

今後の国民生活基礎調査の 在り方についての一考察（第3報）

ハシモト ヒデキ
橋本 英樹*

目的 国民生活基礎調査の現行のサンプリングデザイン・実施体制について、回収率向上や匿名化データセット作成の観点から問題点を整理した。

方法 国民生活基礎調査室の担当数理官へのインタビューを行い、サンプリングならびに比推定の手法について質疑応答を通じて取材した。また実施の状況について、某都道府県担当部局を通じて、調査実施担当者数名からのグループインタビューを実施させていただいた。匿名化データセットの作成に関する問題点については、平成21年度厚生統計協会研究委託「国民生活基礎調査の匿名データ化に関する研究会」に筆者が参加した際の議論も一部踏まえつつ、問題点を整理した。

結果 大調査年では都道府県別表章の誤差範囲を均一化すること、実施体制・コストの制限から地点抽出・地点内悉皆調査により比推定に基づく母数推計が行われている。全国値の算出にあたって分散の違いが考慮されておらず、確率抽出を採用している小調査年統計値との整合性を再検討する余地がある。また比推定は回収率低下によるバイアスの影響を受けやすいことから、実施体制の見直し、特に実施系統の一本化・調査員の教育・情報普及など検討すべきである。匿名化データ作成にあたっては現行のサンプリングデザインに制限を考慮すれば、世帯員レベルでのリサンプリングが匿名性を保ちつつ地域情報を含めるうえで自由度が高いと思われた。

結論 国民生活基礎調査は、世帯面の基幹統計として、変化する調査環境・ユーザーとそのニーズの多様化に対応するには、統計の継続性を確保しつつも、常にその内容や実施方法について大胆な変革を見通した議論を継続していく必要がある。対象者である国民に対して「国民の共有財産としての統計」としての正当性を明確に説明できることが求められる。

キーワード 国民生活基礎調査、サンプリングデザイン、調査実施手法、回収率向上、匿名化データセット

I はじめに

国民生活基礎調査は、国民生活に関わる世帯の社会・経済・健康・医療・介護福祉などの諸側面を包括的に捉える基幹統計調査である。著者らは平成19年度ならびに20年度厚生統計協会委託研究事業の一環として国民生活基礎調査のあり方について健康票を中心とした検討を行い、

これまで2報を報告した¹⁾²⁾。第1報では、本調査における健康の位置づけとして、アウトカムとしての健康と、社会生活を支える基本的資本としての健康の2つの側面があることを指摘したうえで、世帯を単位として健康に影響する社会経済的要因を包括的に測定し、「社会的健康決定要因」²⁾に対する政策的取り組みとの位置づけを明確にしつつ、測定項目の削除・追加を検討する必要があると主張した。第2報では、健康状態の測定指標として本調査で中核的な役

* 東京大学大学院医学系研究科公共健康医学専攻教授

割を果たしてきた自覚的健康状態の報告バイアスの検討と、平成19年度調査より導入されたメンタルヘルス尺度であるK6の表章の在り方について検討を行った。

平成21年度事業では、これら先行の検討を踏まえたうえで、国民生活基礎調査の直面する3つの根幹の問題について論点整理を行うこととした³⁾。第1に、同調査のサンプリングデザインについて、検討を行った。第2に、調査実施方法について、特に回収率の問題を意識しながら、現行手法の課題を整理した。第3に、本統計を用いた匿名化データセットの公開について、検討すべき点について提言を行うこととした。

国民生活基礎調査の大調査年（3年ごと）は、小調査年ならびに他の世帯面統計（総務省の全国消費実態調査など）と異なるサンプリングデザインを施している。国民生活基礎調査の結果と国勢調査の結果で、年齢・性構成が一致しない原因のひとつとして、回収率の低下とあわせて、このサンプリングデザインが影響している可能性がある（図1）。このサンプリングデザインが採用されているひとつの理由は、同調査の実施方法と関係している。同調査は都道府県の保健所ルートと、社会福祉事務所ルートの2系統で、前者が世帯票・健康票を、後者が所得票を同年の別の時点で実施している。調査実施コストを低く抑えることができる一方、必ずしも社会調査遂行に必要な技術訓練を受けているとは限らない調査員を通じて実施していること

が、サンプリングデザインに制限を与え、かつ回収率の問題とも関連している可能性について考察する。さらに、本調査特有のサンプリングデザインは、匿名化データセットの作成においても、後述するようにリサンプリングによる秘匿化を困難にしている。すなわち、上記の3つの問題は相互に関連した問題なのである。

Ⅱ 方 法

国民生活基礎調査の実施方法ならびにサンプリング手法については公表されているものの、その詳細について十分記載されていない。そこで、国民生活基礎調査室の担当数理官へのインタビューを行い、実施手法の詳細について質疑応答を通じて取材した。また実施の状況について、某都道府県担当部局を通じて、調査実施担当者数名からのグループインタビューを実施させていただいた。匿名化データセットの作成に関する問題点については、平成21年度厚生統計協会研究委託「国民生活基礎調査の匿名データ化に関する研究会」（座長；安田聖。一橋大学経済研究所教授）に筆者が参加した際の議論も一部踏まえつつ、整理した。

Ⅲ 結 果

（1） サンプリングデザイン

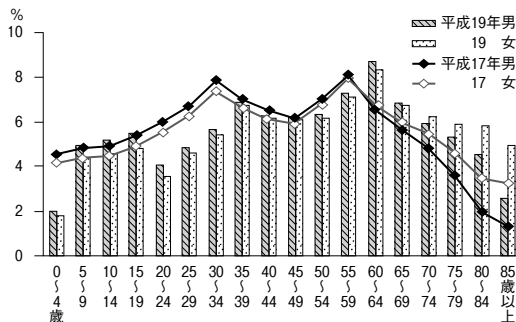
総務省統計などと比較して、国民生活基礎調査大調査年のサンプリングの特徴は、都道府県表章の推計誤差を均等に近づけることを目的として、都道府県ごとにほぼ均等の地点数（調査区）を無作為抽出したうえで、地点内の世帯について悉皆調査を行うという点にある。つまり人口規模による確率論的抽出（多段階クラスターサンプリング）を行っているわけではない。

1) 調査地区の抽出

国勢調査の調査区は全国で約93万区が設定されている。このうち、国民生活基礎調査では5,400区が選ばれ、さらに各調査地区では、その下位の単位区（約12,500）が選ばれている。

1単位区内にはほぼ均等な世帯数が含まれるよ

図1 平成19年国民生活基礎調査人口と抽出元の平成17年国勢調査人口（総人口）の分布比較（性別・5歳階級別）



出典 文献4)

うに設定されている。調査地区の抽出は、約93万地区すべてから無作為抽出するのではなく、調査実施の数年前までの総務省の先行調査（家計調査や全国消費実態調査ほか）と重複している地区をあらかじめ除いて抽出している。平成19年調査では93万地区のうち32万地区が除外されたうえで、残る60万余から無作為抽出がなされている。その理由としては、複数調査が重なることによる回答者負担の増加を避けるためと説明されている。先行調査も調査区を無作為抽出していることから、それでも無作為性は担保されている、というのが総務省ならびに厚生労働省統計情報部の見解である。

2) 比推定による母数推計

大調査年では都道府県表章の数字が必要なので都道府県同士の推計値について誤差を均等に近づけることが要求されている。そこで都道府県ごとに原則的に同数の調査区を抽出し、そこで得られた数字を、その人口サイズに拡大するというのが国民生活基礎調査のサンプリングの基本設計概念である。具体的には、抽出された調査区内に含まれる全世帯を対象に実施された悉皆調査の結果に基づき、各都道府県の母数推計は比推定により求められている。その手順は以下のとおりである。

- ① 各調査区の調査結果から得られた世帯特性（たとえば世帯類型など）の出現数を都道府県レベルで集計する。
- ② 各調査区の調査結果から得られた世帯員数を都道府県レベルで集計する。
- ③ 上記から特性を世帯員数で除して、世帯員数当たりの比率を求める。
- ④ これを国勢調査と人口動態から得られる推計都道府県人口に掛け合わせる。

したがって、精度の高い母数推計を得るためには3つの前提が必要である。

- i) 都道府県を世帯数がほぼ等しいと思われる地区（調査区）に分画できていて、無作為に地区が抽出できている。
- ii) 世帯特性は人口と高い相関をもつ
- iii) 調査対象となった地区について全数が調査できている

上記のうち、i) はおよそ問題ないとしても、ii) そして特にiii) が問題となる。特に、未回答による非系統誤差が生じている場合、比推定では観察標本から得られた比率をそのまま都道府県人口に乗じるため、バイアスを拡大する可能性を含んでいる。

3) 拡大乗数

国民生活基礎調査個票では世帯票・健康票、所得票、そして介護票それぞれに都道府県ごと（特別市は別途）に拡大乗数が計算されている。世帯票・健康票の拡大乗数は、推計人口を世帯票・健康票の観察世帯数で除したものである。介護票の拡大乗数は、推計人口に対して（介護票での観察介護認定者数÷観察世帯人員数）で得られた介護認定者割合を乗し、「認定者推計数」を出し、それを介護票観察認定者数で除して作成している。さらに所得票は、調査地区のさらに下位の単位区が2段階で無作為抽出されているが、その拡大乗数は、都道府県ごとに

$$\left(\frac{\text{国勢調査調査地区数}}{\text{国民生活基礎調査世帯票実施調査地区数}} \right) \times \left(\frac{\text{国民生活基礎調査実施単位区数}}{\text{所得票実施単位区数}} \right)$$

で計算されている。このように拡大乗数は、年齢・性別ほか属性の分布が都道府県（ないし特別市）内で均一であるという強い仮定に基づいており、これも国勢調査や介護給付実態調査の結果と、国民生活基礎調査の結果が食い違う要因になっている可能性がある。

4) 大調査年と小調査年の全国値の表示

大調査年では、都道府県表章を行う要請から比推定の考え方に基づいて都道府県ごとに推計値を求めたうえで、全国集計値は、各都道府県の推計値の単純合計から得られている。しかし都道府県ごとに抽出率が異なることから、人口規模による分散の違いを考慮しないで全国推計値を得ていることは（回収率の問題以外の要素として）国勢調査と推計値が一致しない原因となっている可能性がある。一方、小調査年では、総務省統計（全国消費実態調査など）と同じく、確率論的サンプリングを採用し全国値を求めている。したがって小調査年と大調査年で全国推計値の比較可能性があるかどうかを検証しなお

す必要がある。残念ながら抽出率を揃える作業は、抽出の元となる国勢調査地区情報（93万地区の情報）がないため現時点では正確には行えない。1つの代替方法としては、大調査年のデータから人口比率などで重みづけしてリサンプリングを行い、そのデータを用いて小調査年を含めた経年表示向けの全国値を求めること、などが考えられる。ただし、リサンプリングにより情報損失するデメリットもある。少なくとも、表示にあたっては推計方法が異なることをわかりやすく明記しておくことが必要であろう。

（2）調査実施体制と調査手法

世帯票、健康票、介護票および貯蓄票は、あらかじめ調査員が配布した調査票に世帯が自ら記入し、後日調査員が回収する自計方式により行われている。所得票は調査員が面接聞き取りする他計方式がこれまで採用されてきた。世帯票・健康票・介護票は各都道府県の保健所系統を通じて調査年の6月に実施され、それを受けて同年の7月に福祉事務所の系統を通じて所得票・貯蓄票調査が実施される。すなわち国民生活基礎調査の実施にあたり、都道府県保健所・福祉事務所のボランティア、ないし都道府県自治体の負担に頼っており、調査票によって異なる系統の調査員が別々にあたっているのが現状である。

総務省統計や国民生活基礎調査の小調査年は人口サイズなどを加味した確率的標本抽出を採用している。国民生活基礎調査大調査年でこの方法を取りにくい、消極的理由として、確率論的サンプリングを採用すると調査員の負担が重くなることが挙げられている。限られた地点内を悉皆的に調査するほうが、より多くの地点に散らばる確率的に抽出された対象者との接触を図るより負担が少ないということである。そして地域全体の人口は国勢調査などから事前に得られているので、比推定を用いることが調査コストの観点から望ましいと考えられたと推測される。ただし、調査員インタビューでは、必ずしも地点内悉皆調査のほうが楽かという、そうではないとの意見が聞かれた。国民生活基礎

調査の実施規模とコストをかんがみれば、上記の実施体制と比推定による母数推計を採用したことには一定の正当な理由が認められる。一方、それが抱える問題点も考慮しておく必要がある。

国民生活基礎調査が現在直面する最大の問題は、未回答による非系統誤差である。国民生活基礎調査が現在の地点内悉皆調査・比推定の方法を取る以上、地点内の回答率のばらつき・高い未回答率はバイアスを増幅させ、統計の妥当性・信頼性を脅かす最大の脅威となる。個人情報保護に関する意識の高まりから、調査環境が厳しくなっているとはいえ、国の基幹統計として、回収率の低下に対する積極的な対策を打つことが求められている。しかし、自治体のボランティアに頼る調査体制では、そうした対策を十分取れるか疑問である。

たとえば、近年回収率の低下が著しい所得票について、社会福祉事務所系統の民生委員など地域での信頼性が厚い調査員を通じることによって回収率が保たれている地域もある一方、個人情報保護に対する観点からすれば、顔を知った調査員に詳細な情報を渡すことに抵抗を覚える対象者もいると思われる。また、世帯・健康票と所得・貯蓄票について調査系統が異なり実施時期も分かれていることから、両者の連携が弱いと対象者は、「回答したと思ったら、また来られたので、もういやだ」という反応をとられやすい。実際、所得票の回収率は世帯票・健康票の回収率よりも低い。世帯・健康票の回収の際に「くじ引きの結果によって、また後続の調査をお願いするかもしれません」といった「事前予告」を伝えておくことが望ましいが、そのような系統的指示は出されていない模様である。調査員が対象世帯に調査協力を求めるにあたっては、公的統計調査であることを理由に、改正統計法のもとの「回答義務」を振りかざすだけでは、国民の納得を得られない。調査員が十分当該統計調査の内容を熟知しその意義を十分説明できるように、学習資料の作成や調査員に対する教育訓練を行うべきである。

未回収率が高い20代男性などを取り込み、総体として回答者負担を軽減するためには、訪

問・留め置き調査以外の幅広いオプション（郵送・インターネットなど）を提示し、回答者の便宜性の向上に努める必要もあろう。すでに平成22年国勢調査では試行的にインターネットによる回答が東京都などで実施されている。computer assisted program instrument (CAPI) の導入についても検討をはじめることが求められる。米国のHealth Interview Surveyも、調査項目の増加やサンプル数の増加に効率的に対応するため、1997年よりCAPIの全面導入に踏み切っている。調査負担を減らすために、一部調査項目については、乱数発生させてランダムにサブサンプルで尋ねることもできる。CAPIやインターネットの利用は、調査からデータ提供まで2年近くかかる現状の是正にも貢献することが期待される。ただし、異なる調査モードによる回答バイアスの影響を、事前に比較パイロット調査を実施し十分検証しておくことが必要であり、試験調査などの機会を用いた実験が求められる。また、対象者の中にはコンピューターを使うことで情報漏えいにつながるのではないかと危惧する人もいることから、異なるモードに対応して十分説明がつけられるように調査員訓練が併せて必要となる。

所得票については自計方式により対象を広げることが検討されている。これを機に、調査経路について、従前の保健所・福祉事務所の2系統から1系統化することを検討することが、実施自治体の負担や、個人情報保護などの観点から見ても望ましい。なおCAPI等の導入を考慮すれば、従前のように保健所・社会福祉事務所などの関係者による、事実上のボランティア調査員では、力量的に限界が出てくる。今後、全くの第三者による調査の実施（民間調査会社への委託なども含む）についても、調査精度や回収率・実施可能性（予算など）を包括的に検討する必要がある。回収率を高め精度を上げるために必要であれば調査規模の縮小なども検討し、また政策的有用性を考慮するならば、一部縦断化についても、検討すべき時期に入ったと思われる。

回収率の向上にあたっては、調査員が協力を

求めやすい環境を整備することが必要であり、これは国民生活基礎調査のみならず、公的統計調査すべてに関わる問題である。物理的に対象者との接触を図るうえで、近年問題となるのが集合住宅のゲートセキュリティである。マンションなどの管理者・管理人について、公的調査に対する協力義務を法的に課すなどの措置は早急に取りられるべきである。また集合住宅の住民管理組合などについても、管理規約に公的調査に対応する責任者の配置を定めるなど、管理規約に関する法律についても是正の余地がある。国民生活基礎調査の国民における周知の度合いが高くないことから、調査実施に先駆けて、広報活動を積極的に行うべきである。国民生活基礎調査のデータをもとに得られた研究成果やメディアリリースの内容を、調査員が提示資料として持ち歩けるようにすることも、協力要請をしやすくする重要なアイテムとなる。

以上のように非系統誤差の影響については、回収率を上げるためのあらゆる手段を尽くすことがベストであるが、一方で、そうした影響を考慮したうえで、より妥当と思われる統計指標の公表の在り方についても、セクションモデルやベイズ統計などの最新の統計手法を用いた技術的検討がなされるべきである。

（3）公開利用のための匿名化データセット

国民生活基礎調査の健康票の匿名標本データ提供は現時点では平成22年度内に実施される方向で準備が進められていたが、本稿を起こしている平成23年1月現在、統計委員会匿名化データ部会でまだ審議中である。健康関係の政府統計データとして初めてのケースとなる。基本的には国内における先行事業である総務省データ公開利用制度に倣うことになるが、サンプリングデザインの違い・ユーザー側の多様なニーズへの対応が必要であること、そしてなによりもユーザー側の成熟度と国民の理解を段階的に高めていくなかで、当初は慎重に進めざるを得ないと考えられる。

匿名化データセットを作成するうえで重要となるのが、リサンプリングやトップコーディン

グなどを用いて、個人同定を防ぐための秘匿処理を行うことである。しかし、国民生活基礎調査は、地区内の世帯悉皆調査を実施していることから、秘匿処理が確率論的抽出を行っている総務省調査の場合と異なり、より困難となっている。特に世帯を単位としてリサンプリングする場合は、いくら情報を絞っても世帯員の詳細な情報を伴うことから世帯を特定される危険性を避けるためには、リサンプリング率を下げて、地域情報を落とさざるを得ない。しかし利用者の便宜からすれば地域情報が欠落することはデータの利用価値を損なうことになる。さらに、いくら地域情報を落としても、拡大乗数を見れば、どの都道府県かを同定できてしまうために、乗数情報の提供にあたっては、本来の乗数とは異なる計算をせざるを得なくなり、母数推計が匿名化データセットと本体調査と一致しない可能性が出てくる。その点では、世帯ではなく、世帯員レベルでのリサンプリングを行い、限定的に世帯特性の集計情報を付加したデータセットを作成するほうが、地域情報なども含みやすい。

データ項目の削除、再カテゴリー化・リサンプリングなど施してマスクしても、データが詳細になるほど、個人を特定することは全く不可能ではない。したがって運用規定や罰則ルールなどの明確化、管理体制の組織化などを並行して進めることが重要である。公開利用制度の初期段階にあつては、利用者の成熟度や国民の理解の度合いなどとのバランスを取りつつ、利用便宜性よりはセキュリティを重視した制度設計になることは、やむを得ない。段階的な制度の見直しが必要になるであろう。その際、関連学会などの積極的な関与がキーになると思われる。当面、海外雑誌などに発表するようなレベルの研究については、改正統計法33条に基づく個票利用申請の枠を活用することとして、公開利用データとの役割を明確にすみ分けることも必要であろう。

最後に、総務省ワーキンググループで検討されている社会保障番号については、再度本格的議論が再開したことは歓迎されるべきと考える。

ユニークIDが存在することで、他統計（人口動態）や医療介護給付情報などの行政業務データとリンケージが張れるようになれば、回答者負担を減らし、より精度の高い客観的で信頼性の高い情報が得られることにもつながる。特に人口動態、医療給付情報や特定健診情報とのリンケージは、すでに全国データベースが構築されていることと併せれば、世帯の社会経済的状况と医療サービス利用の関係、その生命予後への影響などを明らかにすることが可能となり、厚生行政の政策策定上重要な知見をもたらすと期待される。

Ⅳ 考察ならびに結論

国民生活基礎調査大調査年のサンプリング・調査実施体制・匿名化データセットの課題について検討してきた。除外された地区についても無作為に抽出されたものであることから、抽出の「無作為性」については理論上保たれていることになるが、除外地区が3分の1に上ることから、全国代表性を十分確保できているかどうかについて、それを示すデータの公表が望まれる。少なくともデータの適切な解釈のためにも調査方法について国勢調査調査区などの範囲から抽出しているのかなど明記しておくことが必要であると考えられた。

調査の目的やコストをかんがみてサンプリングは設計されるものであり、一概に確率論的サンプリングがよいとはいえない。しかし、現在採用している地点悉皆・比推定の方法が持っている問題点について、統計の読者・利用者が十分理解したうえで、データの解釈が行えるような情報公開が強く求められる。また、最大の問題となる回収率の低下については、セレクションモデルなどによる統計的補正など適切な統計値の処理を施すことも早急に考慮すべきであるが、まず原則論として回収率向上のための努力を図ることが先決である。そのうえで、現行の実施体制がその努力を十分果たしうるのか、特に系統の一本化、調査員トレーニング、統計情報の周知徹底など、検討の余地がある。すでに

系統の見直しやトレーニング、集合住宅向けの説明文書の発出など、検討ないし一部実施に向けた動きが見られている。

国民生活基礎調査の匿名化データの作成にあたっては、世帯レベル・世帯員レベルでのリサンプリング、項目制限・トップコーディングなどの秘匿処理を施しても、完全な匿名化は困難であることから、管理運用ルールの明確化や関連制度・組織の充実が求められる。制度運用の初期においては、当面セキュリティ対策を重視しつつも、利用者側の成熟度に併せて、制度の見直しを図り続けることが肝要であろう。またそのためにも、本統計調査が国民の共有財産であるとの公益性を重視する観点から、研究者、政策関係者だけでなく、一般市民への理解・啓蒙を図りつつ、開かれた議論を段階的に踏みながら進めていくことが求められるであろう。

国民生活基礎調査は、世帯面の基幹統計として貴重なデータを発信しつづけてきたし、これからの機能は高く期待され続けるだろう。一方、変化する調査環境・ユーザーとそのニーズの多様化、変化のスピードに対応するには、統計の継続性を確保しつつも、常にその内容や実施方法について大胆な変革を見通した議論を継続していく必要がある。その作業は、対象者である国民に対しても、またデータの利用者に対しても「国民の共有財産としての統計」としての正当性を明確に説明できるように、常に準備しておくことに他ならない。

謝辞

本稿の内容は（財）厚生統計協会による平成21年度厚生統計研究委託金研究事業「国民生活基礎調査の新体系の構築に関する調査研究」（主任研究者：橋本英樹）に基づく。本研究事業の実施に当たり厚生労働省大臣官房統計情報部社会統計課国民生活基礎調査室の上田響室長はじめ、関係各位から資料提供などご協力をいただいた。併せて御礼申し上げたい。なお、本稿の内容は上記研究成果に基づいた筆者の見解を示したものであり、研究班メンバー、国民生活基礎調査室、ならびに厚生統計協会のものでは

はない。事実誤認など残っていれば、すべては著者個人の責任に属する。

平成21年度 厚生統計研究委託 研究報告書
国民生活基礎調査の新体系の構築に関する調査研究

川上 憲人 東京大学大学院医学系研究科精神保健学分野教授

山岡 和枝 国立保健医療科学院技術評価部室長

近藤 尚己 山梨大学大学院医学工学総合研究部講師

橋本 英樹（代表研究者）東京大学大学院医学系研究科臨床疫学・経済学分野教授

注1 平成21年度事業では、他にメンタルヘルス尺度（K6）の系統的欠損（自覚的健康状態が悪いもの、高齢層、男性、低所得層で欠損率が高い）による都道府県表章を行った場合の推計バイアスの問題とサンプルセレクションモデルを用いた補正の可能性、国民生活基礎調査の後続調査として国民健康栄養調査をリンケージした場合の、サンプルバイアスの問題（若年層欠損、高齢層の過剰表出）などについても検討したが、別途稿を起こすこととした。

文 献

- 1) 橋本英樹. 今後の国民生活基礎調査の在り方についての一考察；健康票を中心として 厚生指標 2009；56（1）：1-8.
 - 2) 橋本英樹. 今後の国民生活基礎調査の在り方についての一考察；第2報 厚生指標 2010；57（1）：1-7.
 - 3) 平成21年度厚生統計研究委託金研究事業報告書「国民生活基礎調査の新体系の構築に関する研究事業」（研究代表者：橋本英樹）2010.5.
 - 4) 平成20年度厚生統計研究委託金研究事業報告書「国民生活基礎調査の新体系の構築に関する研究事業」（研究代表者：橋本英樹）2009.4.
- ※文献1）2）の抄録については、当協会HPでも公開しています。