생명과학석사 학위논문

한국 성인 당뇨병 병형에 따른 최적 약물 제안 모델 개발

2021년 12월

서울대학교 대학원

생명과학부 시스템생물학 전공

신유례

한국 성인 당뇨병 병형에 따른 최적 약물 제안 모델 개발

지도 교수 황 대 희

이 논문을 생명과학석사 학위논문으로 제출함
2021년 11월

서울대학교 대학원
생명과학부 시스템생물학 연구실
신 유 례

신유례의 생명과학석사 학위논문을 인준함
2021년 12월

위 원 장 _____ 박주홍 _____ (인)

부위원장 _____ 황대희 _____ (인)

위 원 _____ 김종서 _____ (인)

초 록

당뇨는 고혈당을 증상으로 하는 대사 질환군으로, 현대 사회의 대표적인 만성 질환으로 꼽힌다. 당뇨병은 크게 제1형 당뇨병과 제2형 당뇨병으로 구분되는데, 제2형 당뇨병이 이질적인 환자군의 집합이라는 사실은 잘 알려져 있다. 따라서 제2형 당뇨병의 정교한 분류와 환자 맞춤형 치료법이 필요한 상황이다.

본 연구에서는 2004년부터 2020년까지 서울대병원 CDM에 수집된 데이터를 활용하여 제2형 당뇨병 코호트(n=55,602)를 구축하고, 진단, 임상 및 약물 정보를 수집하여 데이터 기반 당뇨 환자 재분류 및 약물 최적화를 시도했다. 환자 정보 기반 추정사구체여과율 예측 장단기 기억 순환 신경망 모델(R^2=0.8842)을 학습시켰고, 장단기 기억 임베딩 벡터에서 K-평균 군집화를 수행하여 7개의 제2형 당뇨 환자군을 분류했다.

분류된 환자군은 추정사구체여과율 기저 수치 및 당뇨 진행에 따른 추이가 구분되었으며, 단백뇨 및 심뇌혈관 발생에 있어서 유의한 차이를 보였다. 개인별 약물을 최적화했을 때 모델은 기존 연구 결과에 부합하는 약물을 선택하였으며, 만성 신장병의 비가역적인 특징을 잘 재현했다.

주요어 : 제2형 당뇨, 당뇨 합병증, 임상 정보, EMR, 약물 처방, 약물 제안 모델

학 번 : 2020-23101

목 차

표 목차

그림 목차

제1장 서론

제1절 연구의 배경

당뇨병은 고혈당을 증상으로 나타내는 다양한 대사 장애의 집합으로, 여러 개의 독립적인 발병 기전과 위험 요인을 가진다(Association, 2010). 2018년 기준 국내 30세 이상 성인 7명중 1명(13.8%)이 당뇨병을 갖고 있으며 65세 이상에서는 약 30%의 유병률을 보여, 고령의 인구가 증가하는 국내 실정에서 당뇨병은 보건사회적으로도 중요한 질환이다(Jung et al., 2021).

당뇨 환자는 크게 자가면역성 β-세포 파괴로 인해 발생하는 제1형 당뇨와 β-세포 분비 기능 이상으로 인해 발생하는 제2형 당뇨로 나뉘어지지만(Association, 2020), 전체 환자의 약 90%를 차지하는 제2형 당뇨 역시 이질적인 환자군의 집합이라는 사실은 잘 알려져 있다. 그러나 제2형 당뇨 환자를 구분하는 명확한 임상적 기준이 존재하지 않는 현실로 인해, 당뇨 환자 치료법의 결정과 처방은 대개 임상의의 판단에 의존한다.

당뇨 환자의 대표적인 임상 변수를 활용하여 환자를 클러스터링하고, 이를 기반으로 환자를 재분류하고자 하는 시도는 스웨덴의 Ahlqvist 연구팀을 시작으로 하여 중국, 미국, 독일 등 다양한 국가의 환자 코호트를 이용해 계속되어 왔다(Ahlqvist et al., 2018; Zaharia et al., 2019; Zou, Zhou, Zhu, & Ji, 2019). 그러나 이들 연구는 연구를 통해 새로이 정의한 분류군에 속하는 환자들이 치료 요법에 보이는 반응에서 차이가 나는지 검증하지 않았으며, 이러한 분류군이 치료 요법 선택에 실제로 유용한지 확인하지

않았다는 점에서 비판 받기도 했다(Dennis, Shields, Henley, Jones, & Hattersley, 2019).

또한 기존 연구는 당뇨 환자 진단에 유용하다고 알려진 임상 변수 가운데 대표적인 변수 6~20개를 선정하여 해당 변수에 대해서만 분석을 진행하였으며, Zou 연구팀을 제외한 나머지는 환자 개인에 대하여 이러한 임상 변수를 일정 기간에 걸쳐 여러 번 수집하지 않고 한 번만 수집했다. 약물 처방에 관한 데이터는 환자 분류에 사용되지 않았다. 당뇨병이 장기간에 걸친 약물 치료 및 혈당 관리를 동반하는 만성 질환이라는 점을 고려할 때, 기존 연구에서 수집한 임상 정보가 환자의 상태를 잘 대변한다고 보기 어려운 상황이다.

당뇨병은 환자의 식생활습관 및 약물 처방에 따라 예후가 크게 달라진다. 특히 혈당을 정상 수준으로 관리하는 것이 중요하다고 알려져 있으며, 만성적인 고혈당은 당뇨 관련 합병증 발생과 유의한 상관관계를 가진다(Nathan et al., 2009). 당뇨 합병증은 크게 대혈관 합병증 및 미세혈관 합병증으로 나뉘는데(Papatheodorou, Banach, Bekiari, Rizzo, & Edmonds, 2018), 심혈관계 질환 및 만성신질환(Chronic kidney disease)이 사망률에 크게 영향을 끼치는 대표적인 합병증으로 알려져 있다(Braunwald, 2019). 때문에 미국 당뇨병 협회(American Diabetes Association)와 NIH는 신장 합병증의 징후를 감시하기 위해 당뇨 환자에 대해 매년 1회 이상 추정사구체여과율(eGFR) 검사를 실시하도록 권고하고 있으며(Dabla, 2010), 미국 당뇨병 협회와 유럽 당뇨병 학회가 발표한 혈당강하제 처방 권고안 역시 약물을 선택할 때 환자의 신부전 위험을 고려하도록 하고 있다(Buse

et al., 2019).

추정사구체여과율은 나이, 성별, 당뇨병 유병기간, 흡연, 비만, 혈압, 혈당과 혈중 지질 등 다른 위험 인자와 독립적으로 신장 기능을 예측하는 인자다(Dabla, 2010). 미국신장재단(The Kidney Disease Outcomes Quality Initiative)에서 제공하는 지침에 따르면, 추정사구체여과율은 미세알부민 수치와 함께 만성 신질환 진단의 기준으로 사용되며 3개월 이상 추정사구체여과율이 저하된 경우 만성 신질환의 위험이 높은 것으로 보았다(Lok et al., 2020). 본 연구에서는 추정사구체여과율을 당뇨병성 신증 발생의 지표로 보고 추정사구체여과율 예측 모델을 개발하고자 하였다.

제2절 연구의 내용

본 연구는 서울대병원 본원에서 얻을 수 있는 임상 정보를 최대한 활용하여 당뇨 환자의 임상적 특성을 다각도에서 살펴볼 수 있는 데이터를 구축하고, 이를 활용하여 환자군에 따른 최적의 치료제 처방 모델을 개발하고자 했다. 이를 위해 2004년부터 2020년까지 서울대병원 CDM에 등록된 임상 정보로부터 제2형 당뇨 환자 코호트를 구축하고, 당뇨 환자의 임상 및 약물 처방 정보를 시계열로 수집했다.

환자 데이터는 선형 지도 학습과 비선형 지도 학습 두 가지 모델을 이용하여 분석했다. 부분 최소 제곱법(Partial least squares)은 다른 선형 회귀 모델에 비해 다중공선성 문제의 영향을 적게 받는다는 특징이 있기 때문에(Stoica & Söderström, 1998) 독립변수 간의 높은 선형관계가 존재하는 환자 데이터 분석에 적합할 것으로 보았다. 장단기 기억(Long short-term

memory) 순환 신경망(Recurrent neural network)은 장기간에 걸쳐 의존성을 보이는 시계열 데이터 분석에 뛰어난 성능을 보이는 것으로 알려져 있다(Beaufays, 2014). 본 연구에서는 다중 블록 부분 최소 제곱법과 다층 장단기 기억 순환 신경망을 이용하여 당뇨병 환자의 연도별 추적 데이터로부터 추정사구체여과율 회귀 모델을 구현하고, 모델을 기반으로 환자의 약물 반응성을 예측하고자 했다.

제2장 결과

제1절 추정사구체여과율 회귀 모델

2004년 이후 서울대병원 본원에서 1년에 2회 이상 당뇨로 진단받은 18세 이상의 환자 89,603명 가운데 최초 당화혈색소 수치가 6.5 미만인 환자 32,923명은 분석에서 제외되었다. 제1형 당뇨병으로 분류된 환자 566명(1.0%)과 LADA로 분류된 환자 371명(0.7%)을 제외하고, 제2형 당뇨병 환자 56,539명(98.3%)을 대상으로 회귀 모델을 학습시켰다(Table 1). 제2형 당뇨병 환자 가운데 내원 검진 기록이 존재하지 않는 환자는 860명(1.5%)이었으며, 이 환자들은 분석에서 제외되었다. 최초 당뇨 진단 당시 환자의 나이는 18세부터 103세 사이였고, 남성의 비율이 55.8%로 다소 높았다.

본 연구는 제2형 당뇨병 환자의 약물 처방 모델에 집중하고자 하므로, 제2형 당뇨병 환자가 최초 내원 이후 처방 받은 약물 정보를 수집하여 약제 성분에 따라 분류했다. 수집 대상 약물은 경구 혈당강하제(metformin, sulfonylurea, dipeptidyl peptidase-4 inhibitors, thiazolidinedione, sodium-glucose

cotransporter 2 inhibitors, α-glucosidase inhibitors, meglitinide), 주사 혈당강하제(glucagon-like peptide 1 receptor agonists, insulin), 혈압강하약제 (ACEi, ARB, alpha blockers, beta blockers, calcium-channel inhibitors, diuretics) 및 고지혈증 제제(statin, ezetimibe, fibrate/omega-3)로 성분 및 투여방법에 따라 총 18개 약물 클래스로 구분되었다. 2004년부터 2020년까지 서울대병원에 등록된 제2형 당뇨병 환자군에 대한 수집 대상 약물 처방 건수는 1,500,782건이었으며, 전체 제2형 당뇨병 환자군 가운데 92.8% (n=43,873)가 서울대병원에서 당뇨로 진단받은 이후 한 종류 이상의 약물을 처방 받았다(Table 1).

추정사구체여과율 예측 모델은 다중 블록 부분 최소 제곱법(MB-PLS)과 다층 장단기 기억 순환 신경망(LSTM-RNN)을 사용하여 학습되었다. 각 모델에서 환자 데이터는 임상과 약물의 두 가지 블록으로 나뉘어 입력되었으며, 종속 변수는 양적 변수로 주어졌다. 독립 변수는 모델의 특성을 고려하여 일부만 사용하였다(재료 및 방법 참조). 추정사구체여과율 예측 성능은 전반적으로 LSTM-RNN 모델이 MB-PLS 모델에 비해 뛰어났다(LSTM-RNN: mean squared error loss=0.0249; R^2=0.8842. MB-PLS: sum of squares residual=0.3879; R^2=0.7867).

MB-PLS 모델의 변수 중요도 상위 10개 변수는 혈중 크레아티닌, 철 결합 능력, 메트포르민(Metformin), 소변 단백/크레아티닌 비, 요단백, 혈중 칼륨, 혈중 이산화탄소, 적혈구 용적 백분율, 적혈구수, 요미세알부민으로 나타났다(Figure 1C). LSTM-RNN 모델의 변수 중요도 상위 10개 변수는 피브레이트/오메가3(Fibrate/Omega-3), 혈중 단백질, 알칼리인산분해효소,

혈중 인산염, 이뇨제(Diuretics), 메트포르민, 진단 당시 나이, 혈중 칼슘, 혈중 칼륨, 혈중 알부민으로 나타났다(Figure 1D).

임상 변수 가운데 적혈구 용적 백분율, 진단 당시 나이, 혈중 알부민, 혈중 칼륨은 추정사구체여과율 예측에 공통적으로 많이 기여하는 것으로 나타났다. 약물 변수 가운데 메트포르민은 두 예측 모델에서 공통적으로 기여도가 높게 나타났으나, 이뇨제와 피브레이트/오메가3는 LSTM-RNN 모델에서 상대적으로 더 중요한 변수로 나타났다.

MB-PLS 모델은 환자 데이터와 추정사구체여과율 사이의 비선형적 관계를 학습하는 데 있어서 상대적으로 취약했으며(Figure 1A) 혈중 크레아티닌 농도가 독립변수로 주어지지 않았을 때 추정사구체여과율을 거의 예측하지 못하는 한계가 있었다. LSTM-RNN 모델은 혈중 크레아티닌 농도가 독립변수로 주어지지 않았음에도 추정사구체여과율 예측에 있어서 MB-PLS 모델보다 더 나은 성능을 보였다. 또한, 추정사구체여과율에 직접적으로 영향을 미칠 것으로 예상되는 이뇨제가 LSTM-RNN 모델에서 더 높은 변수 중요도를 가졌다.

제2절 LSTM 기반 환자군 분류

LSTM 블록의 임상 정보 임베딩 벡터(embedding vector)는 LSTM 블록의 최종 산출 벡터로써, 다년간의 임상 정보 추적 데이터를 요약하여 나타낸다고 볼 수 있다. 따라서 임베딩 벡터를 활용하여 전체 코호트를 추정사구체여과율 수치 추이에 있어서 구분되는 환자군으로 분류할 수 있을 것으로 보고, 최초 진단 후 8년간 추정사구체여과율 추적 데이터가

존재하는 환자 15,037명의 임상 정보 임베딩 벡터를 이용하여 K-평균 군집화(Likas, Vlassis, & J. Verbeek, 2003)를 수행했다(k=10). 이 중 UMAP 상의 위치 관계 및 추정사구체여과율 추적 데이터 추이가 유사한 군집을 병합하여 전체 환자를 7개의 환자군으로 분류했다(Figure 2, Figure 3).

환자군 0번은 대상 환자 15,037명 중 156명(1.04%)의 환자를 포함하며, 모든 환자군 가운데 가장 낮은 추정사구체여과율 추이를 보였다(최초 진단 시 45.75 ± 22.36; 8년차 추적 시 13.33 ± 14.83; 평균 ± 표준편차로 표기함). 또한 다른 환자군에 비해 최초 진단 시 연령과 HOMA2-IR이 낮은 경향이 있었다(Figure 4). 환자군 1번은 대상 환자 중 541명(3.60%)의 환자를 포함하며, 최초 진단 시 추정사구체여과율 수치가 낮고(59.41 ± 25.11) 점차 악화되는 경향을 보였다(8년차 추적 시 29.90 ± 18.28). 환자군 2번은 대상 환자 중 917명(6.10%)을 포함하며, 최초 진단 시 추정사구체여과율 수치가 낮지만(63.24 ± 22.06) 상대적으로 수치가 잘 유지되었다(8년차 추적 시 58.09 ± 22.54). 환자군 3번은 대상 환자 중 1,658명(11.03%)을 포함하며, 최초 진단 시 추정사구체여과율 수치가 양호하나(79.72 ± 19.46) 시간이 지남에 따라 점차 신장 기능이 떨어지는 것으로 보였다(8년차 추적 시 57.77 ± 19.05). 환자군 4번과 5번은 다른 환자군에 비해 신장 기능이 양호하며, 잘 유지되는 경향을 보였다. 환자군 6번은 대상 환자 중 1,344명(8.94%)을 차지하며, 최초 진단 시 추정사구체여과율 수치는 다소 낮지만 시간이 지남에 따라 신장 기능이 점차 회복되는 것으로 나타났다.

Figure 3B는 각 환자군의 예측된 추정사구체여과율 수치를 나타낸다.

최초 진단 후 1년 까지의 예측 값은 60에 가까운 수치로 다소 왜곡되는 경향을 보였으나, 이러한 경향은 2년차 추적부터 거의 사라졌다.

각 환자군의 기저 특성(baseline characteristics)은 Figure 4에 요약되어 있다. 환자군에 따른 연령 분포는 크게 차이 나지 않았으나, 추정사구체여과율 수치가 떨어진 환자군에서 혈중 인산염이 다소 높은 경향을 보였으며 혈중 칼슘은 다소 낮았다. 적혈구용적률 분포는 환자군에 따라 큰 차이를 보였는데, 이는 신장 기능 저하로 인한 적혈구생성인자(erythropoietin) 생산 저하의 결과로 보인다(Zachée, Vermylen, & Boogaerts, 1994). HOMA2-IR은 환자군 0번에서 다소 떨어졌고 환자군 1번에서 다소 높았다.

본 연구에서 분류한 환자군이 실제 합병증 발생 빈도에 있어서 차이를 보이는지 검증하기 위하여 만성 신질환 및 심혈관계 질환 발생 분석을 진행했다(Figure 5). 만성 신질환(chronic kidney disease)은 사구체여과율 또는 미세알부민 수치에 따라 진단, 분류된다("Chapter 1: Definition and classification of CKD," 2013). 본 연구는 환자 분류 모델 구축에 추정사구체여과율을 사용했기 때문에, 공정한 평가를 위해 만성 신질환 발생 분석에 미세알부민 기준을 사용했다.

환자군 0번은 미세단백뇨(HR=3.144 (2.461-4.017); p-value=5.19E-20; modified by age, gender, and mean HbA1c)를 제외한 단백뇨(HR=17.45 (13.27-22.94); p-value=4.909E-93), 관상동맥질환(HR=2.277 (1.752-2.959); p-value=7.61E-10), 뇌혈관질환(HR=3.608 (2.634-4.941); p-value=1.27E-15) 발생 위험비가 모든 환자군 중 가장 높았다. 그러나 미세단백뇨와 단백뇨에서

환자의 기저 특성이 합병증 발생의 주요한 변인으로 나타났다. 기저 미세알부민 수치를 공변량으로 추가했을 때 미세단백뇨 위험비는 0.1148 (0.04345-0.3035)까지, 단백뇨 위험비는 1.859 (0.634-5.449)까지 떨어졌다(Table 3).

환자군 1번은 기저 미세알부민 수치를 공변량으로 추가했을 때 가장 높은 미세단백뇨(HR=3.292 (2.709-3.999); p-value=3.856E-33) 및 단백뇨(HR=8.355 (6.45-10.82); p-value=3.376E-58) 발생 위험비를 보였다. 환자군 2번은 단백뇨(HR=1.369 (1.135-1.651); p-value=0.001001) 발생 위험비가 상대적으로 낮았으나 미세단백뇨(HR=2.326 (1.71-3.162); p-value=7.31E-08) 발생 위험비는 다소 높았으며, 환자군 3번도 비슷한 경향을 보였다. 환자군 4번과 6번은 5번 대비 미세단백뇨 및 단백뇨 위험비가 비슷하거나 더 낮았다.

제3절 최적 약물 제안 모델

기존 약물 데이터에 존재하는 약물 처방 조합을 모두 추출하여 조합 목록을 만들고, 이를 기존 임상 데이터와 함께 모델에 입력하여 가능한 약물 조합 가운데 추정사구체여과율을 최대화하는 조합을 찾도록 했다. 약물 성분 개수를 점진적으로 늘리도록 하는 미국당뇨병협회 권고안을 참고하여 환자가 처방 받는 성분 개수는 유지되도록 제한하였다(Buse et al., 2019).

모델에 의해 최적화된 약물과 최적 약물을 처방했을 때 예상되는 추정사구체여과율 수치는 Figure 6에 요약되어 있다. 한 예로 2007년부터

2008년까지 Figure 6A 환자의 약물 처방을 최적화했을 때, LSTM-RNN 모델은 실제로 처방된 당뇨 약물 가운데 설포닐우레아(Sulfonylurea)를 메트포르민으로 대체할 것을 제안했다. 2007년 이 환자의 추정사구체여과율은 만성 신장병 3A 단계의 기준인 60ml/min/1.73m^2 미만으로 감소했으나, 메트포르민을 처방했을 때에는 기준선 이상으로 유지될 것으로 예측했다. 이는 설포닐우레아 단일 처방을 받은 환자가 메트포르민 단일 처방을 받은 환자에 비해 상대적으로 큰 추정사구체여과율 감소를 보였다는 기존 연구 결과와 일치한다(Hung et al., 2012).

Figure 6B의 환자는 당뇨 진단 초기부터 신장병이 상당히 진행된 환자로, 약물 최적화를 거친 뒤에도 추정사구체여과율이 만성 신장병 3B 단계의 기준인 45ml/min/1.73m^2 이상으로 회복되지 않았다. 신장병의 진행 단계에 따라 비가역적인 신장 기능 손실을 발생하는 특징을 모델이 잘 반영하고 있음을 알 수 있다.

제3장 토의

종합적으로, 본 연구를 통해 개발한 약물 제안 모델은 기존의 임상적 지식에 부합하는 환자 개인별 약물 최적화를 가능하게 했다. 이는 데이터를 기반으로 한 약물 제안이라는 기존의 임상적 요구에 대한 하나의 해결책이 될 수 있을 것이다. 기존의 데이터 기반 당뇨 정밀 의학 연구와 달리, 본 연구는 서울대병원 CDM 데이터를 활용하여 최대 17개년의 장기간 추적이 이루어진 한국인 제2형 당뇨 코호트를 정의하고, 임상 검사와 약물 처방을

포함하여 환자의 상태를 다각도에서 살펴볼 수 있는 데이터를 구축했다. 본 연구에서 사용한 모델은 OMOP CDM을 사용하는 다른 기관에서도 비교적 쉽게 재현이 가능할 것으로 예상된다.

본 연구에서는 지도 학습 모델로 다중 블록 부분 최소 제곱법과 장단기 기억 순환 신경망을 사용했다. 두 예측 모델에서 변수 중요도 상위 20개 변수 가운데 6개 변수가 약물 정보에 해당했고, 특히 장단기 기억 순환 신경망 모델에서 약물 정보가 변수 중요도 최상위를 차지했다. 이는 추정사구체여과율 예측에 있어서, 나아가 만성 질환 연구에 있어서 약물 정보가 임상 정보 이상으로 중요한 예측 인자임을 시사한다.

연구를 통해 찾은 환자군에 대한 병인학적인 분석은 이루어지지 않았다. 가능한 포괄적으로 환자의 임상적 상황을 반영하는 데이터를 구축하고자 했으나, 당뇨병의 두 가지 중요한 인자인 유전체와 식생활습관에 관한 데이터는 연구에 포함되지 않았다. 후속 연구를 통해 이에 대한 보완이 이루어진다면, 환자군의 추정사구체여과율 추이가 구분되는 원인을 보다 명확히 규명할 수 있을 것으로 기대된다.

제4장 재료 및 방법

제1절 임상 데이터 수집

2004년부터 2020년까지 서울대병원 본원 CDM에 등록된 데이터로부터 환자의 당뇨 진단 기록 및 당화혈색소(HbA1c), GAD(Glutamic Acid Decarboxylase) 자가항체, C-peptide 검진 기록을 추출하여 제2형 당뇨병에

대한 조작적 정의를 내렸다(Figure 7). 대상 기간 내에 제2형 당뇨병 및 당뇨 관련 질환으로 1년에 2회 이상 진단받았으며 당화혈색소 수치가 6.5 이상이었던 18세 이상의 환자를 제2형 당뇨병 환자로 정의했으며, 이 중 GAD 자가항체 수치가 1.0 이상인 환자와 모든 C-peptide 검진 수치가 0.6 미만인 환자는 LADA로 분류하여 제거했다. 또한 제2형 당뇨병 진단 없이 제1형 당뇨병 또는 당뇨 관련 질환으로만 진단받은 환자는 제1형 당뇨병으로 분류하여 제거했다. 등록된 데이터를 이용하여 분류할 수 없는 환자는 연구에서 제외시켰다.

서울대병원 본원 CDM에 존재하는 임상 변수 가운데 당뇨병 환자에 대한 기록이 존재하는 모든 검진을 검토하여 105개의 임상 변수를 선정하고(Table 4) 당뇨병 환자의 내원 기록을 수집했다. 전체 제2형 당뇨병 환자 55,602명 가운데 54,742명이 위 105개 검사 중 하나 이상에 대한 내원 기록을 가지고 있었다. 응급 및 입원 기록은 내원 환자의 상태를 추적하고자 하는 연구 목적에 적합하지 않기 때문에 제외하였다.

제2절 약물 데이터 수집

본 연구에서는 당뇨 환자의 혈당 수치 및 신장 기능 관리를 위해 처방하는 약물 정보를 수집하고 임상 정보와 결합함으로써 장기적인 치료 관점에서 환자의 혈당 및 합병증 관리 수준을 파악하고자 했다. 이를 위해 서울대병원 본원 내분비내과에서 당뇨 환자에게 처방하는 혈당강하제, 혈압강하약제 및 항고지혈증제 성분을 수집 대상 성분으로 정의하고(Table 5) 당뇨병 환자의 약물 처방 기록을 수집했다. 임상 정보와 마찬가지로,

응급 및 입원 기록은 내원 환자의 상태를 추적하고자 하는 연구 목적에 적합하지 않기 때문에 제외하였다. 약물 처방 기록은 각 환자의 당뇨 진단 내역과 결합하여 환자가 내원하였으나 약물을 처방 받지 않은 날짜에 대한 정보를 보완했다. 과거 약물 처방 이력이 있는 당뇨병 환자가 180일 이상의 기간 동안 2회 이상 내원하여 한 번도 약물을 처방 받지 않았을 경우 이를 약물 중단 기록으로 분류했으며, 위 조건을 충족하지 않는 경우 단순 내원으로 보았다.

제3절 분석 데이터 구축

HOMA 지수와 BMI는 당뇨병에 있어서 임상적으로 중요하다고 알려져 있으나 CDM에 등록되어 있지 않다. 분석 데이터 구축 시 두 임상 변수는 별도로 계산하여 반영하였다. HOMA 지수는 공복혈당 수치 및 C-peptide 수치로부터 계산했다. 추정사구체여과율은 CDM에 등록되어 있으나, 도입 시기가 혈중 크레아티닌 보다 늦기 때문에 보다 많은 데이터를 확보하기 위하여 혈중 크레아티닌 수치로부터 계산하여 사용했다.

각 환자의 임상 데이터와 약물 데이터를 방문일자 기준으로 병합하고 연도별로 중앙값을 취하여 분석 데이터를 구축했다. 독립 변수 별로 상위 및 하위 5%의 값은 이상치로 간주하여 결측 값으로 치환했다. 각 독립 변수는 분위수 정규화(quantile normalization) 기법으로 연도별 분포를 정규화했다(Bolstad, Irizarry, Astrand, & Speed, 2003).

제4절 다중 블록 부분 최소 제곱법

NIPALS 알고리즘 기반(Wold, 1966)의 다중 블록 부분 최소 제곱법(Westerhuis, Kourti, & MacGregor, 1998)을 이용하여 임상 데이터와 약물 데이터를 각각 입력 블록으로 받는 회귀 모델을 구축했다. 전체 데이터 가운데 종속변수가 결측 값이거나 독립변수 중 85% 이상이 결측 값인 데이터는 분석에서 제외했다. 결측 값은 치환하지 않고 행렬 계산 시 결측 값이 포함된 계산은 결과에서 제외하도록 했다. 각 독립 변수가 전체 모델에 기여하는 정도는 변수중요도척도(Variable importance in projection, VIP)를 통하여 나타냈다(Galindo-Prieto, Eriksson, & Trygg, 2014).

제5절 다층 장단기 기억 순환 신경망 모델

다중 블록 부분 최소 제곱법과 달리 결측 값을 치환할 필요성이 있기 때문에, 90% 이상의 레코드에서 결측 값인 변수는 분석에서 제외하였다. 단, 임상적으로 중요하다고 알려진 일부 변수(microalbumin to creatinine ratio, blood glucose PP2, c-peptide, insulin)는 분석에 포함하였다. 종속 변수의 계산에 직접 사용된 변수(blood creatinine) 및 동등하다고 여겨지는 변수는 분석에서 제외하였다.

독립 변수가 결측 값인 경우, 직후 년도 및 직전 년도의 값이 존재하면 이를 이용해 보간했다. 보간을 마친 데이터에서 5년 이상의 추적 데이터가 존재하는 환자를 분석 대상으로 삼았다. 종속 변수는 보간하지 않고 직전 년도의 종속 변수 값을 독립 변수로 추가했다. 모든 변수는 최소-최대 정규화하고 결측 값을 0으로 치환했다.

PyTorch(Paszke et al., 2019) 라이브러리가 제공하는 장단기 기억 모듈을 이용하여 임상 데이터와 약물 데이터를 입력으로 받는 다층 장단기 기억 블록을 각각 만들고, 이로부터 다대다 순환 신경망 모델을 구축했다. 임상 및 약물 장단기 기억 블록의 출력 벡터는 하나로 병합하여 통상적인 다층 퍼셉트론(Multilayer perceptron)을 거쳐 추정사구체여과율 회귀값을 출력하도록 구성했다(Figure 8). 각 독립 변수가 예측에 끼치는 영향은 통합 그래디언트(Integrated gradients, IG) 방법으로 계산했다(Sundararajan, Taly, & Yan, 2017).

제6절 합병증 발생 분석

통계 분석은 R 3.6.1, survival(ver. 3.2.3), survminer(ver. 0.4.9) 패키지를 이용하여 진행했다. Cox-회귀 모형에서 HR은 cluster 5에 대하여 계산되었다.

참고 문헌

Ahlqvist, E., Storm, P., Käräjämäki, A., Martinell, M., Dorkhan, M., Carlsson, A., . . . Groop, L. (2018). Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. *The Lancet Diabetes & Endocrinology, 6*(5), 361-369. doi:10.1016/S2213-8587(18)30051-2

Association, A. D. (2010). Diagnosis and Classification of Diabetes Mellitus. *Diabetes Care, 33*(Supplement 1), S62-S69. doi:10.2337/dc10-S062

Association, A. D. (2020). 2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2020. *Diabetes Care, 43*(Supplement 1), S14-S31. doi:10.2337/dc20-S002

Beaufays, H. S. a. A. W. S. a. F. (2014). Long short-term memory recurrent neural network architectures for large scale acoustic modeling. *INTERSPEECH*, 338-342.

Bolstad, B. M., Irizarry, R. A., Astrand, M., & Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics, 19*(2), 185-193. doi:10.1093/bioinformatics/19.2.185

Braunwald, E. (2019). Diabetes, heart failure, and renal dysfunction: The vicious circles. *Progress in Cardiovascular Diseases, 62*(4), 298-302. doi:https://doi.org/10.1016/j.pcad.2019.07.003

Buse, J. B., Wexler, D. J., Tsapas, A., Rossing, P., Mingrone, G., Mathieu, C., . . . Davies, M. J. (2019). 2019 Update to: Management of Hyperglycemia in Type 2 Diabetes, 2018. A Consensus Report by the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD). *Diabetes Care*, dci190066. doi:10.2337/dci19-0066

Chapter 1: Definition and classification of CKD. (2013). *Kidney international supplements, 3*(1), 19-62. doi:10.1038/kisup.2012.64

Dabla, P. K. (2010). Renal function in diabetic nephropathy. *World journal of diabetes, 1*(2), 48-56. doi:10.4239/wjd.v1.i2.48

Dennis, J. M., Shields, B. M., Henley, W. E., Jones, A. G., & Hattersley, A. T. (2019). Disease progression and treatment response in data-driven subgroups of type 2 diabetes compared with models based on simple clinical features: an analysis using clinical trial data. *The Lancet Diabetes & Endocrinology, 7*(6), 442-451. doi:https://doi.org/10.1016/S2213-8587(19)30087-7

Galindo-Prieto, B., Eriksson, L., & Trygg, J. (2014). Variable influence on projection (VIP) for orthogonal projections to latent structures (OPLS). *Journal of Chemometrics, 28*. doi:10.1002/cem.2627

Hung, A. M., Roumie, C. L., Greevy, R. A., Liu, X., Grijalva, C. G., Murff, H. J., . . . Griffin, M. R. (2012). Comparative effectiveness of incident oral antidiabetic drugs on kidney function. *Kidney International,*

81(7), 698-706. doi:https://doi.org/10.1038/ki.2011.444

Jung, C.-H., Son, J. W., Kang, S., Kim, W. J., Kim, H.-S., Kim, H. S., . . . Yoon, K. H. (2021). Diabetes Fact Sheets in Korea, 2020: An Appraisal of Current Status. *Korean Diabetes J, 0*. doi:10.4093/dmj.2020.0254

Likas, A., Vlassis, N., & J. Verbeek, J. (2003). The global k-means clustering algorithm. *Pattern Recognition, 36*(2), 451-461. doi:https://doi.org/10.1016/S0031-3203(02)00060-2

Lok, C. E., Huber, T. S., Lee, T., Shenoy, S., Yevzlin, A. S., Abreo, K., . . . Valentini, R. P. (2020). KDOQI Clinical Practice Guideline for Vascular Access: 2019 Update. *American Journal of Kidney Diseases, 75*(4, Supplement 2), S1-S164. doi:https://doi.org/10.1053/j.ajkd.2019.12.001

Nathan, D. M., Buse, J. B., Davidson, M. B., Ferrannini, E., Holman, R. R., Sherwin, R., & Zinman, B. (2009). Medical Management of Hyperglycemia in Type 2 Diabetes: A Consensus Algorithm for the Initiation and Adjustment of Therapy. *A consensus statement of the American Diabetes Association and the European Association for the Study of Diabetes, 32*(1), 193-203. doi:10.2337/dc08-9025

Papatheodorou, K., Banach, M., Bekiari, E., Rizzo, M., & Edmonds, M. (2018). Complications of Diabetes 2017. *Journal of Diabetes Research, 2018*, 3086167. doi:10.1155/2018/3086167

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., . . . Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. arXiv:1912.01703. Retrieved from https://ui.adsabs.harvard.edu/abs/2019arXiv191201703P

Stoica, P., & Söderström, T. (1998). Partial Least Squares: A First-order Analysis. *Scandinavian Journal of Statistics, 25*(1), 17-24. doi:https://doi.org/10.1111/1467-9469.00085

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic Attribution for Deep Networks. arXiv:1703.01365. Retrieved from https://ui.adsabs.harvard.edu/abs/2017arXiv170301365S

Westerhuis, J. A., Kourti, T., & MacGregor, J. F. (1998). Analysis of multiblock and hierarchical PCA and PLS models. *Journal of Chemometrics, 12*(5), 301-321. doi:https://doi.org/10.1002/(SICI)1099-128X(199809/10)12:5<301::AID-CEM515>3.0.CO;2-S

Wold, H. (1966). Estimation of principal components and related models by iterative least squares. *Multivariate analysis*, 391-420. Retrieved from https://ci.nii.ac.jp/naid/20001378860/en/

Zachée, P., Vermylen, J., & Boogaerts, M. A. (1994). Hematologic aspects of end-stage renal failure. *Annals of Hematology, 69*(1), 33-40. doi:10.1007/BF01757345

Zaharia, O. P., Strassburger, K., Strom, A., Bönhof, G. J., Karusheva, Y., Antoniou, S., . . . Ziegler, D. (2019). Risk of diabetes-associated diseases in subgroups of patients with recent-onset diabetes: a 5-

year follow-up study. *The Lancet Diabetes & Endocrinology, 7*(9), 684-694. doi:https://doi.org/10.1016/S2213-8587(19)30187-1

Zou, X., Zhou, X., Zhu, Z., & Ji, L. (2019). Novel subgroups of patients with adult-onset diabetes in Chinese and US populations. *The Lancet Diabetes & Endocrinology, 7*(1), 9-11. doi:https://doi.org/10.1016/S2213-8587(18)30316-4

Abstract

Development of Data-driven Optimization Model for Treatment in Korean Diabetic Patients

Yourae Shin
Biological Sciences, System biology and medicine
The Graduate School
Seoul National University

Diabetes is a group of metabolic disorders characterized by high blood sugar as its symptom, considered as a representative chronic disease of modern society. Diabetes is broadly classified into two groups, type 1 and type 2 diabetes. Type 2 diabetes accounts for 90% of all diabetic patients and is highly heterogeneous.

In this study, I developed a data-driven optimization model for treatment in Korean diabetic patients (n=55,602) using Seoul National University Hospital CDM data, collected from 2004 to 2020. The model used a long-short term memory recurrent neural network and was trained to predict eGFR in type 2 diabetic patient data(R^2=0.8842).

I identified 7 subgroups in type 2 diabetes, which were distinct in baseline eGFR, a longitudinal transition of eGFR, and renal and development of complications. The optimization model selected reasonable treatment and reproduced the irreversible nature of chronic kidney disease.

Keywords : (6단어 이내) Type 2 diabetes, Diabetic complications, Clinical data, EMR, Optimized treatment, Data-driven optimization model

Student Number : 2020-23101

Table 1 Baseline characteristics in SNUH.

Variables	Type 2 Diabetes
N	55,602 (54,742)
Male, n (%)	30600 (55.8)
Age at first visit, years	61.1 ± 11.7
Body mass index, kg/m2	25.3 ± 4.22
HbA1c, %	7.95 ± 1.48
HbA1c, mmol/L	63.4 ± 16.2
C-peptide	2.43 ± 1.75
HOMA2-%B (by C-peptdie)	54.7 ± 38.5
HOMA2-IR	2.03 ± 2.67
Creatinine, mg/dL	1.08 ± 0.87
eGFR, mL/min/1.73m2	75.2 ± 24.6
Total cholesterol, mg/dL	183 ± 46.2
Triglyceride, mg/dL	165 ± 140
LDL cholesterol, mg/dL	107 ± 39.5
Spot urine albumin/creatinine, mg/g	0.219 ± 0.839

Table 2 Drug exposure in SNUH. DPP4i, Dipeptidyl Peptidase-4 Inhibitors; TZD, Thiazolidinedione; SGLT2i, Sodium-glucose Cotransporter 2 Inhibitors; α-GI, α-Glucosidase Inhibitors; GLP1a, Glucagon-Like Peptide 1 Receptor Agonists.

		Number of records	**Number of patients**
Total		1500782 (100.0)	43873 (100.0)
Oral Hypoglycemic Agents	Metformin (%)	713581 (62.6)	27784 (63.3)
	Sulfonylurea (%)	500836 (43.9)	20479 (46.7)
	DPP4i (%)	406696 (42.7)	19021 (43.4)
	TZD (%)	53619 (4.7)	3417 (7.8)
	SGLT2i (%)	24134 (5.1)	2835 (6.5)
	α-GI (%)	57297 (5.0)	3413 (7.8)
	Meglitinide (%)	24504 (2.1)	1245 (2.8)
Injectable Hypoglycemic Agents	GLP1a (%)	8967 (1.0)	649 (1.5)
	Insulin (basal) (%)	124792 (10.9)	3386 (7.7)
	Insulin (bolus) (%)	25378 (2.2)	789 (1.8)
	Insulin (premixed) (%)	140042 (12.3)	5170 (11.8)
Antihypertensive Agents	ACEi, ARB (%)	671665 (55.9)	20418 (46.5)
	Alpha Blockers (%)	250416 (20.8)	11377 (25.9)
	Beta Blockers (%)	37001 (3.1)	2448 (5.6)
	Calcium Channel Blockers (%)	438852 (36.5)	16020 (36.5)
	Diuretics (%)	290983 (24.2)	11806 (26.9)
Antihyperlipidemic Agents	Ezetimibe (%)	795003 (66.2)	26762 (61.0)
	Fibrate/Omega-3 (%)	65530 (5.5)	4099 (9.3)
	Statin (%)	58101 (4.8)	3470 (7.9)

Table 3 Cox regression analysis comparing hazard ratio (HR) of chronic kidney disease.

		Covariates = age + gender + mean HbA1c					Covariates = age + gender + mean HbA1c + baseline ACR		
	Cluster	N	Event	Event(%)	HR	p-value	N	Event	Event(%)
CKD A2	Cluster 0	5720	2065	36.10%	3.144 (2.461-4.017)	5.19E-20	3645	1428	39.18%
	Cluster 1	6046	2322	38.41%	3.358 (2.975-3.79)	1.59E-85	3821	1578	41.30%
	Cluster 2	6294	2338	37.15%	1.509 (1.343-1.694)	3.85E-12	3877	1532	39.52%
	Cluster 3	7055	2759	39.11%	1.605 (1.474-1.747)	1.43E-27	4429	1861	42.02%
	Cluster 4	9149	3404	37.21%	0.9558 (0.8923-1.024)	0.1972	4812	1865	38.76%
	Cluster 5	-	-	-	1	-	-	-	-
	Cluster 6	6734	2373	35.24%	0.9141 (0.8183-1.021)	0.1115	4192	1595	38.05%
CKD A3	Cluster 0	5823	516	8.86%	17.45 (13.27-22.94)	4.91E-93	3736	341	9.13%
	Cluster 1	6158	701	11.38%	11.98 (10.16-14.12)	1.85E-101	3919	444	11.33%
	Cluster 2	6431	631	9.81%	3.703 (3.105-4.417)	4.82E-48	3997	366	9.16%
	Cluster 3	7193	787	10.94%	2.93 (2.537-3.384)	2.22E-48	4552	517	11.36%
	Cluster 4	9305	847	9.10%	1.169 (1.021-1.34)	0.02389	4945	429	8.68%
	Cluster 5	-	-	-	1	-	-	-	-
	Cluster 6	6858	561	8.18%	1.263 (1.023-1.561)	0.03021	4302	362	8.42%

HR	p-value
0.1148 (0.04345-0.3035)	1.27E-05
3.292 (2.709-3.999)	3.86E-33
1.369 (1.135-1.651)	0.001001
1.458 (1.305-1.628)	2.41E-11
0.8666 (0.7788-0.9643)	0.008597
1	-
0.799 (0.686-0.9305)	0.003907
1.859 (0.634-5.449)	0.2587
8.355 (6.45-10.82)	3.38E-58
2.326 (1.71-3.162)	7.31E-08
2.49 (2.06-3.01)	4.43E-21
0.9643 (0.7768-1.197)	0.742
1	-
1.014 (0.7449-1.381)	0.9284

Table 4 Demographics, laboratory tests, physical measurements, and other measurements considered in this study.

Category	Variables
Demographics (2)	Age, gender
Measurements (105)	
Blood (72)	alanine aminotranferase, albumin, alkaline phosphatase, alpha-1-fetoprotein, apolipoprotein A-I, apolipoprotein B, aPTT, aspartate aminotransferase, band form neutrophils, basophils, bilirubin, bilirubin, blasts, blood glucose PP2, calcium, chloride, cholesterol, CO2, cobalamin, c-peptide, creatinine, creatinine kinase, eosinophils, erythrocytes, estimated glomerular filtration rate, fasting glucose, ferritin, fibrinogen, folate, gamma glutamyl transferase, glomerular filtration rate (EPI), glucose, HbA1c, HDL, Hematocrit, hepatitis B virus surface antibody, hsCRP, immature cells, immature monocytes, INR, insulin, ionized calcium, iron, iron capacity, lactate dehydrogenase, large unstained cells, LDL, leukocytes, lymphocytes, magnesium, MCH, MCHC, MCV, metamyelocytes, monocytes, neutrophils, Normoblast, phosphate, plateletocrit, platelets, potassium, promyelocytes, protein, segmented neutrophils, sodium, thyrotropin, thyroxine, triglyceride, triiodothyronine, troponin I, vancomycin, variant lymphocytes
Urine (18)	chloride, creatinine (24hrs), creatinine (rand), creatinine clearance, dysmorphic erythrocytes, microalbumin (24hrs), microalbumin (rand), microalbumin to creatinine ratio (24hrs), microalbumin to creatinine ratio (rand), potassium (24hrs), potassium (rand), prealbumin, protein (24hrs), protein (rand), protein to creatinine ratio (24hrs), protein to creatinine ratio (rand), sodium (24hrs), sodium (rand)
Body fluid (3)	chloride, erythrocytes, glucose
Physical measurements (8)	body fat mass, body fat percentage, body height, body muscle mass, body weight, hip circumference, visceral fat, waist circumference
Other measurements (4)	CO2 (partial pressure), diastolic blood pressure, pancreatic elastase, systolic blood pressure

Table 5 Drugs considered in this study.

Category	Drug Class	Drug
Oral Hypoglycemic Agents	Metformin	Metformin
	Sulfonylurea	Gliclazide, Glimepiride, Glipizide, Gliquidone
	Dipeptidyl Peptidase-4 Inhibitors	Alogliptin, Anagliptin, Evogliptin, Gemigliptin, Linagliptin, Saxagliptin, Sitagliptin, Teneligliptin, Vildagliptin
	Thiazolidinedione	Lobeglitazone, Pioglitazone
	Sodium-glucose Cotransporter 2 Inhibitors	Dapagliflozin, Empagliflozin, Ipragliflozin
	α-Glucosidase Inhibitors	Acarbose, Miglitol, Voglibose
	Meglitinide	Nateglinide, Repaglinide
	Glucagon-Like Peptide 1 Receptor Agonists	Exenatide, Lixisenatide, Liraglutide, Dulaglutide
Injectable Hypoglycemic Agents	Insulin (basal)	Insulin degludec, Insulin detemir, Insulin glargine, Insulin human (NPH)
	Insulin (bolus)	Insulin lispo, Insulin aspart, Insulin glulisine, Insulin human (regular)
	Insulin (premixed)	Premixed versions of basal and bolus insulin
Antihypertensive Agents	ACEi, ARB	Candesartan, Captopril, Cilazapril, Enalapril, Eprosartan, Fimasartan, Imidapril, Irbesartan, Losartan, Olmesartan, Perindopril, Ramipril, Telmisartan, Valsartan, Zofenopril
	Alpha Blockers	Alfuzosin, Clonidine, Doxazosin, Phenoxybenzamine, Silodosin, Tamsulosin, Terazosin
	Beta Blockers	Atenolol, Betaxolol, Bisoprolol, Carteolol, Carvedilol, Esmolol, Labetalol, Metoprolol, Nadolol, Nebivolol, Propranolol, Sotalol, Timolol
	Calcium Channel Blockers	Amlodipine, Barnidipine, Diltiazem, Felodipine, Hydralazine, Isradipine, Lacidipine, Lercanidipine, Manidipine, Nicardipine, Nifedipine, Nimodipine, Nisoldipine, Verapamil

	Diuretics	Acetazolamide, Amiloride, Furosemide, Hydrochlorothiazide, Indapamide, Spironolactone, Torasemide
Antihyperlipidemic Agents	Ezetimibe	Ezetimibe
	Fibrate/Omega-3	Bezafibrate, Fenofibrate, Omega-3
	Statin	Atorvastatin, Fluvastatin, Lovastatin, Pitavastatin, Pravastatin, Rosuvastatin, Simvastatin

Table 6 Demographics, laboratory tests, other measurements, and calculated indices considered in RNN model.

Category	Variables
Demographics (2)	Age, gender
Measurements (47)	
Serum or Plasma (38)	alkaline phosphatase, alanine aminotranferase, aspartate aminotransferase, basophils, albumin, bilirubin, calcium, chloride, glucose, blood glucose PP2, phosphate, potassium, protein, sodium, triglyceride, cholesterol, CO2, c-peptide, eosinophils, erythrocytes, fasting glucose, gamma glutamyl transferase, HbA1c, HDL, Hematocrit, hsCRP, insulin, LDL, leukocytes, lymphocytes, MCH, MCHC, MCV, monocytes, neutrophils, plateletocrit, platelets, segmented neutrophils
Urine (3)	creatinine (rand), microalbumin (rand), microalbumin to creatinine ratio (rand)
Other measurements (2)	diastolic blood pressure, systolic blood pressure
Calculated indices (4)	body mass index, HOMA2 %B, HOMA2 %S, HOMA2 insulin resistance

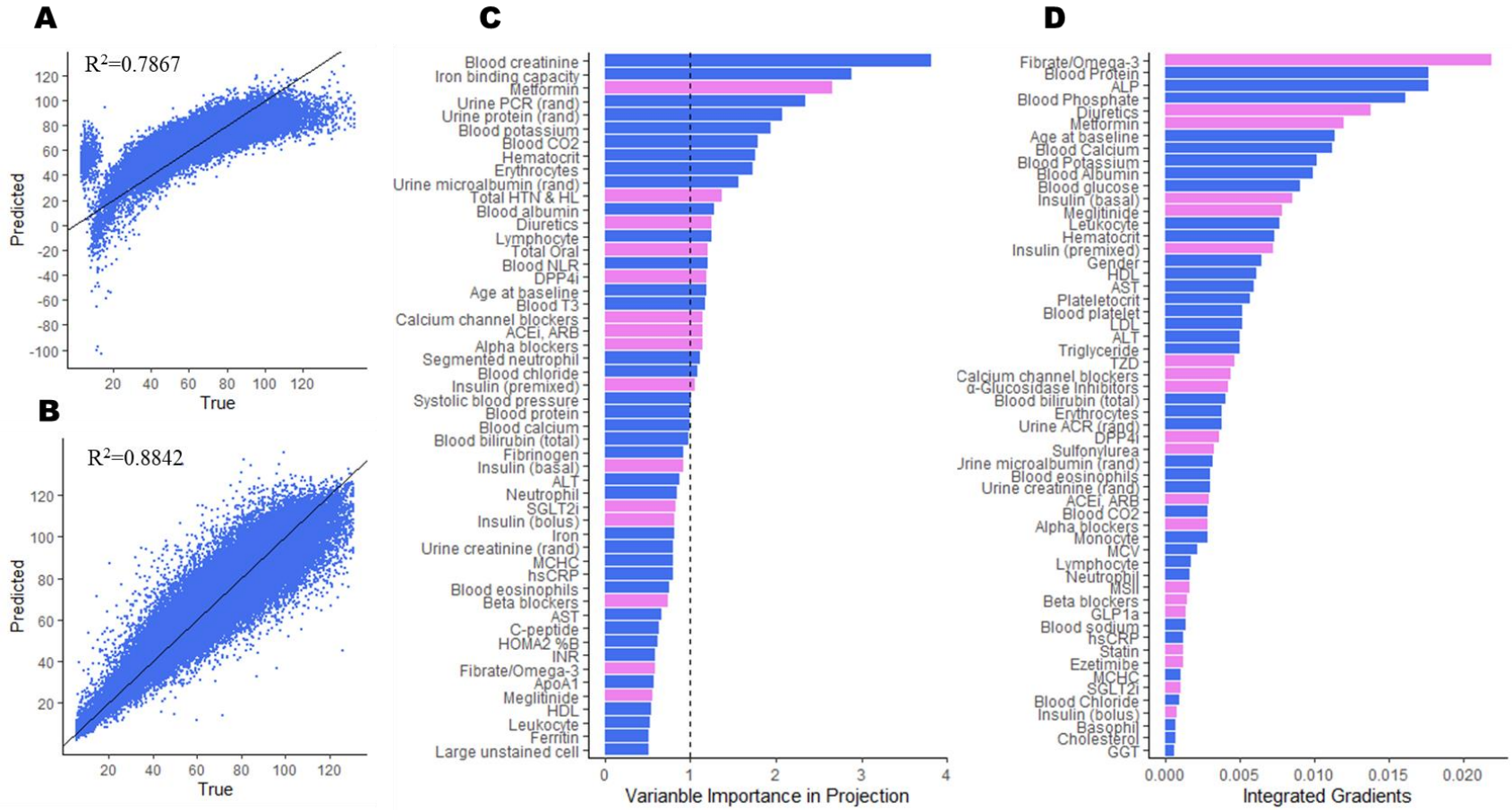

Figure 1 Comparison of PLS and RNN model. (A) Distribution of true and PLS model-predicted eGFR. (B) Distribution of true and RNN model-predicted eGFR. (C) Variable importance projection (VIP) in PLS model (VIP < 0.5 not shown). (D) Integrated gradients in RNN model (IG < 0.0005 not shown).

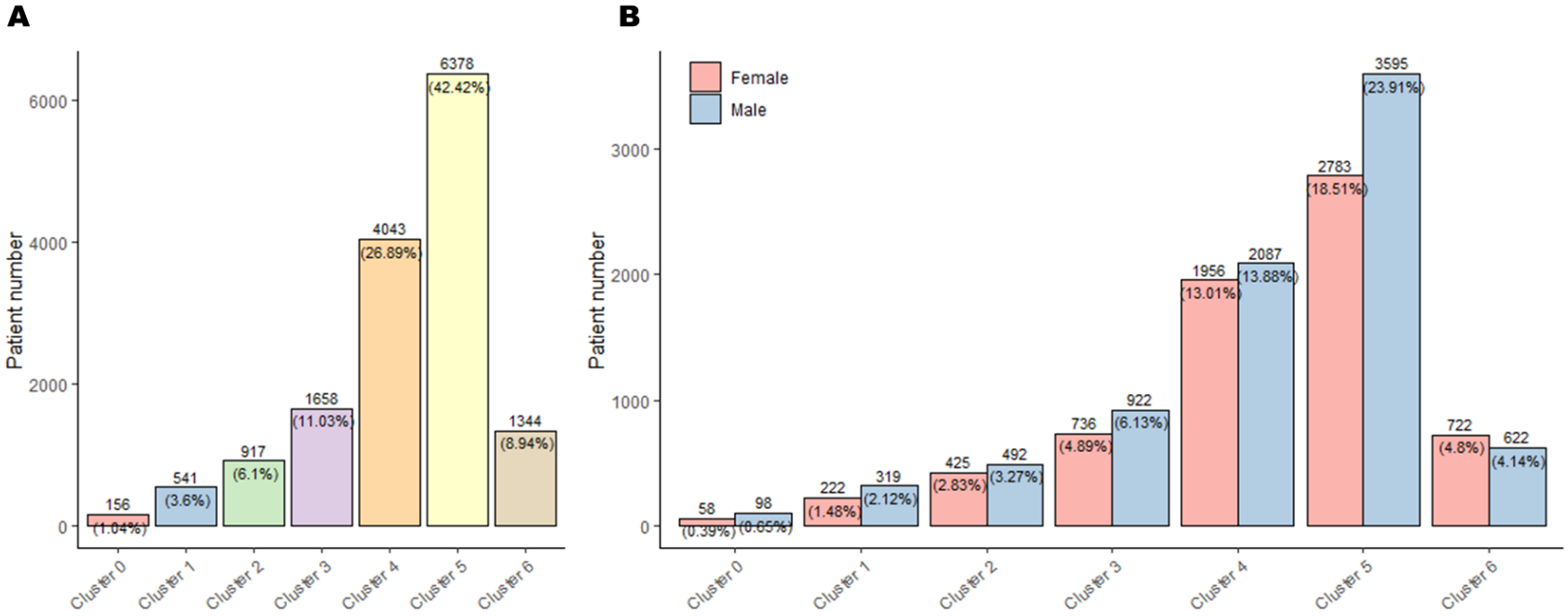

Figure 2 Patient distribution according to LSTM-based cluster. (A) Overall distribution. (B) Stratified distribution by gender.

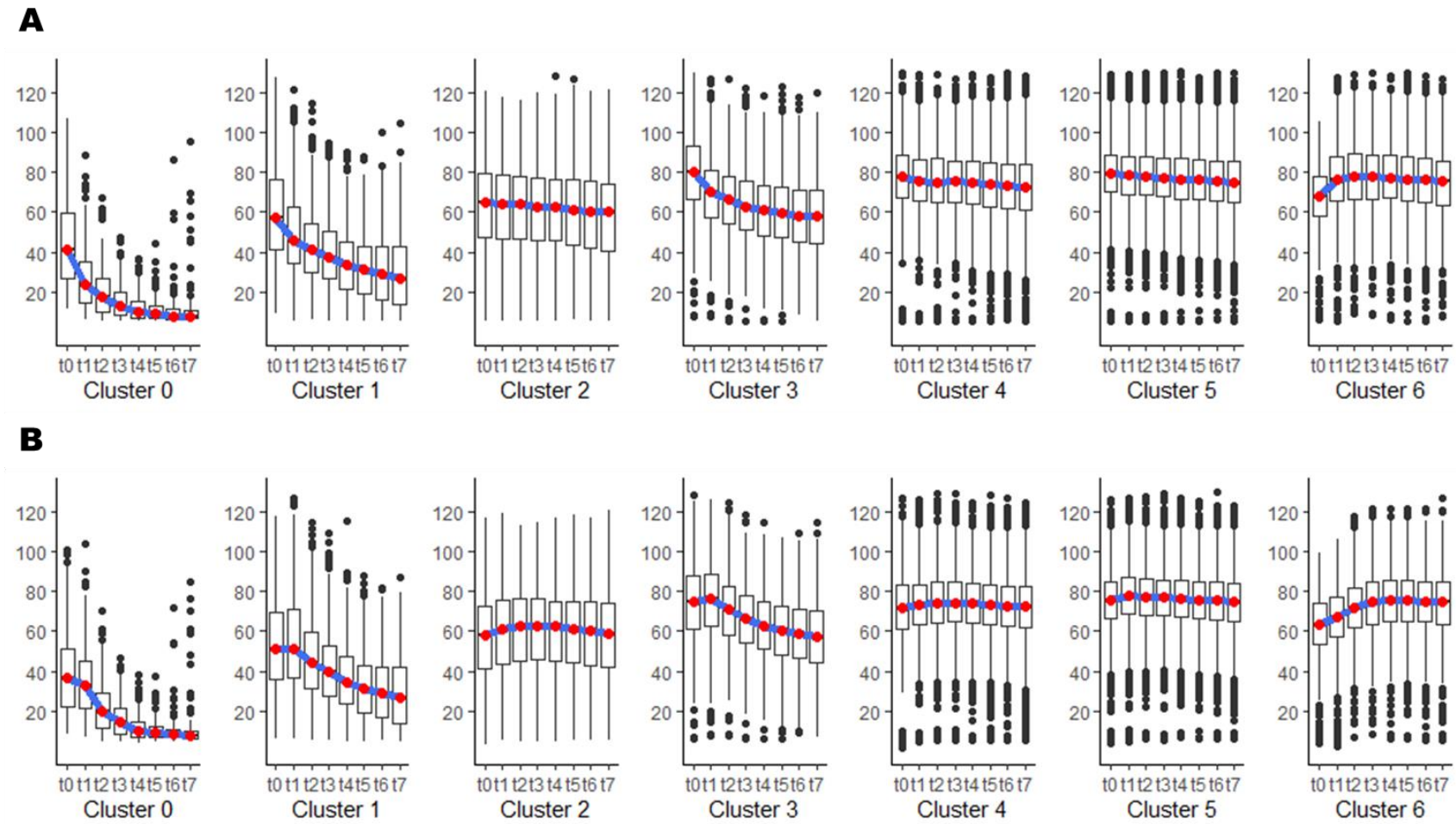

Figure 3 eGFR over time in SNUH cohort. (A) Observed eGFR in LSTM-based clusters. (B) Predicted eGFR in LSTM-based cluster. (Unit=ml/min/1.73m^2)

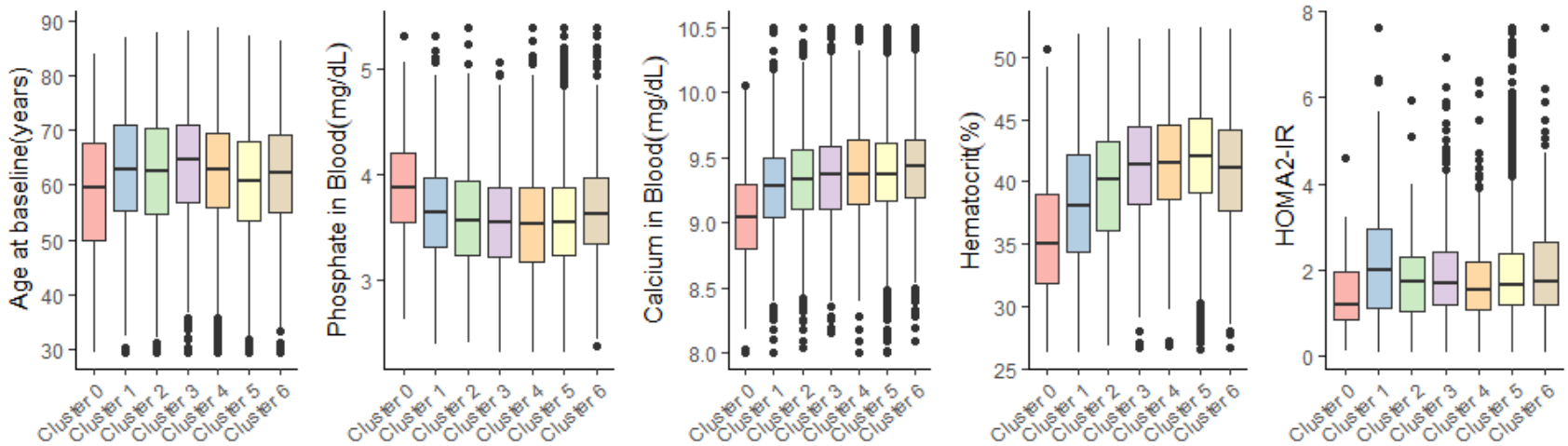

Figure 4 Baseline characteristics of SNUH cohort. The earliest measurement date within 30 days before the first diabetes mellitus diagnosis and afterwards was chosen as a baseline measurement date. All measurements on the baseline measurement date were considered.

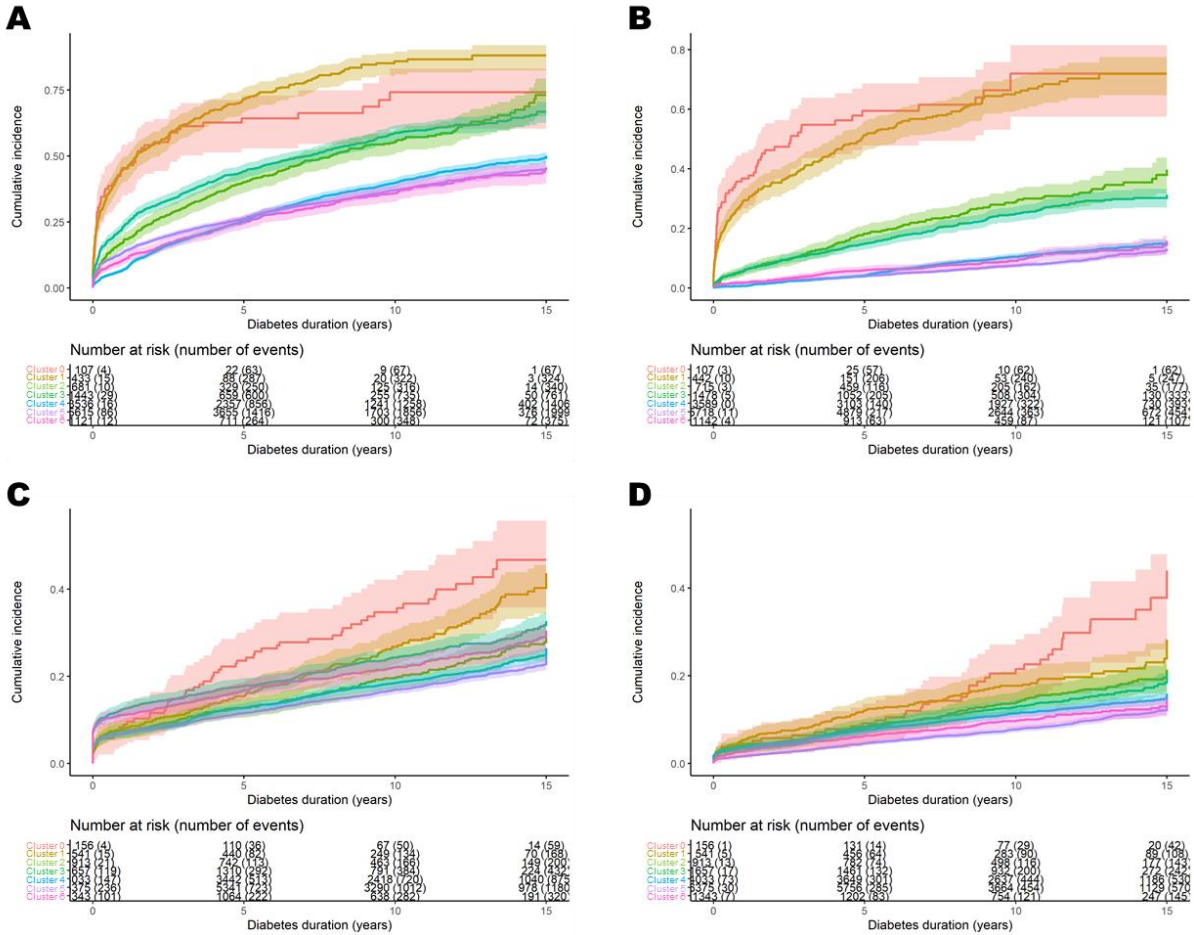

Figure 5 Progression of disease over time by cluster. (A) Time to CKDA2 (Microalbumin-to-creatinine ratio > 30mg/g). (B) Time to CKDA3 (Microalbumin-to-creatinine ratio > 300mg/g). (C) Time to coronary artery disease. (D) Time to cerebrovascular disease; coronary artery disease was defined by ICD-10 codes I20, I21, and coronary intervention. Stroke was defined by ICD-10 codes I60, I61, I62, I63, and I64.

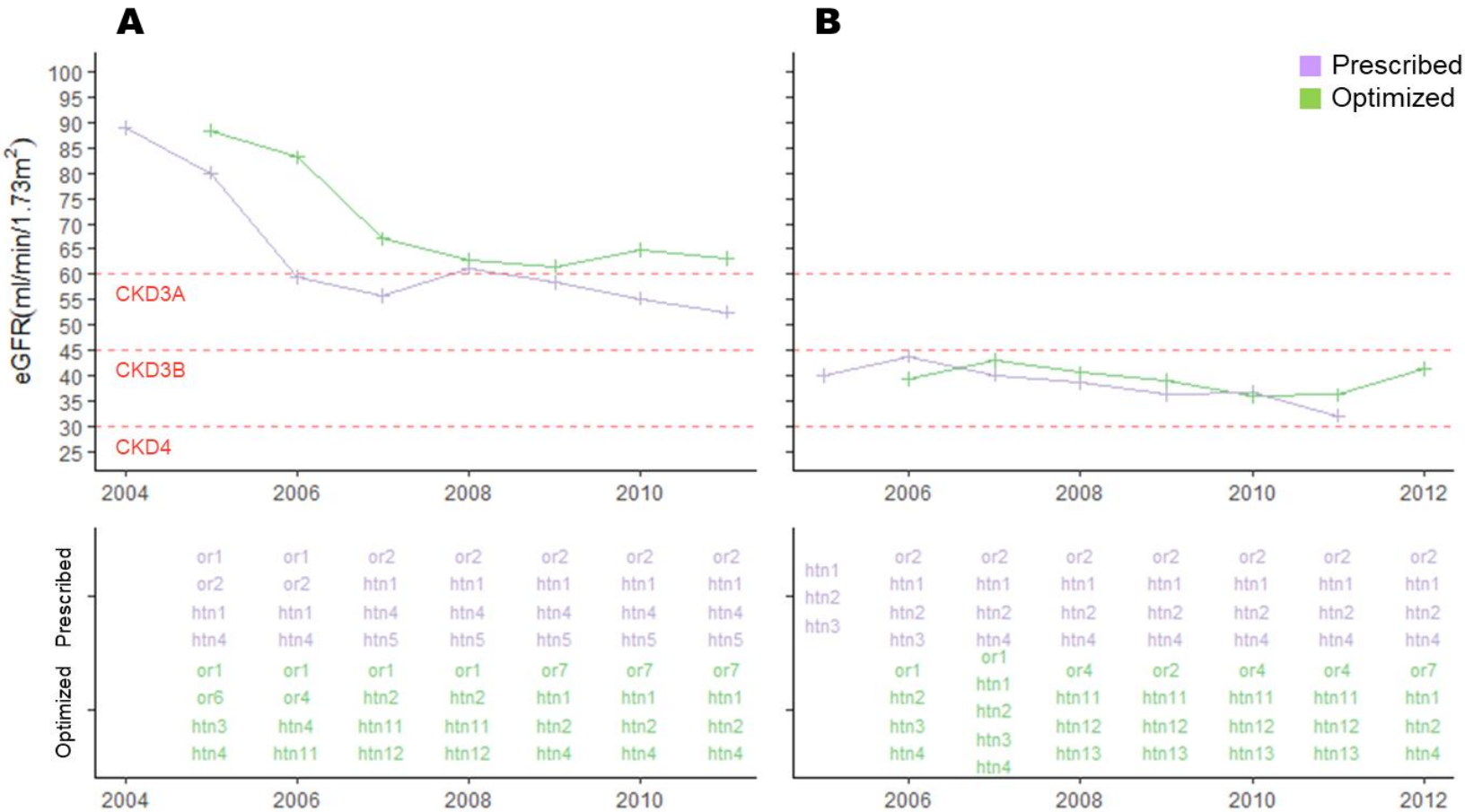

Figure 6 Individually optimized drug treatment and expected outcome. (A) Patient with mildly to moderately decreased eGFR. (B) patient with moderately to severely decreased eGFR. Abbreviation of terms used are as followed; or1 for metformin; or2 for sulfonylurea; or4 for thiazolidinedione; or6 for α-Glucosidase Inhibitors; or7 for meglitinide; htn1 for ACEi, ARB; htn2 for alpha blockers; htn3 for beta blockers; htn4 for calcium channel blockers; htn5 for diuretics; htn11 for statin; htn12 for ezetimibe; htn13 for fibrate/omega-3.

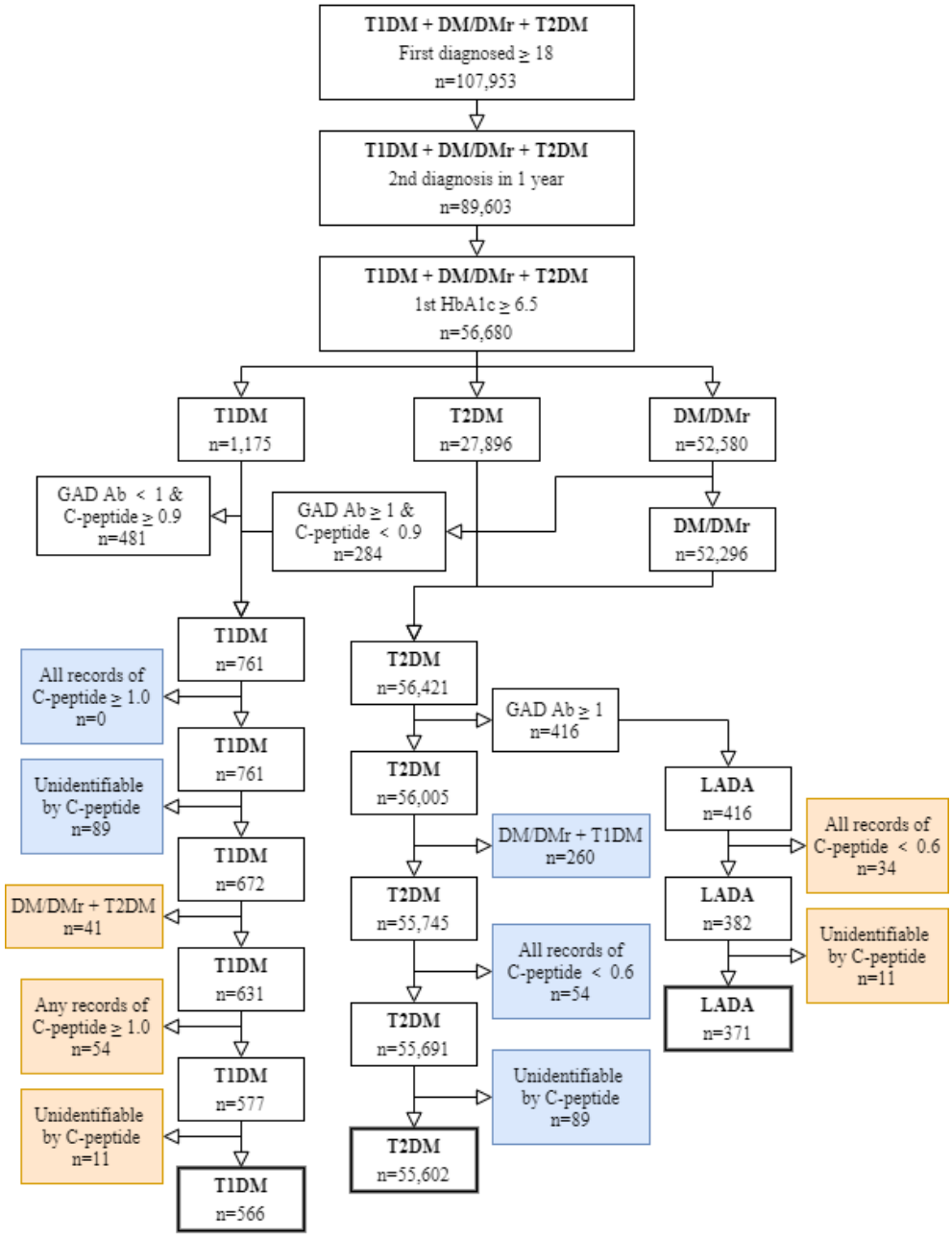

Figure 7 Construction of the study cohort in SNUH. Colored box indicates elimination logic of duplicated patients; Blue indicates duplicated patients in T1DM and T2DM; Orange indicates duplicated patients in T1DM and LADA.

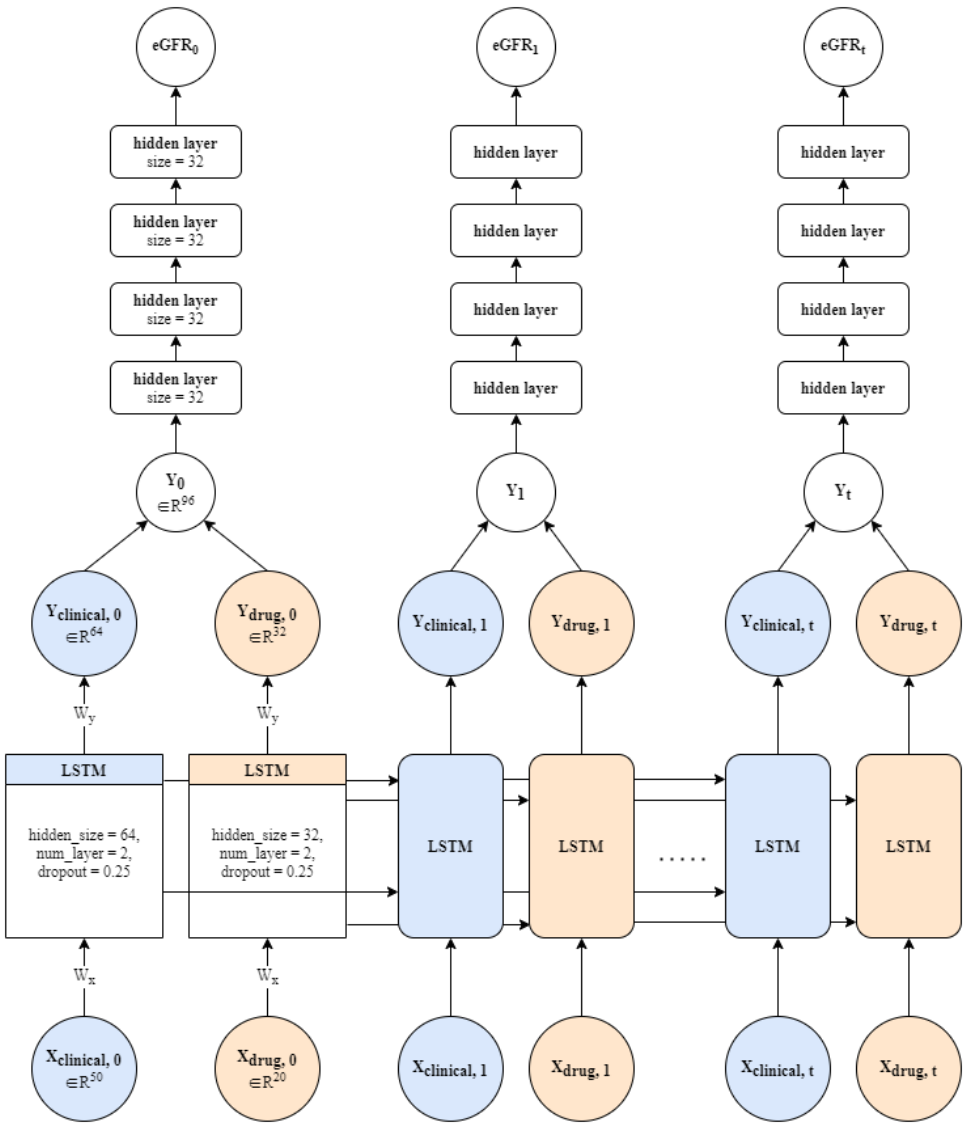

Figure 8 Diagram of recurrent neural network architecture. Hyperparameters are conserved throughout the sequential architecture; epochs = 500; learning rate = 0.0005; weight decay = 0.00001; batch size = 16.