

ディープラーニングによる中高年労働者の疾病ハイリスク化を 予測する手法の開発

群馬大学大学院 医学系研究科 公衆衛生学分野
准教授 内田 満夫

(共同研究者)

九州大学理学研究院 生物科学部門 数理生物学研究室 助教 野下 浩司

はじめに

近年、わが国はどの国も経験したことのない少子高齢化が進み、その結果人口構成の変化と労働力の減少を来している。それに対し、わが国では労働力の損失を防ぎ、さらに社会保障費の増加を抑制するため、高年齢者の雇用を促進している。しかし、その労働者の健康状態を見ると、わが国の労働者の健康診断有所見率は10年連続で50%を超え、また脳・心臓疾患による労災請求も増加し続けている。今後、中高年齢労働者の疾病予防策を構築し、長く健康に働く方法を提案することが必要である。

公衆衛生学分野では、リスクの大小にかかわらず集団全員を対象とした疾病予防策である「ポピュレーション戦略」に力を入れ、特に教育や啓発を中心とした一次予防の活動を行っている。一方でリスクの高い集団には「ハイリスク戦略」として早期発見早期解決という二次予防策を取っている。しかし、ハイリスク戦略は、すでにリスクが上昇している者を対象とするため、治療対象になってしまうことが多く、発病を防ぐ方法とは言い難い。したがって、ハイリスク者になりやすい対象を予め予測して、できる限り早期に介入を行うという取り組みがこれから必要である。

本研究では、蓄積された健康診断データから再帰型ニューラルネットワーク (RNN: recurrent neural network) のモデルを構築し、機械学習の手法の一つであるディープラーニング (DL: deep learning) を用いて対象の将来の健康診断結果を予測する技術を開発することを目的とした。また、社会医学研究においてこのDL技術の活用がいまだ進まない現状を考慮し、なぜ普及が進まないか考察することも目的とした。

結 果

研究 1

本研究では、群馬県内にある物流企業のA社（従業員数490人）より協力を得ることができた。A社の健康診断結果より、過去5年分の健康診断データを得た。データを受け取る際

には、企業側の担当者により個人を特定できる情報は全て削除してもらい、データと個人が連結できない状態にしてから研究者に送付してもらった。本研究内容は、群馬大学倫理審査委員会で承認を得た(HS2019-092)。

入手したデータから欠損値を除き、できるだけ多くの健診データを用いて予測するモデルの作成を検討したところ、174人のBMI (Body Mass Index)、腹囲、収縮期血圧、拡張期血圧、および血液データのHDLコレステロール、LDLコレステロール、中性脂肪、GOT、GPT、 γ GTP、ヘモグロビン、赤血球の12項目を利用することができた。対象をランダムに並べて最初の10名をテストデータとして切り分け、残りの164人を学習データとした。本研究では、モデルの訓練と予測値の推定にWINDOS10、64ビット、CPUCorei7、メモリ16GBのスペックのコンピューターを用いた。モデルの実装はDL向けのライブラリであるTensorflow (Google) およびKerasを用いた。RNNのモデルはLSTM (Long-Short term memory) ネットワークを用いた。中間層は128、エポック数は300、バッチサイズは16でモデルを構築し、訓練をおこなった。

本研究では、当初予定していた中高年労働者の循環器系の結果のみを予測するだけでなく、推定に利用した12項目の全てのデータを予測するモデルを作成することができた。164人のデータからRNNを構築して訓練し、予測値の推定を行なった結果、抽出した10人のBMI (図1)、腹囲 (図2)、収縮期血圧 (図3)、拡張期血圧 (図4)、HDLコレステロール (図5)、LDLコレステロール (図6)、中性脂肪 (図7)、GOT (図8)、GPT (図9)、 γ GTP (図10)、ヘモグロビン (図11)、赤血球 (図12) においてそれぞれ実測値 (左バー) と予測値 (右バー) を表示した。数値データを完全に一致させることは困難であるため、実データの $\pm 20\%$ の値を得た場合に予測が的中したとみなした。その結果、訓練モデルデータ内で評価した精度は92%の高さを示したが、汎化を評価するために予め抽出しておいたテストデータ10名の予測精度は63%に留まった。本研究のモデルのパラメータは、中間層やエポック数を試行錯誤で適切な値を探して設定したが、それぞれの値を少なくするとテストデータの精度は下がった。しかし同時に、これ以上値を大きくしても精度の向上には寄与しなかった (図13)。

以上より、中高年労働者の過去の健康診断データを用い、RNNのモデルを用いて予測を行なった結果、本研究は将来のデータを約60%の精度で予測することができた。血液データの予測では、データの変動のトレンドから逸脱した対象者 (例えば最終年だけ突然値が上昇した場合など) は予測が困難な例も存在した (例えば図10のケース4など)。

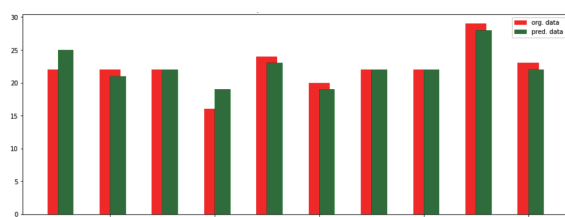


図1 BMI

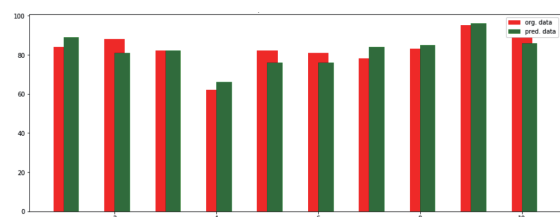


図2 腹囲

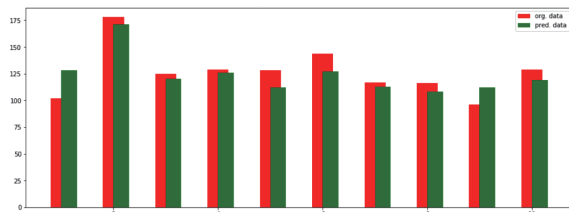


図3 収縮期血圧

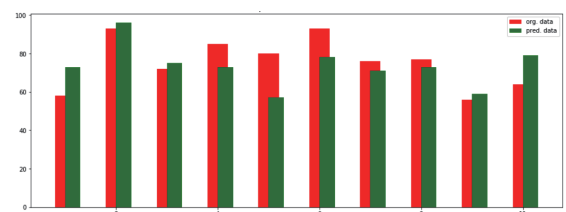


図4 拡張期血圧

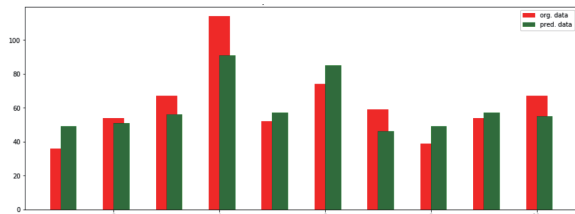


図5 HDLコレステロール

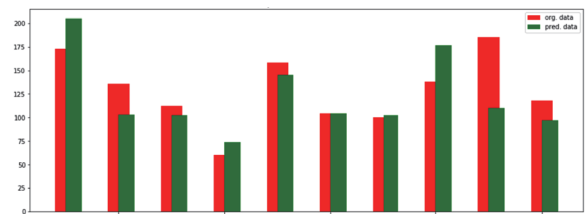


図6 LDLコレステロール

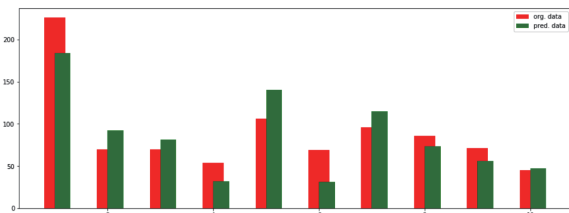


図7 中性脂肪

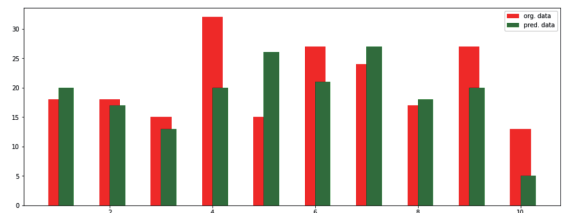


図8 GOT

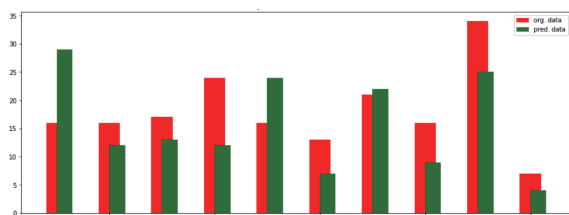


図9 GPT

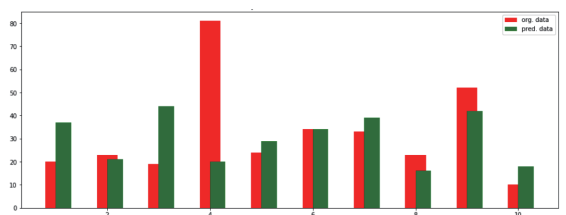


図10 γ GTP

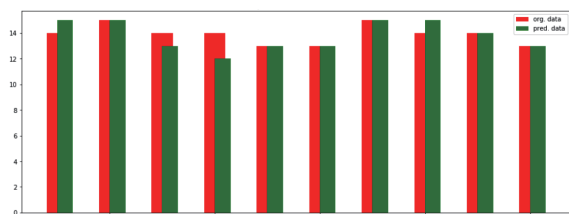


図11 ヘモグロビン

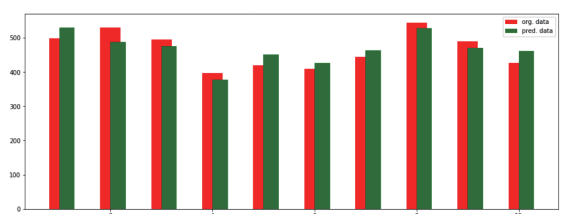


図12 赤血球

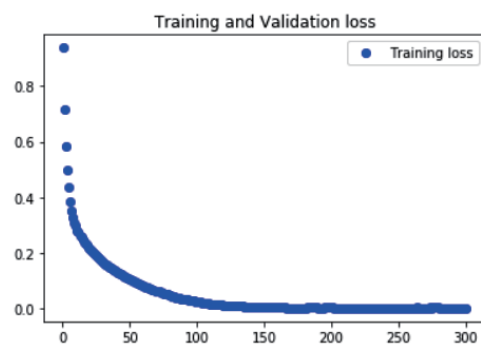


図13

訓練の繰り返しによる損失関数の値の低下

エポック数150回あたりから

損失値はほとんど変化しなかった。

研究2

本研究では、上記と並行して、わが国の社会医学におけるAIの活用が多くない現状を受け、解説論文を作成した^(1, 2)。これらの論文では、医学研究におけるAIの位置づけ、社会医学におけるAI研究、AI研究の現在の課題を論じ、今後社会医学におけるAI研究の必要性について述べた。

考 察

このたび、1つ目の研究では、RNNを用いたDLにより、過去の健康診断データから最新の健康診断結果を予測することを試みた。その結果、過去4年のデータから60%ほどの精度で次年の結果を予測することが可能であった。疾病のハイリスク化を予測するためには各種健康指標の上昇トレンドを察知して早期にアラートを示すことが重要である。本研究では約60%の精度で1年先の結果を予測することができたため、例えば健康診断結果の説明欄に“現状の生活を継続した場合の1年後の予測値”を記載して返却することで、ハイリスク者に該当する対象がいた場合は生活改善を促すための資料とすることができる可能性がある。本研究の成果は、今後のわが国の疾病リスクを抑制するための取り組みとして活用が期待される。

当初の計画では蓄積データを用いて、訓練年数や予測年数のバリエーションを評価することを目的にしていた。しかし、協力企業から得られたデータは5年分と予想を下回ったため、予測年数の代わりに予測項目を増やすこととした。その結果循環器以外の項目も予測することが可能となり、脂質や肝機能といった代謝系臓器のハイリスク化も予測することができた。ただし、対象の観測値が予測トレンドから著しく逸脱するケースもあり、本予測が困難なケースも存在した。これは、本研究で使用したデータが200例弱であり、また訓練に使用したデータが4年分と短かったことが影響している可能性がある。今後は対象サンプル数を蓄積し、また健康診断のデータの蓄積年数を増加させることで精度を向上させなければならない。

また、2つめの研究成果として、社会医学におけるAIの活用の現状を調査した結果、臨床医学と比較するとその活用の機会は乏しく、調査報告は不足していることを明らかにした。これは、DLに利用できるデータが上手く活用されておらず、また社会医学系の研究者の多くが研究に携わっていないためと推察された。研究ツールとして使用するAIは、これまで伝統的に使用されてきた頻度論に基づく医学統計とは異なる概念であり導入することに未だ抵抗を感じる研究者も存在する。これからAIを広く導入するためには、伝統的手法と比較したときにどの程度有用であるかという実績を蓄積すること、また利活用するのに適切な方面を明確にすること（例えば画像解析や時系列データ）が重要である。本研究報告では社会医学において利用できるデータおよび活用方面を報告することができた。今後は社会医学系の研究分野でもAIの利活用が進められることが期待される。

要 約

本研究では、労働者の将来的な健康障害を予防する施策を構築するために、定期健康診断結果とディープラーニングの技術を用いて将来のデータを予測することを目的とした。協力企業1社から健康診断に関連する情報を提供してもらい、174名の5年分のデータを研究に使用した。再起型ニューラルネットワークを構築して、予測値を推定した結果、次年のデータを約60%の精度で的中させた。本研究の結果より、蓄積された健康診断データを用いて将来予測を行うことができる可能性が示唆された。今後さらに研究を継続してデータを収集し、モデルの精度を向上させることが必要である。

謝 辞

本研究を進めるにあたり、技術的支援および研究内容に適切な意見を示していただいた筒井保博先生（日立金属）に深く感謝申し上げます。

文 献

1. 内田満夫, 野下浩司, 渡辺由佳子, 小山洋. 社会医学におけるAI研究. BIO Clinica 34 (3) ; 89-93. 2019
2. 内田満夫, 野下浩司, 渡辺由佳子, 小山洋. 社会医学におけるAI研究. 細胞 51 (7) ; 42-45. 2019