OCRによるがん登録

川村　範夫*

はじめに

岡山県では1992年から胃・大腸（結腸・直腸）、肺、乳房、子宮の5部位のがん（以下、5がんという）について、がん登録を開始した。これが軌道に乗った1996年から対象を全がんに拡大した。この機会に、それまで手入力に頼っていたデータ入力を全面的にOCR利用に切り替えた。移行から1年半が経過したが、登録業務は順調に行われている。OCRの導入は、単に入力作業の効率化にとどまらない、色々なメリットがあることも明らかになった。

方法

導入したOCRは沖電気製の手書きOCRシステムIR620VJである。導入経費は、OCR装置及び添付ソフトに379万円、システム設計に192万円、届出票印刷費172万円、届出票配布のための送料39万円である。

がん登録届出表はB4版2枚で構成され、1枚目には記入上の注意、診断名・発生部位、組織診断名の主要なものについてICD-10コード対比表が印刷してある。2枚目が記入様式で、記入項目は39項目からなっている。発生部位、診断名はICD-10コードによる記入を原則としているが、一覧表にない場合には、自由記入欄に日本語で記載するようにした。

届出票は岡山県医師会情報センターに集められ、記入不備をチェックした後、1回100枚単位でOCRに読み込む。100枚の読み取りには約15分、1枚平均9秒を要する。OCRが読み込んだデータには、届出票の記載内容をスキャナー機能で画像データとしてそのまま取り込んだものと、OCRの変換ソフトにより文字データに変換したものの両方を、モニタ画面に上下に並べて表示し、それを見比べながら確定入力していく。この処理には100枚あたり約150分、1枚平均1分半かかるが、オペレータの仕事は入力作業から点検作業へと質的に変わり、ストレスは非常に軽減される。

確定入力したデータは、CSV形式のテキストファイルに加工されて、出力される。がん登録データベースの中へ、このテキストファイルを取り込むことにより、データが追加されてデータ入力業務は完了する。

OCRでの読み取り精度を上げるために、いろいろな試みをした。システム設計に当たっては、次のようなオプションを採用している。(1) 人名、地名の漢字読み取りには、人名辞書、地名辞書を使用する。変換には学習機能も利用する。(2) 行政単位の市町村コード、保健所コードは自動生成する。(3) 生年月日を和暦から西暦に変換した後、診断日における満年齢を計算し、診断日との関係の論理チェックに役立てる。(4) 届出票に記載された医療機関名と、同じく記載された医療機関コードから医療機関辞書によって変換した医療機関名とを並列表示し、医療機関コードのチェックに使用する。(5) 人名、地名の漢字を新字体（略字）に統一して、出力する。

OCRは予め定められた位置に記載された画

*岡山県医師会理事

〒703 岡山市古京町1丁目1-10　岡山県医師会情報センター　TEL 086-272-7744　FAX 086-272-3356

像データを読み取って文字に変換するもので、もし書き間違えたものを線で消して欄外に記載されるとOCRでは読み取れず、いちいち原票から再入力する必要が出てくる。このような事態を避けるため、書き間違った場合には消しゴムで消して書き直すことを求めている。

一方で、OCRにとっての泣き所は、読み取り画面の汚れで、鉛筆で書いた届出票を大量に読み込むと、どうしても鉛筆の汚れが溜まって読み取りエラーの原因になり、定期的に読み取り面を清掃する必要がある。病歴士が書いてくれる一部の医療機関には、個別に、ボールペンの使用をお願いしている。

がん登録データベースへデータを取り込んだあと、このデータが新規登録であるかどうかを調べる。既登録データがあれば、その全部を新既登録データと合わせて画面表示し、キーとなる項目の内容を比較して、どちらをマスターのデータとするかを決め、マスターとならない方のデータの診断方法、治療方法、受診の動機等の内容を、マスター側の予備フィールドに追加する。診断日は原則として、最も早い日付を採用する。診断方法、治療方法、受診の動機等については、最初に診断した医療機関からのデータをマスターデータの予備フィールドに転写する。既登録患者の死亡年月日、死因等の死亡情報が得られたときは、これをその患者の全ファイルの該当欄に転写する。その際、死亡年月日がその届出票によって報告されたものか、登録室で転写したものかの区別がつくような方法をとっている。

成績

5がんを対象とした、1993～95年の3年間の平均では、DCO割合は15.1%であったが、全がんが対象となりOCR用の届出票に変わった1996年のDCO割合は、現時点で15.2%となり、それまでとほぼ同じ精度の届出が得られている。このことから、OCRによるがん登録が医療機関に受け入れられて、定着したものと考えている。

考察

文字をスキャナー機能で画像データとして取り込んだものを、同時に画面に表示することで、点検のため、いちいち原票に目を移さなくてもよいようになった。このことは、このシステムの特徴の一つであるが、このCSV形式のファイルに名前を付けて保存しておけば、原票をずっと保管しなくてもよくなり、保管スペースが足りなくなったときには、大きなメリットになる。

システム設計に当たって発注した前述の5種類のオプション仕様は、何れもOCR読み取りの精度向上、がん登録データの論理チェック等に役立っている。

人名、地名の漢字を新字体（略字）に統一してOCRが変換するようにしたことで、検索の効率は著しく向上した。これは、OCRを採用したことの2次効果である。

我々が最初にOCRでの登録を計画した段階では、届出に対する協力を得るためには出来るだけシンプルな様式にすること、それまでの5がんの登録とデータの互換性をもたせること、の2点を最優先にした。したがって、自由記入欄は極力減らしたが、自由記入欄があった方が医療機関との意志の疎通にもなり、照会などの件数も減らすことが出来るのではないかと考える。

集計項目も、文字記入欄はさておき、単に「レ」マークを入れる項目は100%正確に変換されるので、診断方法、治療方法等の選択項目の数が増えても困らないところが、OCRの強みである。OCRの長所をうまく利用できるような、届出票の設計を考えたい。

結論

OCR によるがん登録を始めて1年半になるが、登録は順調に行われている。今後は、現行の届出票について、OCR の特性やパソコン性能向上のメリットを取り入れた、限られたスペースの中に更に効率のよい、親しみやすい様式をもった届出票に改定したいと考えている。

文献

藤本伊三郎編：地域がん登録の手引き．厚生省がん研究助成金「がん予防におけるがん登録の役割に関する研究」班（主任研究者 福間誠吾），千葉, 1986.