, 이의 개선 정도가 하나의 성과 지표가 된다.

② LDL-콜레스테롤 실측값과 추정값의 상관계수 혹은 R^2

- **Data Set 2 [Validation Data Set]**을 이용하여 세 가지 공식(Friedewald 공식, Martin 공식, Sampson 공식)의 "LDL-콜레스테롤 추정값"과 "LDL-콜레스테롤 실측값(**LDL-C**)"의 상관계수 혹은 이의 제곱값(설명력)을 각각 계산하여 비교한다.
- **Data Set 2 [Validation Data Set]**을 이용하여, 여러분이 개발한 "LDL-콜레스테롤 추정값"과 "LDL-콜레스테롤 실측값(**LDL-C**)"의 상관계수 혹은 이의 제곱값(설명력)이 얼마인지 계산하고, 이의 개선 정도가 하나의 성과 지표가 된다.
- <Table 3>은 Friedewald 공식의 "LDL-콜레스테롤 추정값"과 "LDL-콜레스테롤 실측값(**LDL-C**)"의 상관계수 혹은 이의 제곱값(설명력)을 계산한 것이다. 상관계수 = 0.968, 그리고 $\mathbf{R^2}$ = 0.937임을 알 수 있다.

- 여러분이 개발한 새로운 추정식이 이러한 두 가지 성과 지표에서, 세 가지 공식(Friedewald 공식, Martin 공식, Sampson 공식)보다 높은 스코어를 얻는다면 개선을 이룬 것이다
- 이러한 두 가지 성과 지표 이외에서, 여러분이 개발한 추정식의 성과를 제시할 수 있다면 자유롭게 제출하면 된다.

<Table 1>

CG_6_LDLC_Friedewald * CG_6_LDLC_direct 교차표[a]

			CG_6_LDLC_direct						
			1	2	3	4	5	6	전체
CG_6_LDLC_Friedewald	1	빈도	340	117	0	0	0	0	457
		CG_6_LDLC_Friedewald 중 %	74.4%	25.6%	0.0%	0.0%	0.0%	0.0%	100.0%
	2	빈도	79	1350	270	1	0	0	1700
		CG_6_LDLC_Friedewald 중 %	4.6%	79.4%	15.9%	0.1%	0.0%	0.0%	100.0%
	3	빈도	0	144	1580	181	0	0	1905
		CG_6_LDLC_Friedewald 중 %	0.0%	7.6%	82.9%	9.5%	0.0%	0.0%	100.0%
	4	빈도	0	0	132	817	68	0	1017
		CG_6_LDLC_Friedewald 중 %	0.0%	0.0%	13.0%	80.3%	6.7%	0.0%	100.0%
	5	빈도	0	0	0	61	253	17	331
		CG_6_LDLC_Friedewald 중 %	0.0%	0.0%	0.0%	18.4%	76.4%	5.1%	100.0%
	6	빈도	0	0	0	0	18	55	73
		CG_6_LDLC_Friedewald 중 %	0.0%	0.0%	0.0%	0.0%	24.7%	75.3%	100.0%
전체		빈도	419	1611	1982	1060	339	72	5483
		CG_6_LDLC_Friedewald 중 %	7.6%	29.4%	36.1%	19.3%	6.2%	1.3%	100.0%

<Table 2>

CG_8_TG_50_100_150_200_300_400_800 * YN_CG_6_D_Friedewald 교차표[a]

			YN_CG_6_D_Friedewald		
			0	1	전체
CG_8_TG_50_100_150_200_300_400_800	1	빈도	400	65	465
		CG_8_TG_50_100_150_200_300_400_800 중 %	86.0%	14.0%	100.0%
	50	빈도	1808	316	2124
		CG_8_TG_50_100_150_200_300_400_800 중 %	85.1%	14.9%	100.0%
	100	빈도	1225	283	1508
		CG_8_TG_50_100_150_200_300_400_800 중 %	81.2%	18.8%	100.0%
	150	빈도	557	167	724
		CG_8_TG_50_100_150_200_300_400_800 중 %	76.9%	23.1%	100.0%
	200	빈도	332	167	499
		CG_8_TG_50_100_150_200_300_400_800 중 %	66.5%	33.5%	100.0%
	300	빈도	73	90	163
		CG_8_TG_50_100_150_200_300_400_800 중 %	44.8%	55.2%	100.0%
전체		빈도	4395	1088	5483
		CG_8_TG_50_100_150_200_300_400_800 중 %	80.2%	19.8%	100.0%

<Table 3>

모형 요약[a]

					통계량 변화량				
모형	R	R 제곱	수정된 R 제곱	추정값의 표준 오차	R 제곱 변화량	F 변화량	자유도1	자유도2	유의확률 F 변화량
1	.968[b]	.937	.937	7.925	.937	81873.141	1	5481	.000

a. CG_3_year = 2

b. 예측자: (상수), E_LDLC_Friedewald