MiniReview

Templates for Analysis of Individual-Level Prescription Data

Jesper Hallas\(^1\) and Henrik Støvring\(^2\)

\(^1\)Department of Clinical Pharmacology, and \(^2\)Research Unit of General Practice, University of Southern Denmark, Odense, Denmark

(Received June 20, 2005; Accepted October 5, 2005)

Abstract: The advent of large population-based prescription databases has enabled us to study drug use with the individual user as our unit of analysis. This review presents three non-specific analytic templates that may be applied to individual-level prescription data. The ratio of prevalence odds to incidence rate can estimate the average duration for drug use. Limitations and pitfalls are discussed. Although it should be cautiously interpreted, it provides a reasonable ranking of drugs with respect to their retention in users. The Lorenz curve is an analytic tool to express skewness in drug use. It shows the proportion of drug use that is accounted for by percentiles of drug users, ranked according to their volume of drug intake. It may express the extent of heavy users as well as sporadic small-volume users and may, for example, be used to screen for an unsuspected abuse potential of a drug. The waiting-time distribution is a frequency distribution of first occurrences of drug use within a time-window. It forms the basis for a theoretical model for robust estimates of prevalence and incidence rate. On an intuitive level, it displays visual correlates of epidemiological prescribing parameters such as period prevalence, point prevalence, incidence rate, duration, prescription renewal rate, relapse of treatments and seasonality. Each measure may be incorporated into an integral matrix that reflects various traits in utilization of every drug or drug class, thereby possibly finding abnormalities that suggest sub-optimal prescribing.

The Scandinavian countries have a long tradition for gross volume statistics, whereby the drug use on the entire population level is described without knowledge about the distribution among users (Lee & Bergmann 1999). These gross volume statistics allow us to follow crude measures of the population's drug use and morbidity over time (Sartor & Walckiers 1995), but rarely allow any conclusion about the quality of prescribing.

A vast improvement lies in the possibility of analysing individual-level prescription databases and thereby provide a characteristic of how each drug is used (Hallas & Nissen 1994). This review presents three non-specific analytics templates that may be applied to any drug or class of drugs, and may possibly be used to screen for anomalies in prescribing patterns; the prevalence/incidence rate relationship, the Lorenz curve, and the waiting-time distribution.

The data used are from the Odense University Pharmacoepidemiologic Database, which covers half a million inhabitants of the County of Funen (Gaist et al. 1997). It has complete capture of all reimbursed prescriptions, as well as a demographic module to account for migrations and deaths for all County of Funen residents, drug users as well as non-users. Data from 2003 were used and categorised according to the 2003 version of the anatomic therapeutic chemical classification (ATC) and defined daily dose (DDD) (WHO 2003). Insulin is used as a recurrent example in the calculations below, as it is a prototype of a drug for chronic use and is given only for diabetes.

Prevalence: incidence rate relationship

It is surprisingly difficult to determine the point prevalence of use for a drug from prescription data. The underlying problem is that prescription data are dispersed in time and that many users show an irregular pattern of periods on and off the drug. It is difficult to distinguish between intermittent use and continuous use with irregular redeeming of prescriptions. Most methods entail assigning a period of usage to each prescription record (Sartor & Walckiers 1995; Mantel-Teeuwisse et al. 2001; Støvring et al. 2003). The point prevalence on a given day is then estimated straightforwardly from the number of persons who have a prescription whose period of usage crosses this date. There is a number of variants; assigning a constant period to each prescription, assuming a constant use of, e.g., 1 defined daily dose (DDD) or assuming other fixed amounts (Bjerrum et al. 1997). If such data are available, it is also possible to estimate the period of usage on the basis of the prescribed dose or the legend duration (van Staa et al. 1994). Usually, authors using such methods make a minor allowance for
non-compliance (Lau et al. 1997). Finally, complex statistical models such as exponential smoothing may be applied to the prescription data (Chatfield et al. 2001). Similar variations may arise from using different denominators: people living in a specified area or users of a specific health care provider.

If we assume a fixed daily intake of 1 DDD for insulin, equivalent to 40 IU, the residents of Funen County had periods of usage that crossed July 1, 2003. According to census data, there were 473,471 residents, yielding a point prevalence of 7.17 per 1000 for this day.

The incidence rate is determined straightforwardly as a count of new users after a certain drug-free run-in period divided by the person-time at risk. Our demographic module allows us to distinguish between incident users and prevalent users that move into the county, both of which would appear as new users. It also allows us to establish the denominator, person-time at risk, accurately. For example, we may assume that persons who redeem insulin prescriptions during the last two months of 2003 without having redeemed insulin prescriptions before that and who are not immigrants, are in fact incident users. If we employ a ten month run-in period, we arrive at 89 new users during a 61 day period November-December 2003. The corresponding incidence rate is 1.13 per 1000 person-years, calculated as

\[
\frac{89}{61 \times 365} = \frac{89}{3,395}. 
\]

### Table 1.

Analyses of epidemiological measures of drug use for the major drug classes. Data from the County of Funen 2003. The background population is 473,471 persons as of January 2003.

<table>
<thead>
<tr>
<th>Drug class (ATC-code)</th>
<th>Count of users</th>
<th>Therapeutic intensity, DDD/1000/d</th>
<th>Average age, years</th>
<th>Percentage Male users</th>
<th>Prevalence /1000</th>
<th>Incidence rate /1000 person-years</th>
<th>Duration Years</th>
<th>Lorenz 1-percentile</th>
<th>Lorenz 50-percentile</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ulcer drugs (A02B)</td>
<td>29412</td>
<td>29.00</td>
<td>59.31</td>
<td>44.31</td>
<td>23.59</td>
<td>30.87</td>
<td>0.78</td>
<td>5.96</td>
<td>89.76</td>
</tr>
<tr>
<td>Insulin (A10A)</td>
<td>5045</td>
<td>11.71</td>
<td>56.63</td>
<td>54.99</td>
<td>7.17</td>
<td>1.13</td>
<td>6.38</td>
<td>4.67</td>
<td>75.82</td>
</tr>
<tr>
<td>Oral antidiabetics (A10B)</td>
<td>7541</td>
<td>15.44</td>
<td>65.75</td>
<td>52.55</td>
<td>9.92</td>
<td>3.52</td>
<td>2.84</td>
<td>4.52</td>
<td>81.79</td>
</tr>
<tr>
<td>Oral anticoagulants (B01AA)</td>
<td>5111</td>
<td>5.37</td>
<td>70.21</td>
<td>58.38</td>
<td>5.04</td>
<td>3.57</td>
<td>1.42</td>
<td>3.61</td>
<td>74.14</td>
</tr>
<tr>
<td>Digoxin (C01A)</td>
<td>5935</td>
<td>7.15</td>
<td>77.40</td>
<td>47.97</td>
<td>6.88</td>
<td>2.38</td>
<td>2.91</td>
<td>3.32</td>
<td>73.50</td>
</tr>
<tr>
<td>Diuretics (C03)</td>
<td>44895</td>
<td>106.84</td>
<td>69.31</td>
<td>34.04</td>
<td>64.74</td>
<td>25.34</td>
<td>2.73</td>
<td>10.02</td>
<td>80.29</td>
</tr>
<tr>
<td>Betablockers (C07)</td>
<td>24676</td>
<td>25.40</td>
<td>64.36</td>
<td>45.02</td>
<td>23.11</td>
<td>15.32</td>
<td>1.54</td>
<td>4.45</td>
<td>80.13</td>
</tr>
<tr>
<td>Calcium blockers (C08)</td>
<td>20772</td>
<td>45.12</td>
<td>67.70</td>
<td>46.86</td>
<td>32.35</td>
<td>10.23</td>
<td>3.27</td>
<td>3.09</td>
<td>73.96</td>
</tr>
<tr>
<td>ACE inhibitors (C09)</td>
<td>35149</td>
<td>86.45</td>
<td>65.92</td>
<td>48.53</td>
<td>56.28</td>
<td>17.02</td>
<td>3.50</td>
<td>4.33</td>
<td>75.53</td>
</tr>
<tr>
<td>Statins (C10AA)</td>
<td>17365</td>
<td>47.43</td>
<td>64.41</td>
<td>54.24</td>
<td>27.25</td>
<td>12.06</td>
<td>2.32</td>
<td>5.34</td>
<td>75.89</td>
</tr>
<tr>
<td>Thyroid replacement (H03A)</td>
<td>7604</td>
<td>9.90</td>
<td>61.26</td>
<td>14.18</td>
<td>8.72</td>
<td>2.40</td>
<td>3.67</td>
<td>7.28</td>
<td>72.22</td>
</tr>
<tr>
<td>Antithyroid drugs (H03B)</td>
<td>2470</td>
<td>2.93</td>
<td>65.40</td>
<td>16.11</td>
<td>2.51</td>
<td>1.46</td>
<td>1.73</td>
<td>8.90</td>
<td>75.29</td>
</tr>
<tr>
<td>Antibiotics (J01)</td>
<td>147967</td>
<td>12.10</td>
<td>40.50</td>
<td>42.01</td>
<td>11.02</td>
<td>231.65</td>
<td>0.05</td>
<td>9.68</td>
<td>78.85</td>
</tr>
<tr>
<td>NSAIDs (M01A)</td>
<td>72398</td>
<td>36.58</td>
<td>51.53</td>
<td>42.95</td>
<td>30.84</td>
<td>98.06</td>
<td>0.32</td>
<td>10.02</td>
<td>91.40</td>
</tr>
<tr>
<td>Opioids (N02A)</td>
<td>26700</td>
<td>16.14</td>
<td>61.31</td>
<td>41.46</td>
<td>11.12</td>
<td>33.20</td>
<td>0.34</td>
<td>19.29</td>
<td>96.48</td>
</tr>
<tr>
<td>Antiepileptics (N03A)</td>
<td>6557</td>
<td>9.28</td>
<td>52.34</td>
<td>48.57</td>
<td>6.32</td>
<td>3.54</td>
<td>1.80</td>
<td>6.90</td>
<td>88.13</td>
</tr>
<tr>
<td>Antipsychotics (N05A)</td>
<td>10046</td>
<td>9.75</td>
<td>59.03</td>
<td>41.20</td>
<td>6.37</td>
<td>7.38</td>
<td>0.85</td>
<td>12.79</td>
<td>95.12</td>
</tr>
<tr>
<td>Antidepressants (N06A)</td>
<td>31762</td>
<td>50.81</td>
<td>56.78</td>
<td>34.57</td>
<td>37.04</td>
<td>21.69</td>
<td>1.77</td>
<td>5.14</td>
<td>82.69</td>
</tr>
<tr>
<td>Asthma drugs (R03)</td>
<td>43548</td>
<td>64.34</td>
<td>39.43</td>
<td>47.57</td>
<td>32.83</td>
<td>43.62</td>
<td>0.78</td>
<td>11.72</td>
<td>94.33</td>
</tr>
<tr>
<td>Inhaled beta-agonists (R03A)</td>
<td>29431</td>
<td>33.28</td>
<td>44.75</td>
<td>47.13</td>
<td>21.95</td>
<td>26.18</td>
<td>0.86</td>
<td>11.94</td>
<td>90.67</td>
</tr>
<tr>
<td>Inhaled corticosteroids (R03B)</td>
<td>20084</td>
<td>24.59</td>
<td>43.64</td>
<td>47.44</td>
<td>18.42</td>
<td>14.73</td>
<td>1.27</td>
<td>7.00</td>
<td>86.59</td>
</tr>
</tbody>
</table>

ATC=Anatomic Therapeutic Chemical classification. DDD=Defined Daily Doses.
the available sample to the population, who have not recently immigrated.

- Prevalence is to be understood with respect to current use, and so incidence rate captures both first-time-ever use and re-initiation of treatment. As a consequence, the estimate of duration is an estimate of current treatment spell duration, not of disease duration.
- It is biased towards lower values by heterogeneity in prevalence and incidence rate among drug users (Alho 1992).
- It is biased in non-equilibrium states. If incidence rate is rising, duration estimates are biased towards lower values.
- Its interpretation may be problematic. A claim that the average duration of insulin use was 6.4 years in 2003 can not be taken literally. Its interpretation is equivalent to that of life-table techniques: the average life-time for a newborn child experiencing the same age-specific mortality as the present population.

Notwithstanding its limitations, the duration derived from such calculations yield a satisfying ranking of drugs with respect to retention in users. Among the major therapeutic groups, insulin, ACE inhibitors and thyroid replacement therapy have the highest durations, while the lowest are found for antibiotics, NSAIDs and opioids (table 1). An example of a drug prescribing problem that might have been captured by a low prevalence:incidence rate ratio is the remarkably short duration of statin use that has been observed repeatedly over the world (Avorn et al. 1998; Larsen et al. 2000). The estimated duration by our analysis is 2.32 years. Although it has improved from previous analyses, this is still far too low to achieve full benefit from the drug. When finding such low duration by an prevalence:incidence rate ratio, more robust and elaborate analyses should be undertaken, such as survival analyses (Larsen et al. 2000; Rosholm et al. 2001).

Lorenz curve

Lorenz was a Swiss economist who used this approach around 1900 to describe skewness in income. In its original form, the x-axis represents a given proportion of the population ranked with respect to their income, the y-axis represents the equivalent proportion of the income that would be accounted for by this part of the population. Lorenz ranked the population with low income first, yielding a convex graph (Lee 1997). For the purpose of describing drug users, it is more intuitive to rank the heavy users first, thereby obtaining a concave graph. If all users take similar doses, then the curve would be a diagonal line. If few drug users accounted for the entire drug volume, then the curve would approach the upper left corner. The actual graph for insulin is shown in fig. 1.

We have suggested that we could use the 1-percentile of users to express the extent of heavy use (Hallas & Nissen 1994). For insulin, 1% of users accounted for 4.7% in 2003, which is a comparatively low value, suggesting that heavy users do not abound for insulin. Other drug classes with low values are calcium blockers, digoxin and oral anticoagulants (table 1). The corresponding figure for opioid analgesics is 19.3%, suggesting the existence of some heavy users of opioid. Other drugs with high 1-percentiles are neuroleptics and asthma drugs.

The 50-percentile of opioid users accounts for 96.5% of the opioid consumption. In other words, the 50% of users that consume the least account for only 3.5% of use. For opioids, along with a group of heavy users there is a large group of sporadic small-quantity users. Thus, high values of the 50-percentile suggest that sporadic use is highly prevalent (Hallas & Nissen 1994). This pattern is found for nitrates, asthma drugs and neuroleptics, while low values of the 50-percentile, suggesting few sporadic users is found for calcium blockers, digoxin and thyroid replacement. Both the 1- and 50-percentile can be calculated for the entire spectrum of drugs and drugs classes for the entire population by a simple algorithm.

Not surprisingly, the 1- and 50-percentiles correlate strongly. One reason is a simple mathematical property of the method. If the 1- and 2-percentiles are very high, then there are limits to how low the 50-percentile can be. There are, however, also clinical reasons why such correlations exist, one being the toxicity of the drug. Very toxic drugs can not have heavy users, and very toxic drugs are not prescribed sporadically.

Another closely related measure is the quantile plot, which is available in several statistical packages. Here, the actual quantities consumed by persons located at the 1-percentile, 2-percentile etc., are plotted (Altman & Bland 1994). For example the quantity of insulin consumed by a person located at the 10-percentile is 385 DDD/year. It has the advantage that the y-values for each percentile is not dependent on the values for the lower percentiles, contrary to the Lorenz curve. On the other hand, an experienced clinician’s judgement is needed to interpret each of the values. It is therefore not as directly amenable to automatic screening for skewness as the Lorenz curves.
In a theoretical extension of the Lorenz curve, Gini has proposed a single measure to express skewness, the Gini coefficient (Brown 1994). Obviously, if the skewness is expressed in one single compound measure, we lose the specific insight provided by the 1- and 50-percentiles, i.e. specific information about the heavy users and sporadic users.

An example of a drug use pattern that was described by a Lorenz-like measure was heavy use of sumatriptan. It was observed that a small fraction of sumatriptan users consumed inordinate amounts, a finding which eventually lead to the discovery of an abuse potential (Gaist et al. 1994 & 1998).

**Individual-level drug utilization statistics**

Other measures have not been mentioned, for which the calculation is straightforward, e.g., the one-year prevalence, the therapeutic intensity (number of DDD/1000 inhabitants/day), the average age of users and the sex distribution. Each of these measures along with the point-prevalence, the incidence rate, the duration, and the 1- and 50-percentile in the Lorenz curves may be incorporated into an integral matrix that describes utilization of every drug and drug class at all levels of the ATC-system (Hallas & Nissen 1994). The algorithm is fairly simple, and the data for the County of Funen for a whole year may be processed in as little as 20 min. on a personal computer.

Such tables are produced annually. Since each drug or drug class is measured by several indices that reflect various traits in their utilization pattern, it is possible to achieve a good impression of their use and possibly to find abnormalities that suggest suboptimal prescribing. Obviously, findings that arouse the interest of researchers should be followed by other, more specific analyses.

**Waiting-time distribution**

The waiting-time distribution is a graphical approach to analyses of prevalence/incidence rate relations (Hallas et al. 1997). It is more complicated than the other measures, but also very informative. The waiting-time distribution is a frequency distribution of first occurrences of drug use within a time-window. If we graph all persons who redeemed insulin prescriptions in 2001–2003 according to the appearance of their first prescription within that period, it would generate a graph as in fig. 2. Intuitively, it shows how long we have to wait before an insulin user appears, hence the term waiting-time distribution. In the beginning of the curve, continuous, prevalent users dominate, but after a few months all prevalent insulin users have been captured, and the curve is dominated by incident users. Thus, the graph has two distinct components, provided by the incident and the prevalent users. For insulin, the ratio of prevalent to incident users is high. There is no seasonality and no sign of relapse. This is a pattern typical of a chronic treatment. The principle may be extended to estimating incidence rate and prevalence of other health phenomena than drug use, for example episodes of care in general practice (Schroll et al. 2004).

One of the authors has developed parametric methods to estimate point prevalences and incidence rates from waiting-time graphs (Støvring 2002; Støvring & Vach 2005). The advantages of formulating a formal statistical model is first, that it provides valid estimates of uncertainty in the point estimates of prevalence and incidence rate, where the traditional estimates based on a fixed run-in period will be overly optimistic, as they implicitly assume that treatment status can be determined with certainty from the run-in period (Støvring & Vach 2005). Secondly, the technique avoids truncation by immigration while explicitly accounting for censoring due to emigration and death. Thirdly, the parametric estimation allows use of standard statistical techniques for assessment of model fit, since the fitted density may be compared both to the observed histogram and the empirical quantiles, cf. fig. 4. The validity of the method has been studied and confirmed in simulation studies (Støvring & Vach 2005), which also showed that the method requires roughly a doubling in the sample size for studying prevalence to obtain the same precision as studies where treatment is directly observed. For incidence rate, the corresponding increase in sample size is 5 to 8. With the large size of pharmacoepidemiological databases, this is not a severe limitation.
Corresponding estimates of prevalence and incidence are 8.12 (95% CI: 7.84; 8.41) per thousand, and 2.16 (1.97; 2.37) per thousand person years, respectively.

Software for applying the methodology in Stata 8 is available from the internet (Støvring 2005). Challenges for using the method is primarily that it does not directly categorize patients as prevalents and non-prevalents. What may be obtained from the model is predicted probabilities of treatment status, but so far the merits of using these in, for example, studies of adverse drug reactions are unknown. Alternatively, as the method rests on general statistical principles, it should allow formal extension for joint models of treatment status and outcomes of interest.

Unlike in Denmark, other health care systems might have an upper limit to the interval between prescription renewals, e.g., one month's supply. It will shift the curve towards the left, but the underlying properties are retained; estimates of prevalence and incidence rate are still valid.

It is also possible to achieve an intuitive impression of other utilization traits from the graph. A steep slope of the first segment suggests rapid prescription renewals, which is typical of expensive drugs. The slope of the terminal portion is an indicator of relapses, at least if the time-window is very wide. Finally, with the prevalence:incidence rate relationship described before, a long duration will manifest as a high ratio between the two areas, representing prevalent and incident users (fig. 2). We have, however, not yet developed formal, parametric methods to estimate duration, prescription renewal or relapse rates. Examples of prescribing problems that we have captured using the waiting-time distribution are lack of prescription renewal for disulfiram which is a drug for supporting alcohol abstinence (Hallas et al. 1997), lack of persistence with statins (Larsen et al. 2000) and extensive use of injected corticosteroids to treat hay fever (Hallas et al. 1997).

Finally, we have used waiting-time distributions in analytic studies of drug effects to provide a qualified guess on the treatment period for each prescription. For example, if we assign a fixed 90-day exposure window to all warfarin prescriptions, waiting-time analyses may reveal that quite many chronic users redeem prescriptions with much larger intervals. Consequently, they would be misclassified as non-exposed during a substantial part of their drug use history.

In conclusion, we can apply an array of non-specific analytical templates to our prescription databases and achieve an impression of how each drug or drug class is utilized. These analyses are fishing expeditions, not hypothesis-
templates for analysis of prescription data

References


