Vol.2018-EIP-79 No.13
2018/2/16

ビッグデータの統計学的意義

早野順一郎† 湯田 恵美† 吉田 豊†

概要：ビッグデータ解析への批判として、統計学は少数例のサンプリングによる母集団特性の推定を追求して発展したものであるから、少数例で得られた統計的知見以上のことがビッグデータから見つかることは希である、と言われる。そこで、心拍ビッグデータを用いてそのような事が実際に見られるかどうかを検証した。24 時間心電図ビッグデータ Allostatic State Mapping by Ambulatory ECG Repository (ALLSTAR) の洞調律例を対象に、仮説 1 :心拍数は気圧の上昇時と下降時で差があるか、仮説 2：心拍数には季節変動があるか、仮説 3：心拍数は人口密度の影響を受けるか、という 3 つの仮説を、男女それぞれからランダム抽出した 10 万、5 万、2 万、1 万、5 千、2 千、千、5 百、2 百、百例で統計解析した。仮説についても、統計学的に有意に達した例数以上に例数を増やしても結果には実質的な違いはなかった。また，有意な結果を得るための必要サンプル数は statistical power 解析(α<0.05、β>0.8)の結果と一致した。本研究の結果，統計学的有意性は、母集団の特性に対するサンプルの代表性の指標となり、それ以上のサンプルを集めても結果は大きく変わらないことが示された。

キーワード：ALLSTAR, 24 時間心電図，ビッグデータ，心拍変動，統計学

Statistical Significance of Big Data

JUNICHIRO HAYANO† EMI YUDA† YUTAKA YOSHIDA †

1. はじめに

情報通信技術の発展に伴って，様々な領域で膨大なデータが比較的容易に収集・蓄積できるようになった．その結果，測定精度を考慮したデータ数がこれまでの既成概念を超越したいわゆるビッグデータが形成されるようになった．かつてない質と量の情報が得られるようになった事で，その解析で得られる新しい知見に対する期待が高まっている．

一方，ビッグデータに対する過剰な期待に対する批判として，そもそも統計学は，母集団から抽出した少数のサンプルから母集団の特性を推定することを追求として発展してきたのだから，ビッグデータを解析しても小数例の解析で既に知られていること以上の知見が得られることは希であると言われている．しかし，この批判自体も仮説であって，その妥当性の検証には母集団のデータを実際に解析し，小数例のサンプルから得られる結果と比較する必要がある．そこで，本研究では 24 時間心電図のビッグデータである Allostatic State Mapping by Ambulatory ECG Repository (ALLSTAR) データベースを用い，小数例から統計学的に導かれる知見の妥当性，つまり，統計学的の判断基準である有意性が，抽出されたサンプルの，母集団の特性に対する代表性の基準となり得るか否かを検証した．

2. 方法

2.1 ALLSTAR プロジェクト

本研究では，ビッグデータの例として，ALLSTAR プロジェクト[1-3]の 24 時間心拍時系列のデータベースを使用した．本プロジェクトでは，日本全国で記録されるホルター心電図の約 5%にあたる年間約 6 万件のホルター心電図データの収集とデータベース化を進めており，現在，約 40 万件が登録されている．

ALLSTAR プロジェクトの使用データは，日本国内の医療機関が(株)スズケンの札幌，東京，名古屋にある心電図解析センターに解析を依頼した 24 時間ホルター心電図の内，検査対象者によるオプトアウトの申し出のあったものを除いた全データである．したがって，これらのホルター心電図は，疾患のスクリーニング，診断，治療効果判定など，何らかの医療目的で記録されたものである．

2.2 対象データ

ALLSTAR データベースの約 40 万件の 24 時間 R-R 間隔時系列データの内，次のいずれかの除外基準に相当するものを除外した男性 113,777 例, 女性 140,580 例を本研究の対象データとした．

(1) 全心拍の 20%を超える拍が洞調律でないもの
(2) 持続性または発作性心房細動，または心房粗動
(3) ペースメーカ埋込例
(4) 年齢，性別，記録日時，郵便番号の欠損例

2.3 データ分析

24 時間 R-R 間隔時系列データより, 連続する洞調律からなる R-R 間隔(normal-to-normal，N-N 間隔)のみを抽出し，その平均値と標準偏差を求めた．60,000/平均 N-N 間隔(ms)を 24 時間平均心拍数(HR, bpm), 標準偏差を SDNN (standard deviation of N-N interval over 24 h，ms)として用いた．

†名古屋市立大学大学院医学研究科
Nagoya City University Graduate School of Medical Sciences

2.4 検証仮説

本研究では，次の 3 つの仮説を題材として設定した．

仮説 1 : :「心拍数および心拍変動は気圧の上昇時と下降時で差があるか」．ホルター心電図は 24 時間に渡って記録されるが，実際の記録は，1 日目の昼間に開始され，2 日目の同時刻に終了する．そこで，気象庁のデータベースから得た 1 日目と 2 日目の平均気圧を比較し，1 日目より 2 日目の気圧が低い場合を下行気圧，そうでない場合を上昇気圧として，各条件で記録された HR と SDNN を比較した．

仮説 2 :「心拍数には季節変動があるか」．HR データを，ホルター心電図の記録された月(1 月～12 月)によって分類し，月による変動を調べた．

仮説 3 :「心拍数および心拍変動は人口密度の影響を受けるか」. HR と SDNN データを郵便番号によって都道府県に分類し，2017 年 10 月 1 日の各都道府県の人口密度(人/km^2)との相関を分析した．

2.5 統計解析

統計解析には Statistical Analysis System (SAS　institute, Carry，NC，米国）のプログラムパッケージを使用した．対象データから，男女それぞれ，10 万，5 万，2 万、1 万、5 千、2 千、千、5 百、2 百、百例のサンプルをランダムに抽出し，各サンプルについて以下の統計解析を行った．仮説 1 では，上昇気圧時と下行気圧時の HR と SDNN を t 検定によって比較した．仮説 2 では，HR に対する測定月の影響を一般線型モデルによって解析した．月はカテゴリカルデータとし，年齢の影響を調整して，月の HR に対する影響を検定した．仮説 3 では，年齢の影響を除外した偏相関係数を用いた．統計学的有意性の基準には，type I error level α <0.05 を用い，statistical power 分析には，type II error level β >0.8 を用いて必要サンプル数を推定した．

3. 結果

3.1 仮説 1

上昇気圧と過去気圧の時の HR の差は，男性ではみられなかったが，女性では 5 万例以上で下行気圧時の 0.2 bpm の増加が有意差として検出された(両群の標準偏差は 10.1 bpm)．SDNN は下行気圧時の 1 ms の有意な低下が，男性では 10 万例のみで，女性では 5 万例以上で検出された(両群の標準偏差は男性 44 ms，女性 40 ms)．これらのサンプル数は statistical power 分析の結果と一致した．

3.2 仮説 2

1 月に頂値 9 月に底値を示す HR 季節変動が，男性では 5 千例以上，女性では 2 千例以上で検出され，その範囲ではサンプル数を増やしてもパターンに変化はなかった．

3.3 仮説 3

HR と人口密度の間に，男性では r = 0.04 の相関が 2 万例以上で，女性では r = 0.03 の相関が 5 万例以上で検出された．SDNN は男性では r = -0.03 の相関が 5 千例以上で，女性では r = -0.01 の相関が 5 万例以上で検出された．

4. 考察

ビッグデータより抽出した 100 から 100,000 例までの様々なサイズのサンプルを用いて，統計的有意性と母集団の特性に関するサンプルの代表性を 3 つの仮説を題材として検証した．その結果，いずれの仮説についても，サンプル数を増やすことで有意な結果が得られたが，結果が有意となった数以上にサンプル数を増やしても，結果には実質的な変化はなかった．この事は，統計学的有意性が，抽出されたサンプルの，母集団に存在する特性に対する代表性の指標となっていることを示している．つまり，適切にサンプリングされた少数例において，統計的有意性を持って検証されている特性は，ビッグデータを用いて分析してもその結果に実質的な差異を生ずる可能性は低いと言える．

ビッグデータは，統計的手法を使用することで，極めて小さな差異や微弱な関連をも捉えうる強力な検出力を発揮する．本研究でも，仮説 1 では，標準偏差 10 bpm の 2 群間の 0.2 bpm の平均値の差が 5 万例以上のサンプルで検出された．サンプル数を何処までも増やすことができれば，どのような僅かな差異も微弱な関連も，いずれは有意な差として検出される．したがって，ビッグデータの解析においては，どの程度の差異や関連(effect size)に意味があるのか, 事前に明確にしておく必要がある. また, その effect size によって，その検証に必要なサンプル数も規定されることから，不必要に大きなデータを扱うことによって生ずる無駄なコストを避けることができる．

今後，様々な領域におけるビッグデータの活用が期待される中で，統計理論の重要性はむしろ高まり，ビッグデータの効果的かつ効率的な解析に実質的な効力を発揮することになるものと考える．

参考文献

[1] ALLSTAR Research Group. (April 21). *Allostatic State Mapping by Ambulatory ECG Repository (ALLSTAR)* Available: http://www.med.nagoya-cu.ac.jp/mededu.dir/allstar/index.html

[2] E. Yuda, Y. Furukawa, Y. Yoshida, J. Hayano, and ALLSATR investigators, "Association between regional difference in heart rate variability and inter-prefecture ranking of healthy life expectancy: ALLSTAR Big Data Project in Japan," in *Big Data Technologies and Applications: Procedings of the 7th EAI International Conference, BDTA 2016*, J. J. Jung and P. Kim, Eds., ed Seoul, Korea: Springer Nature, 2017, pp. 23-28.

[3] J. Hayano, E. Yuda, Y. Furukawa, and Y. Yoshida, "Association of 24-hour heart rate variability and daytime physical activity: ALLSTAR big data analysis," *International Journal of Bioscience, Biochemistry and Bioinformatics,* vol. 8, pp. 61-67, 2018.