Journal of Biomedical Informatics 60 (2016) 328–333

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Predictive analysis for identifying potentially undiagnosed post-stroke spasticity patients in United Kingdom

Andrew Paul Cox [a,*], Mireia Raluy-Callado [a], Meng Wang [a], Abdel Magid Bakheit [b], Austen Peter Moore [c], Jerome Dinet [d]

[a] Evidera, Metro Building, 6th Floor, 1 Butterwick, London W6 8DL, United Kingdom
[b] Moseley Hall Hospital, Alcester Road, Birmingham, West Midlands B13 8JL, United Kingdom
[c] Walton Centre NHS Foundation Trust, Lower Lane, Fazakerley, Liverpool, Merseyside L9 7LJ, United Kingdom
[d] Ipsen Pharma, 65, quai Georges Gorse, 92650 Boulogne Billancourt Cedex, France

ARTICLE INFO

Article history:
Received 3 April 2015
Revised 11 February 2016
Accepted 21 February 2016
Available online 27 February 2016

Keywords:
Spasticity
Stroke
Machine learning
Random forest
Electronic medical records

ABSTRACT

Purpose of the research: Spasticity is one of the well-recognized complications of stroke which may give rise to pain and limit patients' ability to perform daily activities. The predisposing factors and direct effects of post-stroke spasticity also involve high management costs in terms of healthcare resources, and case-control designs are required for establishing such differences. Using 'The Health Improvement Network' (THIN) database, such a study would not provide reliable estimates since the prevalence of post-stroke spasticity was found to be 2%, substantially below the most conservative previously reported estimates. The objective of this study was to use predictive analysis techniques to determine if there are a substantial number of potentially under-recorded patients with post-stroke spasticity.
Methods: This study used retrospective data from adult patients with a diagnostic code for stroke between 2007 and 2011 registered in THIN. Two algorithm approaches were developed and compared, a statistically validated data-trained algorithm and a clinician-trained algorithm.
Results: A data-trained algorithm using Random Forest showed better prediction performance than clinician-trained algorithm, with higher sensitivity and only marginally lower specificity. Overall accuracy was 75% and 72%, respectively. The data-trained algorithm predicted an additional 3912 records consistent with patients developing spasticity in the 12 months following a stroke.
Conclusions: Using machine learning techniques, additional unrecorded post-stroke spasticity patients were identified, increasing the condition's prevalence in THIN from 2% to 13%. This work shows the potential for under-reporting of PSS in primary care data, and provides a method for improved identification of cases and control records for future studies.

1. Introduction

Stroke remains a devastating neurological disease, often causing severe physical impairment or death [1]. In the United Kingdom (UK), it is one of the top three causes of death in the population and the largest cause of adult disability, with approximately 110,000 strokes per year in England alone [2].

This considerable disease burden reflects the wide range of complications associated with stroke. Among these is spasticity, a combination of symptoms and clinical signs that follow lesion formation in sensorimotor brain areas and tracts in the central nervous system [3], and which are typified by increased muscle tone during active or passive movements. More specifically, the underlying mechanism in spasticity has been defined as a velocity-dependent hyperexcitability of muscles to stretch, characterized by exaggerated tendon reflexes, increased resistance to passive movement, and hypertonia, resulting from loss of upper motor neuron inhibitory control [4].

The related functional changes in the limbs affected by post-stroke spasticity (PSS) may give rise to pain and limit patients' ability to stand, walk, eat, care for themselves, work, or perform other daily activities. Potential consequences therefore include dependency; complications such as falls and fractures [5]; decline in

Abbreviations: GP, general practitioner; PSS, post-stroke spasticity; QOF, Quality and Outcomes Framework; THIN, The Health Improvement Network; UK, United Kingdom.

* Corresponding author. Tel.: +44 (0)208 576 5006.
E-mail address: andrew.cox@evidera.com (A.P. Cox).

http://dx.doi.org/10.1016/j.jbi.2016.02.012
1532-0464/

the patient's self-esteem and health-related quality of life [6]; and a significant burden on caregivers of stroke survivors [3].

Such outcomes are particularly significant given that the development of spasticity may relate to, and so might compound, the severity of neurological impairment caused by the stroke. For example, a prospective longitudinal study concluded that predictors of the development of PSS included a severe degree of paresis and hemihypesthesia at stroke onset [7]. Also, another study showed that National Institutes of Health Stroke Scale (NIHSS) scores were higher (i.e., indicated worse stroke-related impairment) in patients with spasticity compared to those without [8]. Furthermore, another study identified that key risk factors associated with the development of spasticity included lower Barthel Index scores (indicating greater dependence), a severe degree of paresis, stroke-related pain, and sensory deficits [3]. Other possible risk factors for PSS include early arm and leg weakness, left-sided weakness, and early reduction in activities of daily living (as well as a history of smoking) [9].

The predisposing factors and direct effects of PSS make it unsurprising that the condition carries high treatment costs [5]. Evidence of this burden includes the suggestion from one study that direct costs for stroke survivors with spasticity are around four times those for individuals without PSS during the first year after the event ($84,195 vs. $21,842 in 2003) [10]. Also, the same research estimated that, assuming 20% of stroke survivors experience spasticity, direct costs for stroke with spasticity would be approximately $4.2 million per 100,000 inhabitants per year [10]. Spasticity may thereby be a significant contributor to the huge costs of stroke, which are driven largely by expenditure associated with management of the disabilities common in stroke survivors.

Such estimates of the socioeconomic problems posed by PSS make it crucial to have reliable information on the prevalence of the condition. In reality, however, there is no such epidemiological clarity. While there is general agreement that spasticity is a common after stroke, the occurrence of PSS has been difficult to quantify. Reasons for this include the heterogeneity across studies in methods for assessing spasticity, the lack of published population-based data on spasticity, and to some extent, the absence of a consensus on the diagnosis of spasticity in general, which reflects the complexity and the diversity of phenomena associated with the condition [11]. Key insights into such issues have come from Wissel et al. [3], whose systematic literature review of studies on the prevalence of PSS concluded that this ranged from 4% to 46%, with a similarly broad range of patients having disabling spasticity (2–33%). In considering these variations, it is important to note that the timeframe of analyses also differed widely across the studies, from 2 to 10 days to 18 months after stroke. However, even among the studies that assessed patients specifically at 12 months, the proportion with spasticity still ranged from 17% to 38% [3]. The authors suggested that this might be because, while the diagnosis of spasticity was mainly based on the Modified Ashworth Scale, the score used to define spasticity on this scale varied between studies. Also, the number of patients included was rather small because each study was based on stroke survivors from a single hospital unit.

The absence of clear evidence from published literature on the prevalence of PSS invites questions about whether exploration of large population databases might help to address this key knowledge gap. An obvious candidate for this approach in the UK is The Health Improvement Network (THIN) database, a population-based primary-care collection of information about each patient's visits to their general practitioner (GP). Specifically, this includes diagnosis, symptoms and other relevant details, which are coded by the GP using Read codes, a standardised clinical coding system. As a result, the THIN database offers key advantages as an evidence source, since GPs are often the first point of contact when patients seek medical attention and act as gatekeepers of the health system. Accordingly, this dataset is routinely used for health economics and epidemiological studies, and we had previously selected it for conducting a study on costs post-stroke that compared patients who developed spasticity with those who did not. Of note, our exploratory analysis in THIN to inform the cost study revealed that the proportion of patients with a record of spasticity, as identified by Read codes within 12 months after a stroke event, was only around 2% [unpublished data], which is substantially below the most conservative estimates of 17–38% cited in the published literature [3]. One inference, therefore, is that spasticity could be under-recorded in the UK primary care setting. This might happen if, for instance, less severe cases do not receive a diagnostic code for spasticity, or if GPs treat spasticity as a symptom of the stroke itself, or, again, it could stem from the lack of a uniform definition of spasticity that can be used across clinical research settings.

A large number of undiagnosed patients with PSS in THIN would make it difficult to perform studies of this condition, especially to determine differences in costs for patient with and without PSS. A method that can identify PSS cases in this type of source data would allow less biased studies of the condition. With this in mind, the objective of the current study was to use machine learning and expert opinion to determine whether there is a sizeable number of potential PSS cases that do not receive a diagnostic code, and to identify cases and controls for a subsequent study of costs and resource utilization.

2. Materials and methods

This study used data from the THIN UK primary care database, which covers a population representative of the UK, with 3.7 million registered patients as of September 2011. Available information included anonymised medical records with demographic characteristics, prescriptions of medication issued by GPs, and medical diagnoses from 532 primary care practices covering over 5% of the UK population. The THIN data was selected for capturing all primary care elements of patients' management and for its representativeness to the UK population. Although stroke events are typically managed initially in the hospital setting, related, longer-term, information in the primary care records was expected to be adequately recorded, given that stroke and transient ischaemic attack are part of the indicators of the Quality and Outcomes Framework (QOF) [12], a financial incentive scheme used in the UK National Health Service since April 2004 to reward GPs for good practice. The study period was chosen to post-date the introduction of the QOF so as to help ensure consistency of recording across all time periods. The study population included adult patients who had at least one record of a stroke event as identified by the appropriate Read codes [Table A1, Supplementary material]) between 1 January 2007 and 31 December 2011 and who were registered with a GP practice that contributed to THIN. Read codes are a coded thesaurus of clinical terms and have been used in the United Kingdom's National Health System (NHS) since 1985 and include sociodemographic data, medical signs and symptoms, diagnostic and laboratory tests, referrals, and prescriptions of any medications issued by the GPs and have been validated for epidemiological research. No further inclusion or exclusion criteria were applied.

Full ethical approval was granted for this work (SRC Protocol: 14-025).

2.1. Classification of stroke events

All stroke events between 1 January 2007 and 31 December 2011 were included in the analyses.

The data consisted of records of stroke events and the twelve-month period before and after the stroke event. Within this study