Prediction of postoperative nausea and vomiting

Jolanda E. van den Bosch

Van den Bosch, Jolanda E.
Prediction of postoperative nausea and vomiting
Thesis, University of Utrecht, Faculty of Medicine

ISBN 90-393-4149-4

Cover design: puking pumpkin, photographed by Chris Timmers
Lay out: FEBO Druk BV, Enschede, The Netherlands (by Nicole Nijhuis)
Printed by: FEBO Druk BV, Enschede, The Netherlands

Prediction of postoperative nausea and vomiting

Voorspellen van postoperatieve misselijkheid en braken

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor
aan de Universiteit Utrecht
op gezag van de rector magnificus, prof. dr. W.H. Gispen,
ingevolge het besluit van het college voor promoties

in het openbaar te verdedigen op

vrijdag 7 juli 2006
des ochtends om 10.30 uur

door

Jolanda Eveline van den Bosch

geboren op 25 september 1978 te Giessenburg

Promotoren: Prof. dr. C.J. Kalkman
Universitair Medisch Centrum Utrecht

Prof. dr. K.G.M. Moons
Universitair Medisch Centrum Utrecht

Co-promotoren: Dr. W.A. van Klei
Universitair Medisch Centrum Utrecht

Dr. Y. Vergouwe
Universitair Medisch Centrum Utrecht

The studies described in this thesis were financially supported by the Department of Perioperative Care and Emergency Medicine, University Medical Center Utrecht, Utrecht, The Netherlands, and by GlaxoSmithKline BV, Zeist, The Netherlands.

The randomized trial from which part of the data originated was financially supported by a fund for investigative medicine (Project Code OG96-038) from the Dutch Health Insurance Council, College voor Zorgverzekeringen, Amstelveen, The Netherlands.

We also gratefully acknowledge the support by the Netherlands Organisation for Health Research and Development (ZON-MW 917.46.360).

The printing of this thesis was financially supported by

- GlaxoSmithKline BV, Zeist, The Netherlands
- ProStrakan, Amersfoort, The Netherlands
- Department of Perioperative Care and Emergency Medicine, University Medical Center Utrecht, Utrecht, The Netherlands
- Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands

Quid illud, quod … ?

Manuscripts based on the studies described in this thesis:

Chapter 2
JE van den Bosch, GJ Bonsel, KGM Moons, CJ Kalkman. Effect of postoperative experiences on willingness to pay to avoid postoperative pain, nausea, and vomiting. Anesthesiology 2006; 104: 1033-9.

Chapter 3
JE van den Bosch, CJ Kalkman, Y Vergouwe, WA van Klei, GJ Bonsel, DE Grobbee, KGM Moons. Assessing the applicability of scoring systems for predicting postoperative nausea and vomiting. Anaesthesia 2005; 60: 323-331.

Chapter 4
JE van den Bosch, KGM Moons, GJ Bonsel, CJ Kalkman. Does measurement of preoperative anxiety have added value for predicting postoperative nausea and vomiting? Anesthesia and Analgesia 2005; 100: 1525-32.

Chapter 5
JE van den Bosch, Y Vergouwe, WA van Klei, L van Wolfswinkel, CJ Kalkman, KGM Moons. Prediction of PONV in surgical inpatients: which model to use in clinical practice? Submitted.

Chapter 6
JE van den Bosch, WA van Klei, Y Vergouwe, L van Wolfswinkel, KGM Moons, CJ Kalkman. Can prediction models for postoperative nausea and vomiting be generalized from surgical inpatients to outpatients? Submitted.

Chapter 7
JE van den Bosch, Y Vergouwe, WA van Klei, CJ Kalkman, KGM Moons. Boundaries to the generalizability of clinical prediction models. Submitted.

Table of Contents

Chapter 1: General Introduction 9

Chapter 2: Effect of postoperative experiences on willingness to pay to avoid postoperative pain, nausea and vomiting 17

Chapter 3: Assessing the applicability of prediction models for postoperative nausea and vomiting 35

Chapter 4: Does measurement of preoperative anxiety have added value for predicting postoperative nausea and vomiting? 53

Chapter 5: Prediction of PONV in surgical inpatients: which model to use in clinical practice? 71

Chapter 6: Can prediction models for postoperative nausea and vomiting be generalized from surgical inpatients to outpatients? 87

Chapter 7: Boundaries to the generalizability of clinical prediction models 103

Chapter 8: General Discussion 117

Summary 129

Samenvatting 135

Dankwoord 141

Curriculum Vitae 145

Chapter 1

General Introduction

Postoperative nausea and vomiting (PONV) are among the most important concerns for patients after surgery[1;2]. Many patients find PONV more distressing than postoperative pain[2] and are even willing to pay substantial amounts out-of-pocket in order to avoid PONV[3-5]. Persistent PONV causes extreme patient discomfort, tension on suture lines and increases risk of pulmonary aspiration. It may also delay patient discharge, particularly after ambulatory surgery[6-8]. Therefore, anesthesiologists have attempted to reduce the incidence of PONV. However, despite advances in anesthetic practice, the PONV incidence after general anesthesia is still about 25% to 30%[6;9] and can reach over 70% in patients who are at high risk[10]. Thus, there is still room for improvement in finding predictive factors (predictors) of PONV as well as in the prevention and treatment of this “big little problem”.

The occurrence of PONV is multifactorial, including patient, anesthetic and surgical factors[9;11]. Patient factors related to PONV are gender, smoking behavior, motion sickness, and a history of nausea and vomiting after previous surgery. The duration of anesthesia, the use of inhaled anesthetic agents, the type of surgery, and postoperative administration of opioid analgesics are also associated with PONV[10;12-18]. Several authors have suggested that psychological factors, such as preoperative anxiety, might play an additional role[9;19]. However, data supporting this hypothesis are lacking.

Possible strategies for PONV prevention include risk-based prophylactic administration of anti-emetics[3;10;12;20-22], use of propofol instead of volatile anesthetics, and substitution of nitrogen for nitrous oxide[11;18;23]. Routine application of these strategies to all surgical patients is contra-indicated because of potential side-effects and costs[9;24]. Instead, anti-emetic strategies should be selectively applied to patients at high risk of PONV. Such high-risk patients can be identified before surgery using a so called prediction model for PONV.

Prediction models combine predictors of the outcome (e.g. PONV) and the relative weights of these predictors (estimated with statistical techniques, such as logistic regression) to calculate the risk of the outcome for individual patients. This risk calculation is a cumbersome procedure to perform by hand. Therefore, prediction models are often simplified to so called risk scores. The anesthesiologist can then simply count the number of predictors present and read the corresponding risk estimate from a risk table. Alternatively, in clinics equipped with an electronic patient record (EPR) the prediction model can be implemented in computer software. Relevant information can then be retrieved automatically from the EPR and the risk of the outcome can be directly presented on screen, possibly accompanied by a treatment advice. In either way, prediction models for PONV can be used

in preoperative assessment to classify patients into subgroups of risk, which makes an institutional policy of risk-based prophylaxis of PONV possible.

So far, several prediction models for PONV have been proposed[10;12;14]. The best-known are those developed by Apfel et al.[12] and Koivuranta et al.[10]. Both models have shown good predictive accuracy in the patient sample from which they were developed. However, it is known that a prediction model generally predicts less accurately in new patients[25-27]. Therefore, the predictive accuracy of a model should be tested (validated) in several validation studies, involving various patient groups, before the model is widely implemented in clinical practice. The more diverse the patient groups in which the model has been validated and found to be robust, the more likely it is that the model will also generalize to new patients[28]. It is recommended that such validation studies are performed with patient samples (validation samples) that are different from but related to the patient sample from which the model was developed (development sample)[28;29]. Criteria for a 'different but related´ sample are, however, not clearly defined. Neither have the boundaries of the generalizability of clinical prediction models been fully explored.

The number of outpatient surgeries has increased in the last decade. As PONV may delay discharge[6-8;16] and return to work, it is increasingly important that risk-based PONV prophylaxis is not limited to inpatients but is applied to outpatients as well. Unfortunately, currently available prediction models were mainly developed in inpatient populations[10;12;14;22;30]. Whether these models predict accurately in outpatients is yet unknown. What is known is that the incidence of PONV in outpatients is lower than in inpatients[16], which may influence a model's predictive accuracy[31]. Therefore, it is indicated to test the generalizability of available inpatient prediction models for PONV to outpatient populations. If the predictive accuracy in outpatients is not satisfying, inpatient models may be modified for outpatient populations or a new model may be developed specifically for outpatients. For practical reasons, it may be preferable to develop one general model for predicting PONV in both in- and outpatients.

Outline of the thesis

The aim of the research described in this thesis was to test the generalizability of currently available prediction models for PONV across populations, including outpatient populations, and to improve the predictive accuracy of these models.

Chapter 2 describes an assessment of patients´ willingness to pay (WTP) for prophylactic anti-emetics. The resulting WTP values are compared

with WTP values for analgesics to obtain insight in the relative importance that PONV prevention has for patients as compared to avoidance of postoperative pain. The trial data that were used in this study included both preoperative and postoperative WTP values. This provided a unique opportunity to investigate whether postoperative experiences of pain, nausea, and vomiting changed a patient´s WTP, when corrected for other factors that may elicit a WTP change.

Chapter 3 describes the validation of the two best-known prediction models for PONV that are currently available. One of these models was already implemented by several groups, despite the fact that only a limited number of validation studies were performed, of which one was quite small and another was performed in the same institution as where the model was developed. Therefore, we performed an additional validation study using data from a large and heterogeneous population of Dutch inpatients, originating from another institution and country.

Chapter 4 describes the development of a new prediction model for PONV that was prompted by the results of the validation study described in chapter 3. This model was developed on the same data as were used for validating the two best-known prediction models for PONV (chapter 3). Further, we quantified the added predictive value of preoperative anxiety to the predictors of PONV that were included in the newly developed model.

Chapter 5 describes the historical and geographical validation of the newly developed prediction model for PONV in a sample of clinical inpatients, collected during another time period (historical validation) in another Dutch hospital (geographical validation).

Chapter 6 describes how the generalizability of available inpatient prediction models for PONV to ambulatory surgery was tested. The hypothesis was that inpatient models would provide similar predictive accuracy in outpatients, under the provision that small adjustments might be needed to account for a lower absolute incidence of PONV in outpatients.

Chapter 7 explores the boundaries to the generalizability of clinical prediction models. The concept of a ‘different but related´ sample is explored, illustrated with examples of prediction models that are developed and validated in different care settings (e.g. primary versus secondary care) or different age categories (e.g. adult versus pediatric patients). The chapter also discusses whether the generalizability of prediction models can be improved by using more heterogeneous development samples.

Chapter 8 discusses methodological issues related to PONV prediction, summarizes clinical implications, and outlines directions for future PONV prediction research.

References

1. Jenkins K, Grady D, Wong J, Correa R, Armanious S, Chung F. Post-operative recovery: day surgery patients' preferences. *British Journal of Anaesthesia* 2001;**86**:272-4.
2. Macario A, Weinger M, Carney S, Kim A. Which clinical anesthesia outcomes are important to avoid? The perspective of patients. *Anesthesia and Analgesia* 1999;**89**: 652-8.
3. Eberhart LH, Mauch M, Morin AM, Wulf H, Geldner G. Impact of a multimodal anti-emetic prophylaxis on patient satisfaction in high-risk patients for postoperative nausea and vomiting. *Anaesthesia* 2002;**57**:1022-7.
4. Gan TJ, Sloan F, de L Dear G, Wright D, El-Moalem HE, Lubarsky DA. How much are patients willing to pay to avoid postoperative nausea and vomiting? *Anesthesia and Analgesia* 2001;**92**:393-400.
5. Tang J, Wang B, White PF, Watcha MF, Qi J, Wender RH. The effect of timing of ondansetron administration on its efficacy, cost-effectiveness, and cost-benefit as a prophylactic antiemetic in the ambulatory setting. *Anesthesia and Analgesia* 1998;**86**:274-82.
6. Kovac AL. Prevention and treatment of postoperative nausea and vomiting. *Drugs* 2000;**59**:213-43.
7. Fortier J, Chung F, Su J. Unanticipated admission after ambulatory surgery - a prospective study. *Canadian Journal of Anesthesia* 1998;**45**:612-9.
8. Gold BS, Kitz DS, Lecky JH, Neuhaus JM. Unanticipated admission to the hospital following ambulatory surgery. *JAMA: the journal of the American Medical Association* 1989;**262**:3008-10.
9. Watcha MF, White PF. Postoperative nausea and vomiting. Its etiology, treatment, and prevention. *Anesthesiology* 1992;**77**:162-84.
10. Koivuranta M, Läärä E, Snåre L, Alahuhta S. A survey of postoperative nausea and vomiting. *Anaesthesia* 1997;**52**:443-9.
11. Apfel CC, Kranke P, Katz MH, Goepfert C, Papenfuss T, Rauch S et al. Volatile anaesthetics may be the main cause of early but not delayed postoperative vomiting: a randomized controlled trial of factorial design. *British Journal of Anaesthesia* 2002;**88**:659-68.
12. Apfel CC, Läärä E, Koivuranta M, Greim C-A, Roewer N. A simplified risk score for predicting postoperative nausea and vomiting: conclusions from cross-validations between two centers. *Anesthesiology* 1999;**91**:693-700.
13. Cohen MM, Duncan PG, DeBoer DP, Tweed WA. The postoperative interview: assessing risk factors for nausea and vomiting. *Anesthesia and Analgesia* 1994;**78**:7-16.
14. Palazzo M, Evans R. Logistic regression analysis of fixed patient factors for postoperative sickness: a model for risk assessment. *British Journal of Anaesthesia* 1993;**70**:135-40.
15. Pierre S, Benais H, Pouymayou J. Apfel's simplified score may favourably predict the risk of postoperative nausea and vomiting. *Canadian Journal of Anesthesia* 2002;**49**: 237-42.

16. Sinclair DR, Chung F, Mezei G. Can postoperative nausea and vomiting be predicted? *Anesthesiology* 1999;**91**:109-18.
17. Toner CC, Broomhead CJ, Littlejohn IH, Samra GS, Powney JG, Palazzo MG et al. Prediction of postoperative nausea and vomiting using a logistic regression model. *British Journal of Anaesthesia* 1996;**76**:347-51.
18. Visser K, Hassink EA, Bonsel GJ, Moen J, Kalkman CJ. Randomized controlled trial of total intravenous anesthesia with propofol versus inhalation anesthesia with isoflurane-nitrous oxide: postoperative nausea with vomiting and economic analysis. *Anesthesiology* 2001;**95**:616-26.
19. Lerman J. Surgical and patient factors involved in postoperative nausea and vomiting. *British Journal of Anaesthesia* 1992;**69**:24S-32S.
20. Biedler A, Wermelt J, Kunitz O, Müller A, Wilhelm W, Dethling J et al. A risk adapted approach reduces the overall institutional incidence of postoperative nausea and vomiting. *Canadian Journal of Anesthesia* 2004;**51**:13-9.
21. Pierre S, Corno G, Benais H, Apfel CC. A risk score-dependent antiemetic approach effectively reduces postoperative nausea and vomiting - a continuous quality improvement initiative. *Canadian Journal of Anesthesia* 2004;**51**:320-5.
22. Van den Bosch JE, Moons KG, Bonsel GJ, Kalkman CJ. Does measurement of preoperative anxiety have added value for predicting postoperative nausea and vomiting? *Anesthesia and Analgesia* 2005;**100**:1525-32.
23. Apfel CC, Korttila K, Abdalla M, Kerger H, Turan A, Vedder I et al. A factorial trial of six interventions for the prevention of postoperative nausea and vomiting. *The New England Journal of Medicine* 2004;**350**:2441-51.
24. Sung YF. Risks and benefits of drugs used in the management of postoperative nausea and vomiting. *Drug Safety: an international journal of toxicology and drug experience* 1996;**14**:181-97.
25. Laupacis A, Sekar N, Stiell IG. Clinical prediction rules. A review and suggested modifications of methodological standards. *JAMA: the journal of the American Medical Association* 1997;**277**:488-94.
26. McGinn TG, Guyatt GH, Wyer PC, Naylor CD, Stiell IG, Richardson WS. Users' guides to the medical literature: XXII: how to use articles about clinical decision rules. Evidence-Based Medicine Working Group. *JAMA: the journal of the American Medical Association* 2000;**284**:79-84.
27. Wasson JH, Sox HC, Neff RK, Goldman L. Clinical prediction rules. Applications and methodological standards. *The New England Journal of Medicine* 1985;**313**:793-9.
28. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Annals of Internal Medicine* 1999;**130**:515-24.
29. Altman DG, Royston P. What do we mean by validating a prognostic model? *Statistics in Medicine* 2000;**19**:453-73.

30. Gan TJ. Postoperative nausea and vomiting - can it be eliminated? *JAMA: the journal of the American Medical Association* 2002;**287**:1233-6.
31. Vergouwe Y, Steyerberg EW, Foster RS, Habbema JDF, Donohue JP. Validation of a prediction model and its predictors for the histology of residual masses in nonseminomatous testicular cancer. *Journal of Urology* 2001;**165**:84-8.

Chapter 2

Effect of postoperative experiences on willingness to pay to avoid postoperative pain, nausea and vomiting

JE van den Bosch, GJ Bonsel, KGM Moons, CJ Kalkman

Anesthesiology 2006; 104: 1033-9

Abstract

We assessed the willingness to pay (WTP) for 'perfect' prophylactic anti-emetics and analgesics in patients who are scheduled for surgery under general anesthesia. Furthermore, we determined whether postoperative experiences of nausea, vomiting (PONV) and pain changed patients' WTP. Data were collected alongside a randomized clinical trial that investigated the incidence of PONV in patients anesthetized with either inhalation anesthesia or total intravenous anesthesia. A subset of 808 consecutive patients participating in the trial completed WTP-questionnaires one day before and two weeks after surgery. The outcome measure was the maximum amount of money that patients were willing to pay for 'perfect' anti-emetics and analgesics. Preoperative WTP and individual WTP changes after surgery were analyzed in relation to baseline characteristics and postoperative pain and PONV experiences. Prevention of postoperative pain was valued higher than prevention of PONV. The median preoperative WTP for analgesics was US$ 35 (interquartile range: 7-69) versus US$ 17 (interquartile range: 7-69) for anti-emetics. Individual WTP changes for anti-emetics were not related to PONV experience, whereas severe postoperative pain (numerical rating score ≥ 8) was associated with an increase in the WTP for analgesics. Severe postoperative pain experiences increased patients' WTP for analgesics, but postoperative nausea and vomiting did not increase WTP for anti-emetics. The elicited WTP values were lower than reported in previous studies, which is possibly related to differences in market culture or patients´ attitudes towards postoperative nausea, vomiting and pain.

Introduction

Despite advances in anesthetic practice, postoperative pain, nausea, and vomiting (PONV) remain prevalent unpleasant side effects of anesthesia and surgery[1-5]. Possible strategies to reduce the incidence of PONV include risk-based prophylactic administration of anti-emetics[6-12] and the use of newer anesthetic drugs (propofol instead of volatile anesthetics, substitution of nitrogen for nitrous oxide)[7;13;14]. Postoperative pain management may be improved by utilization of an "acute pain service"[15] and perhaps by preoperative identification of patients at high risk of severe pain[16].

It is difficult to evaluate these interventions using a cost-utility analysis[17], because there is no universally accepted method of assigning monetary value to transient disutility such as postoperative pain or PONV. A possible alternative is to use the 'stated preference approach'. This approach establishes how patients value a certain health care outcome by asking them how much they are willing to pay (WTP) for an intervention that yields this outcome[18]. The elicited WTP values may be used as input in cost-benefit analyses of strategies for avoiding PONV and postoperative pain. When the costs of a strategy do not exceed the values that patients attach to its effects, one might decide to implement the strategy.

Several WTP studies have been performed in relation to PONV[9;19-22]. These studies, however, involved children[19], or were performed with small[20-22] or restricted[9;20] study populations, and WTP was measured after surgery only. So far, only one small study addressed patients' WTP for avoiding postoperative pain[22].

The present study assessed the WTP for a hypothetical, 'perfect' anti-emetic or analgesic, both one day before and two weeks after surgery, in a large, unselected mix of adult in- and outpatients scheduled for surgery under general anesthesia. Furthermore, we identified factors that determined patients' preoperative WTP and investigated whether postoperative pain and PONV experiences changed an individual's WTP.

Methods

Patients

Patients were recruited from a single-center patient-blinded observer-blinded randomized trial that compared the incidence of PONV after inhalation anesthesia with isoflurane/nitrous oxide versus intravenous anesthesia with propofol. The design of this trial was approved by the institutional review board of the Academic Medical Center of the University of Amsterdam, Amsterdam, the Netherlands, and is described elsewhere[14]. In short, this trial included in- and outpatients, aged 18-80 years, that were scheduled for various surgical procedures (with the exception of cardiac surgery or intracranial neurosurgery) under general anesthesia. Exclusion criteria were: ASA physical status > 3, use of anti-emetic or pro-emetic medication in the two weeks before surgery, or insufficient command of the Dutch language.

Blinded trained research nurses recorded baseline characteristics such as gender, age, patient status (outpatient versus inpatient), previous surgery, previous PONV, history of motion sickness, anxiety state and trait as measured with the Spielberger State-Trait Anxiety Inventory (STAI)[23], net household income and type of scheduled surgery.

Patients' postoperative experiences of PONV (i.e. nausea or retching or vomiting or a combination of these symptoms) and severe pain (measured as an 11-point numerical rating score [NRS], where 0 indicates no pain at all and 10 indicates the most severe pain imaginable) were recorded every 15 minutes until discharge from the postanesthesia care unit (PACU). Severe pain was defined as a NRS ≥ 8.

WTP-questionnaire

The WTP-questionnaire was presented to the last 852 participants in the trial. The WTP-questionnaire was piloted for comprehensibility and self-report suitability. Identical self-administered WTP-questionnaires were used pre- and postoperatively. The questionnaire that patients filled out postoperatively was returned by mail.

The WTP-questionnaire contained three questions. First, the patient was requested to indicate for a series of predefined amounts of money (US$ 7, 17, 35, 69, 173 and 346; see also 'Analysis') whether (s)he would be willing to pay each of these amounts out-of-pocket for a hypothetical, 'perfect' anti-emetic ("cumulative bidding technique"). This question was presented as follows: "Suppose a pill exists that completely prevents the occurrence of

nausea and vomiting after surgery. This pill will not be reimbursed by your medical insurance company. You have to go and buy this pill yourself at the pharmacy before the operation. How much would you be willing to pay for this pill? Fill out for each amount of money whether you would buy, would doubt to buy or would not buy the pill" (appendix, question 1). A similar scenario was presented for a hypothetical pill that completely avoids postoperative pain (appendix, question 2). A third question was added to measure patients' preferences for PONV versus pain prevention (appendix, question 3). Here, the patient was asked to choose one out of five possible 'combination pills' with different treatment efficacy for PONV and pain, varying from almost perfect treatment of PONV, but minimal effect on pain, to almost perfect treatment of pain, but minimal effect on PONV. This pill was stated to be free of charge. This question served as a check on the questionnaire's convergent validity. This means that preferences for either PONV or pain prevention should be consistent with possible differences in WTP for PONV and pain.

Analysis

The outcome on the individual level was the patient's maximal accepted bid (MAB), i.e. the maximum amount of money that a patient was willing to pay for a 'perfect' anti-emetic or analgesic. The MAB of a patient that did not want to pay the lowest amount of money (US$ 7) was set to US$ 0.

It should be noted that the original WTP values were measured in Dutch guilders. These values were converted to US$ values after a correction for price changes in the period from 1998 (year of data collection) to 2003 (most recent price index available), using a price index of 115.2. This index was calculated by the Dutch Central Bureau of Statistics, Voorburg, The Netherlands.

No formal sample size calculation was performed, but in view of previous studies a sample size in excess of 800 was regarded as sufficient to detect changes under a wide range of assumptions.

The following characteristics were considered as possible covariates for preoperative WTP: gender, age, previous surgery, previous PONV, history of motion sickness, net household income (up to 1382 US$; 1382 < US$ ≤ 2281 ; 2281 < US$ ≤ 3318 ; more than 3318 US$), patient status (in- or outpatient) and STAI anxiety state and trait. Individual WTP changes after surgery were tested for postoperative experiences of PONV and pain, in addition to anesthetic technique and the mentioned baseline characteristics.

Univariable and multivariable analyses for preoperative WTP dependence on baseline characteristics were performed with ordinal logistic regression. Previous to these analyses, the proportional odds assumption was tested with the score test[24;25]. The proportional odds assumption was justified for pain (without any measures). The assumption was also justified for PONV, but only after the variable for preoperative WTP was recoded (the categories US$ 173 and US$ 346 were collapsed to one category) and age was excluded from the multivariable analysis.

Univariable and multivariable analyses for dependence of individual WTP changes on PONV and pain experiences, income and other baseline characteristics, were performed with linear regression. Here, the change of WTP (in bids) was considered as a continuous variable with range -6 bids to +6 bids, in case a patient who was willing to pay US$ 346 before surgery changed its MAB to US$ 0 after surgery, or vice versa.

Variables were included in the multivariable analysis if the p-value in the univariable analysis was ≤ 0.15. All analyses were performed with SPSS for Windows 10.1.0 (SPSS Inc., Chicago IL, USA). The proportional odds assumptions were tested with SAS 8.2 (TS2MO, SAS Institute Inc. Cary, NC, USA).

Results

Response and baseline characteristics of respondents

Of the 852 participants, 808 (95%) responded to the WTP-questionnaire. Preoperative WTP data were obtained from 803 (94%) (anti-emetics) and 808 (95%) (analgesics) patients. Postoperative WTP data were obtained from 681 (80%) patients for both anti-emetics and analgesics. Complete pre- and postoperative WTP data for anti-emetics and analgesics were obtained from 681 (80%) patients.

Of the 808 respondents, 57% were women and 74% were inpatients, with a median age of 42 years (interquartile range [IQR]: 32-54). Eighty-three percent of the respondents had undergone previous surgery of whom 32% reported to have experienced PONV. Motion sickness was reported in 19% of the respondents. The median preoperative STAI anxiety state score was 40 points (IQR: 33-49) and the STAI anxiety trait score 37 points (IQR: 31-45). The net median household income per month was US$ 2074 (IQR: 1382-3110); 37% of the respondents declined to reveal their net monthly household income (table 1). Baseline characteristics of non-respondents were similar to that of respondents.

Pill preference and WTP for 'perfect' anti-emetics and analgesics

The distributions of the respondents' pill preferences and maximal accepted bids for 'perfect' anti-emetics and analgesics are shown in figure 1.

For anti-emetics, the median WTP was US$ 17 (IQR: 7-69) before surgery. Two weeks after surgery, the median WTP for anti-emetics was US$ 17 (IQR: 7-35) (table 2). After surgery, 47% of the respondents were willing to pay the same amount of money for preventing PONV as before surgery, whereas 32% decreased their bid and 21% increased their bid.

For analgesics, the median WTP was US$ 35 (IQR: 7-69) before and after surgery (table 2). After surgery, 48% of the respondents were willing to pay the same amount of money for preventing pain, 27% decreased their bid and 25% increased their bid.

The higher WTP for analgesics compared to anti-emetics was also reflected by the responses to the pill preference question. Respondents generally favored a pill that was more effective against pain than against PONV: 56% preferred pill A or B whereas 27% preferred pill D or E. Sixteen percent of the respondents showed no preference for either pain or PONV prevention (pill C) (figure 1c).

Table 1 Characteristics of the respondents. Data are presented as percentages unless stated otherwise.

characteristic	% of respondents
female gender	57
inpatient status	74
age (years)	42 (32 – 54)[a]
previous surgery	83
history of PONV[b] in patients with previous surgery	32
history of Motion sickness	19
STAI[c] anxiety state score	40 (33 – 49)[a]
STAI[c] anxiety state score > 50 points	20
STAI[c] anxiety trait score	37 (31 – 45)[a]
STAI[c] anxiety trait score > 50 points	12
net household income per month[d]	
quartile[e] 1: ≤ 1382 US$	16
quartile[e] 2: 1382 < US$ ≤ 2281	20
quartile[e] 3: 2281 < US$ ≤ 3318	15
quartile[e] 4: > 3318 US$	13
unwilling to provide income data	18
missing	19
type of surgery[d]	
superficial	75
lower abdominal	4
upper abdominal	4
laparoscopy	10
strabismus	1
middle ear	5

[a] median (interquartile range)

[b] postoperative nausea and vomiting

[c] using Spielberger's State-Trait Anxiety Inventory (STAI), the patient's anxiety state and trait are both scored on a continuous scale from twenty to eighty points. Patients scoring more than 50 points are considered to be very anxious or to have a strong anxiety trait, respectively.

[d] total percentage is not equal to 100% due to rounding error

[e] income quartiles for the population in the Netherlands as defined by the Dutch Central Bureau of Statistics in 1996

Figure 1 Maximal accepted bids and pill preferences before (white bars) and two weeks after (shaded bars) surgery. Figure 1a shows the distribution of respondents' pre- and postoperative willingness to pay for 'perfect' anti-emetics (appendix, question 1). Figure 1b shows the distribution of respondents' pre- and postoperative willingness to pay for 'perfect' analgesics (appendix, question 2). Figure 1c shows the distribution of respondents' pre- and postoperative pill preferences (appendix, question 3).

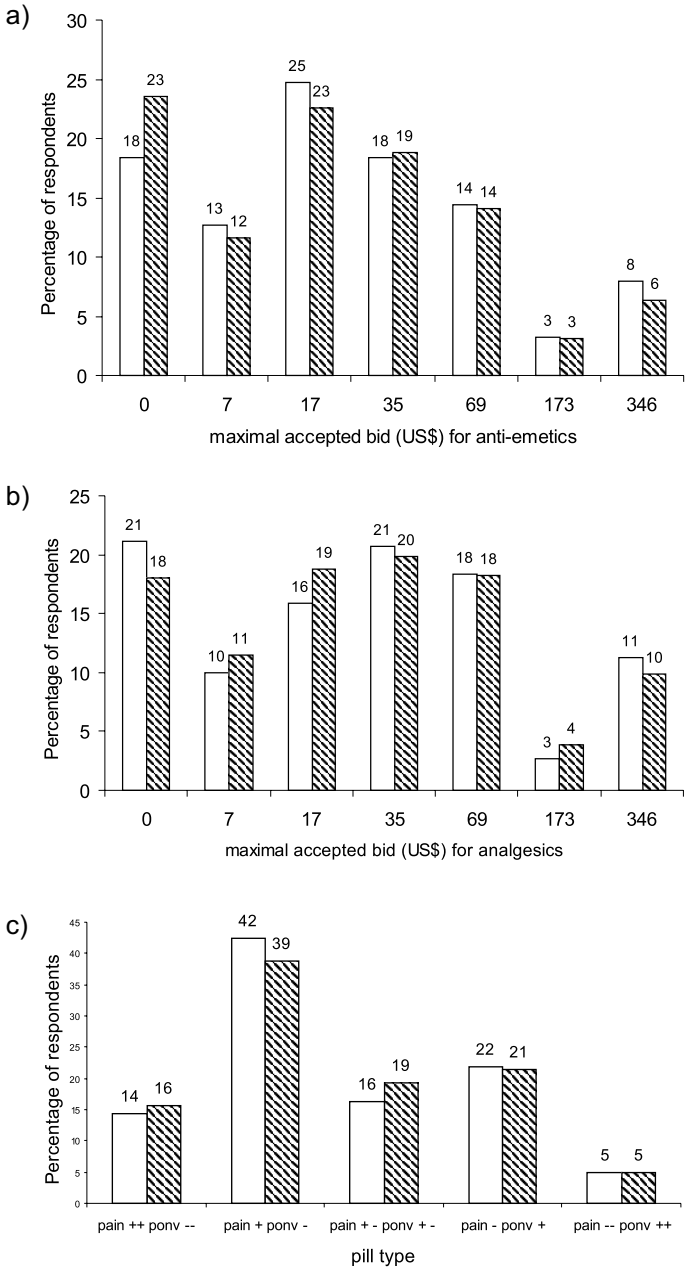

Table 2 Absolute willingness to pay (WTP) before and two weeks after surgery and individual WTP change (in number of bids change). Data are presented as median (interquartile range).

	absolute WTP		individual WTP change
	preoperative WTP	**postoperative WTP**	
PONV[a]	**17 (7 – 69)**	**17 (7 – 35)**	**0 (-1 ; 0)**
no PONV[a]	17 (7 – 69)	17 (0 – 35)	0 (-1 ; 0)
any PONV[a]	17 (7 – 69)	17 (7 – 69)	0 (0 ; 1)
nausea (moderate or severe)	17 (7 – 69)	17 (17 – 69)	0 (-1 ; 0,5)
retching or vomiting	17 (7 – 69)	35 (17 – 69)	0 (0 ; 0,75)
pain	**35 (7 – 69)**	**35 (7 – 69)**	**0 (-1 ; 0)**
NRS[b] > 4	35 (7 – 69)	35 (7 – 69)	0 (-1 ; 1)
NRS[b] ≥ 7	35 (7 – 69)	35 (7 – 69)	0 (0 ; 1)
NRS[b] ≥ 8	35 (7 – 69)	35 (7 – 69)	0 (0 ; 1)

[a] postoperative nausea and vomiting
[b] numerical rating score

Table 3 Postoperative nausea and vomiting (PONV) and pain experiences in the postanesthesia care unit. Data are presented as percentages.

postoperative experience	**% of respondents**
PONV	
no PONV	75
any PONV	25
nausea (moderate or severe)	13
retching or vomiting	13
pain	
NRS[a] > 4	63
NRS[a] ≥ 7	39
NRS[a] ≥ 8	23

[a] numerical rating score

Preoperative WTP in relation to baseline characteristics

According to univariable ordinal regression analysis, previous surgery was negatively related ($p = 0.025$) to preoperative WTP for anti-emetics, whereas STAI preoperative anxiety state score ($p = 0.12$) and net household income ($p < 0.001$) had a positive association with this outcome. A multivariable regression analysis showed that preoperative WTP for anti-emetics was increased if patients were more anxious ($p = 0.023$) or had a higher net household income ($p < 0.001$).

According to univariable ordinal regression analysis, female gender ($p = 0.101$) and age ($p = 0.029$) were negatively related to preoperative WTP for analgesics, whereas inpatient status ($p = 0.021$), STAI preoperative anxiety state score > 50 points ($p = 0.006$) and net household income ($p < 0.001$) had a positive association with this outcome. Multivariable analysis showed that preoperative WTP for analgesics was higher if patients had the inpatient status ($p = 0.004$), were more anxious (i.e. STAI state score > 50 points) ($p = 0.028$) or had a higher net household income ($p < 0.001$).

Individual WTP changes in relation to postoperative experiences of pain and PONV in the PACU

In the PACU, 75% of all respondents were free of PONV, 13% was moderately or severely nauseated and 13% experienced at least one episode of retching or vomiting. At least moderate pain (NRS > 4) was experienced by 63% of patients, whereas 39% (NRS ≥ 7) and 23% (NRS ≥ 8) experienced severe pain in the PACU (table 3).

According to univariable linear regression analysis, experience of any PONV in the PACU was positively related ($p = 0.054$) to a postoperative change (increase) of the maximal accepted bid (MAB) for anti-emetics. Other factors that were associated with an increase of the MAB were female gender ($p = 0.076$), STAI preoperative anxiety state score ($p = 0.025$), STAI anxiety trait score ($p = 0.017$) and isoflurane/nitrous oxide anesthesia ($p = 0.005$). Older respondents ($p = 0.083$) and respondents with higher incomes, however, were more inclined to decrease their MAB ($p = 0.028$). The income effect was confirmed in the multivariable analysis ($p = 0.043$), but the influence of the other factors including experience of any PONV in the PACU could not be demonstrated.

According to univariable linear regression analysis, experience of severe pain (NRS ≥ 8) in the PACU was associated ($p = 0.016$) with a postoperative increase of the MAB for analgesics and so was female gender

(p = 0.142), whereas respondents with a STAI anxiety trait score > 50 points (p = 0.011) tended to decrease their bids. These associations were also evident from the multivariable analysis (p-value of 0.006 for STAI anxiety trait score > 50 points and 0.015 for NRS ≥ 8), although the effect of female gender could not be confirmed.

Discussion

We measured surgical patients' willingness to pay (WTP) for 'perfect' anti-emetics and analgesics before and two weeks after surgery, and investigated which factors altered patients' WTP after surgery. We also assessed whether patients preferred prevention of PONV over avoidance of pain or vice versa.

Patients generally preferred avoidance of pain over prevention of PONV. Before surgery, half of our respondents were willing to pay US$ 35 or more out-of-pocket to avoid postoperative pain versus US$ 17 or more to prevent PONV. Also, after surgery, the MAB was more often decreased for PONV than for pain, whereas the percentage of patients that was willing to pay the same amount of money as before surgery was about the same (48%).

We expected that patients who suffered PONV or pain would increase their bid two weeks after surgery. After surgery, however, the WTP for pain was not changed, irrespective of the severity of early postoperative pain experience (table 2). Although the WTP for PONV was increased for patients who experienced retching or vomiting, it increased only slightly for patients who suffered moderate/severe nausea (table 2). Despite these variations in WTP between subgroups with different postoperative PONV experiences, multivariable analyses showed that only income determined whether individual patients would change their WTP for PONV after the operation. In contrast, individual changes in WTP for pain were not income dependent, but evoked by severe postoperative pain (NRS ≥ 8) and related to a patient's preoperative anxiety trait.

Higher preoperative anxiety levels also increased patients' willingness to pay for PONV and pain prevention before surgery. Other factors that were positively related to preoperative WTP were income (anti-emetics and analgesics) and inpatient status (analgesics). Although our results confirm the relation between income and WTP that was evident from previous studies[19;21], the influence of previous PONV experience on WTP[21] could not be confirmed.

When interpreting our results, it should be noted that 83% of our patients underwent surgery on a previous occasion. On average, these patients had lower preoperative MAB's for PONV (US$ 17) than inexperienced surgical patients (US$ 35). The preoperative MAB's for pain were the same as in patients who had never been operated before (US$ 35).

Five previous studies have investigated the WTP for PONV prevention in surgical patients. One study assessed how much parents (n=162), whose

child had undergone surgery within the previous two years, were willing to pay to reduce the chance that their child would experience PONV from 1 in 3 to less than a 1 in 10 chance, in case the child had to undergo a hypothetical simple surgical procedure requiring general anesthesia[19]. The median WTP in this study was UK£ 50 (95% confidence interval: 20-80). Another study assessed the WTP for preventing PONV in 164 women scheduled for outpatient laparoscopic surgery. The study focused on a hypothetical situation in which the patient had to undergo the same laparoscopic operation in the future[20]. The mean (standard deviation) WTP out-of-pocket in this population was US$ 117 (82). In a third study, 80 elective surgical outpatients were asked (after recovery in the PACU) how much they were willing to pay to receive a hypothetical new drug, that would eliminate PONV, instead of the routine anti-emetic, which would still allow a chance of PONV[21]. The median WTP out-of-pocket in this patient sample was US$ 56 (IQR: 26-97). The WTP ranged from US$ 40 (IQR: 20-95) in patients who did not develop PONV to US$ 100 (IQR: 61-200) in patients who experienced emesis. The fourth study measured WTP in a group of female patients admitted for gynecological and breast surgery of which 108 at high risk of PONV received multimodal anti-emetic treatment, and a control group of 71 women at a lower risk of PONV did not receive anti-emetics[9]. The WTP for receiving the same anesthetic technique in future was high in patients that received multimodal prophylaxis: UK£ 84 (IQR: 33-184) compared to UK£ 14 (IQR: 4-30) for patients in the control group. Finally, in a study of 66 generally young Caucasians, patients were willing to pay US$ 162 and US$ 230 to avoid nausea and vomiting respectively. The same patients were willing to pay US$ 267 to prevent pain[22].

Remarkably, the WTP for avoiding postoperative nausea, vomiting and pain in these previous studies was much higher than that observed in our study (US$ 17 for PONV and US$ 35 for pain, table 2). Although we used a bidding range with preset money values, which is vulnerable to starting point bias, the lower WTP values found in our study can not be explained by the choice of the bidding range, as the highest value in our bidding range (US$ 7 to 346) is even larger than the maximum values of the bidding ranges used in previous studies (US$ 0 to 190 or > 190[19], and US$ 0 to 250 or other[20]). Alternatively, the differences in WTP values may be related to differences in sample size and patient selection (previous study populations were about 5 times smaller and consisted of female patients with an increased risk of PONV, or parents that may be inclined to pay more to prevent PONV in their children). The fact that the median WTP in the outpatient study and in the subsample of patients who did not suffer PONV was much closer to the

WTP values we found, also seems to point in this direction. Differences in (market) culture or patients´ attitudes towards postoperative pain, nausea and vomiting may also play a role. Whether this is actually so and whether this has a substantial influence on WTP values, when corrected for the influence of other factors, remains to be defined. If cultural differences play a role, these may limit the generalizability of the results, but this effect may be counterbalanced by the fact that our study is the first study which assessed the WTP in a non-selected group of patients.

If we accept the WTP values for 'perfect' anti-emetics and analgesics elicited in this study, we may use them as input in cost-benefit analyses of strategies for avoiding PONV and postoperative pain. When the costs of a strategy do not exceed the values that patients attach to its effects, one might decide to implement the strategy. Alternatively, one might decide to ask for a modest co-payment for improved drugs, knowing that both the pre- and postoperative WTP values for anti-emetics and analgesics were in the magnitude of existing co-payments presently employed for some medications in the Netherlands.

In conclusion, severe postoperative pain experiences increased patients' WTP for analgesics, but postoperative nausea and vomiting did not increase WTP for anti-emetics. The elicited WTP values were lower than reported in previous studies, which is possibly related to differences in market culture or patients´ attitudes towards postoperative nausea, vomiting and pain.

References

1. Apfelbaum JL, Chen C, Mehta SS, Gan TJ. Postoperative pain experience: results from a national survey suggest postoperative pain continues to be undermanaged. *Anesthesia and Analgesia* 2003;**97**:534-40.
2. Huang N, Cunningham F, Laurito CE, Chen C. Can we do better with postoperative pain management? *American Journal of Surgery* 2001;**182**:440-8.
3. Kovac AL. Prevention and treatment of postoperative nausea and vomiting. *Drugs* 2000;**59**:213-43.
4. Watcha MF, White PF. Postoperative nausea and vomiting. Its etiology, treatment, and prevention. *Anesthesiology* 1992;**77**:162-84.
5. Macario A, Weinger M, Carney S, Kim A. Which clinical anesthesia outcomes are important to avoid? The perspective of patients. *Anesthesia and Analgesia* 1999;**89**: 652-8.
6. Apfel CC, Läärä E, Koivuranta M, Greim C-A, Roewer N. A simplified risk score for predicting postoperative nausea and vomiting. *Anesthesiology* 1999;**91**:693-700.
7. Apfel CC, Korttila K, Abdalla M, Kerger H, Turan A, Vedder I et al. A factorial trial of six interventions for the prevention of postoperative nausea and vomiting. *The New England Journal of Medicine* 2004;**350**:2441-51.
8. Biedler A, Wermelt J, Kunitz O, Müller A, Wilhelm W, Dethling J et al. A risk adapted approach reduces the overall institutional incidence of postoperative nausea and vomiting. *Canadian Journal of Anesthesia* 2004;**51**:13-9.
9. Eberhart LHJ, Mauch M, Morin AM, Wulf H, Geldner G. Impact of a multimodal anti-emetic prophylaxis on patient satisfaction in high-risk patients for postoperative nausea and vomiting. *Anaesthesia* 2002;**57**:1022-7.
10. Koivuranta M, Läärä E, Snåre L, Alahuhta S. A survey of postoperative nausea and vomiting. *Anaesthesia* 1997;**52**:443-9.
11. Pierre S, Corno G, Benais H, Apfel CC. A risk score-dependent antiemetic approach effectively reduces postoperative nausea and vomiting - a continuous quality improvement initiative. *Canadian Journal of Anesthesia* 2004;**51**:320-5.
12. Van den Bosch JE, Moons KG, Bonsel GJ, Kalkman CJ. Does measurement of preoperative anxiety have added value for predicting postoperative nausea and vomiting? *Anesthesia and Analgesia* 2005;**100**:1525-32.
13. Apfel CC, Kranke P, Katz MH, Goepfert C, Papenfuss T, Rauch S et al. Volatile anaesthetics may be the main cause of early but not delayed postoperative vomiting: a randomized controlled trial of factorial design. *British Journal of Anaesthesia* 2002;**88**:659-68.
14. Visser K, Hassink EA, Bonsel GJ, Moen J, Kalkman CJ. Randomized controlled trial of total intravenous anesthesia with propofol versus inhalation anesthesia with isoflurane-nitrous oxide: postoperative nausea and vomiting and economic analysis. *Anesthesiology* 2001;**95**:616-26.

15. Rawal N. 10 years of acute pain services - achievements and challenges. *Regional anesthesia and pain medicine* 1999;**24**:68-73.
16. Kalkman CJ, Visser K, Moen J, Bonsel GJ, Grobbee DE, Moons KGM. Preoperative prediction of severe postoperative pain. *PAIN* 2003;**105**:415-23.
17. Robinson R. Economic evaluation and health care. What does it mean? *British Medical Journal* 1993;**307**:670-3.
18. Robinson R. Economic evaluation and health care: cost-benefit analysis. *British Medical Journal* 1993;**307**:924-6.
19. Diez L. Assessing the willingness of parents to pay for reducing postoperative emesis in children. *Pharmacoeconomics* 1998;**13**:589-95.
20. Tang J, Wang B, White PF, Watcha MF, Jinhui Q, Wender RH. The effect of timing of ondansetron administration on its efficacy, cost-effectiveness, and cost-benefit as a prophylactic antiemetic in the ambulatory setting. *Anesthesia and Analgesia* 1998;**86**: 274-82.
21. Gan TJ, Sloan F, Dear GL, El-Moalem HE, Lubarsky DA. How much are patients willing to pay to avoid postoperative nausea and vomiting? *Anesthesia and Analgesia* 2001;**92**:393-400.
22. Macario A, Vasanawala A. Improving quality of anesthesia care: opportunities for the new decade. *Canadian Journal of Anesthesia* 2001;**48**:6-11.
23. Spielberger CD, Gorsuch RL, Lushene RE. Manual for the State-Trait Anxiety Inventory. Palo Alto, California: Consulting Psychologists Press, 1970.
24. Peterson B, Harrell FE. Partial proportional odds models for ordinal response variables. *Applied Statistics* 1990;**39**:205-17.
25. Stiger TR, Barnhart HX, Williamson JM. Testing proportionality in the proportional odds model fitted with GEE. *Statistics in Medicine* 1999;**18**:1419-33.

Appendix

Willingness to pay questionnaire

Question 1: willingness to pay for 'perfect' anti-emetics

Suppose a pill exists that completely prevents the occurrence of **nausea and vomiting** after surgery. This pill will not be reimbursed by your medical insurance company. You have to go and buy this pill at the pharmacy before the operation. How much would you be willing to pay for this pill? Fill out for each amount of money whether you would buy, would doubt to buy or would not buy the pill.

US$	7	17	35	69	173	346
Yes	☐	☐	☐	☐	☐	☐
I doubt	☐	☐	☐	☐	☐	☐
No	☐	☐	☐	☐	☐	☐

Question 2: willingness to pay for 'perfect' analgesics

Suppose a pill exists that completely prevents **pain** after surgery. This pill will not be reimbursed by your medical insurance company. You have to go and buy this pill at the pharmacy before the operation. How much are you willing to pay for this pill? Fill out for each amount of money whether you would buy, would doubt to buy or would not buy the pill (see above).

Question 3: pill preference

Unfortunately, a pill that completely prevents both nausea/vomiting and pain does not exist. Suppose we can offer you one of the following pills for free. You do not have to pay for it. Which pill would you choose to make you feel as comfortable as possible after the operation?

- ☐ pill A: relieves pain almost completely; hardly relieves nausea and vomiting
- ☐ pill B: relieves pain well; relieves nausea and vomiting a little
- ☐ pill C: relieves both pain and nausea/vomiting fairly well
- ☐ pill D: relieves pain a little; relieves nausea and vomiting well
- ☐ pill E: hardly relieves pain; relieves nausea and vomiting almost completely

Chapter 3

Assessing the applicability of prediction models for postoperative nausea and vomiting

JE van den Bosch, CJ Kalkman, Y Vergouwe, WA van Klei, GJ Bonsel, DE Grobbee, KGM Moons

Anaesthesia 2005; 60: 323-31

Abstract

We have validated two prediction models for predicting postoperative nausea and vomiting (PONV), developed by Apfel et al. and Koivuranta et al. from 1388 adult inpatients undergoing a wide range of surgical procedures. The predictive accuracy of the prediction models was evaluated in terms of the ability to discriminate between patients with and without postoperative nausea and vomiting (discrimination) and agreement between observed and predicted outcomes (calibration). Discrimination and calibration were less than expected based on previous reports, with both prediction models providing risk predictions that were too extreme. The area under the receiver operating characteristic (ROC) curve was 0.63 for Apfel et al.'s prediction model and 0.66 for Koivuranta et al.'s prediction model. Neither of the prediction models provided a risk threshold for administering anti-emetic prophylaxis that yielded satisfying results in terms of predictive values, sensitivity and specificity. Hence, in their original forms, the prediction models do not guarantee accurate prediction of the risk of PONV in other patient populations. Koivuranta et al.'s prediction model appears to be more robust across different populations.

Introduction

Despite modern anesthetic and surgical techniques, the incidence of postoperative nausea and vomiting (PONV) is still about 25-30%[1;2]. Since PONV may result in extreme discomfort, it is unacceptable to neglect the prophylactic potential of anti-emetics to reduce the risk of PONV. However, routine administration of anti-emetics to all surgical patients is contra-indicated because of potential side-effects and costs[2;3]. Instead, prophylactic anti-emetics should be selectively administered to patients at high risk of PONV, for example those identified before surgery using a prediction model for predicting PONV.

Several prediction models have been proposed for predicting PONV within 24 hours after surgery[4-7]. The best-known prediction models are those developed by Apfel et al. (Germany)[4] and Koivuranta et al. (Finland)[6]. Both prediction models have shown good predictive accuracy in the patient samples from which they were developed. Consequently, Apfel et al.'s prediction model was recently used to select high-risk patients for participation in a randomized trial that evaluated six interventions to reduce PONV[8]. Furthermore, several groups have already implemented Apfel et al.'s prediction model in daily practice, for example, to investigate whether risk-based prophylaxis reduced the overall incidence of PONV[9-11]. However, before widespread implementation in clinical practice, a prediction model should be validated in various patient populations, as it is known that a prediction model generally predicts less accurately in new patients than in the patients that were used to develop the model[12-15]. The more diverse the populations are in which a prediction model has been tested and found to be accurate, the more likely it is that the prediction model will generalize to new populations. Despite this, the predictive accuracy of the two prediction models for predicting PONV has hardly been tested in patient samples other than the samples from which the models were developed[10;16;17]. Moreover, one of these validation studies[16] was performed in the institution where the original prediction model was developed and another[10] in a small number of (predominantly female) patients undergoing a limited number of surgical procedures. We therefore tested the generalizability of Apfel et al.'s and Koivuranta et al.'s prediction models to a large number of patients undergoing a wide range of surgical procedures.

Methods

Prediction models and development samples

Apfel et al.'s prediction model

Apfel et al.'s prediction model was developed from a combined sample of 1040 adult surgical inpatients aged ≥ 18 years, in two centers (Oulu, Finland and Wuerzburg, Germany)[4]. PONV was defined as at least one episode of nausea or vomiting within the first 24 hours after surgery. Nausea was assessed on a binary scale at 2 hours and on an 11-point numeric scale (0-10) at 24 hours after surgery. The number of vomiting episodes was recorded for both intervals. The average incidence of PONV in the combined sample was 44%. Descriptive statistics for the combined sample were not presented in the original paper, but could be estimated by averaging the numbers presented for the Oulu and the Wuerzburg samples (table 1, column 2). The formula of the original prediction model, developed using multivariable logistic regression analysis, was:

Risk of PONV = 1 / (1 + exp (- (-2.28 + 1.27*female gender + 0.65*history of PONV or motion sickness + 0.72* non-smoking + 0.78*postoperative opioid use))) (1)

The expression '-2.28 + 1.27*female gender + 0.65*history of PONV or motion sickness + 0.72* non-smoking + 0.78*postoperative opioid use' is called the linear predictor. In this expression, -2.28 represents the model's intercept and the other numbers represent the regression coefficients (weights) of each corresponding predictor. The formula 1 / (1 + exp(-(-2.28 + …))) represents the patient's baseline risk, which is increased by the presence of any of the other factors.

This logistic regression model was simplified to a four-item risk score, which was defined as the number of predictors present. If none, one, two, three or four of these predictors are present, the predicted risk of PONV is 10%, 21%, 39%, 61% and 79%, respectively. The logistic regression models and simplified risk scores (from Oulu and Wuerzburg) were internally validated by cross-validation (developed in one center and validated in the other center).

Koivuranta et al.'s prediction model

Koivuranta et al.'s prediction model was developed from a sample of 1107 surgical inpatients aged 4-86 years, who were prospectively included during a 3-month period in 1995 in one center in Oulu, Finland[6]. A subset of these

data was part of the combined sample from which Apfel et al.'s prediction model was developed. The original paper reported incidences of nausea and vomiting separately. Both outcomes were assessed for the intervals 0-2 hours and 2-24 hours after surgery, with nausea rated on an 11-point numeric scale (0-10). As the overall incidence of PONV within 24 hours after surgery was not reported, we estimated this incidence using the incidence of nausea, which was 52% (table 1, column 3). The formula of the original prediction model, developed using multivariable logistic regression analysis, was:

Risk of PONV = 1 / (1 + exp (- (-2.21 + 0.93*female gender + 0.82*history of PONV + 0.59* history of motion sickness + 0.61*non-smoking + 0.75*duration of surgery over 60 minutes))) (2)

The intercept (-2.21) was obtained by means of written communication with the author. This baseline risk was about the same as in Apfel et al.'s prediction model.

Koivuranta et al.'s logistic regression model was also simplified to a five-item risk score, which was defined as the number of predictors present. If none, one, two, three, four or five of these predictors were present, the estimated predicted risk of nausea was 17%, 18%, 42%, 54%, 74% and 87%, respectively. Internal validation of the logistic regression model or risk score was not reported.

Validation sample

The present (validation) sample included surgical inpatients aged 18-80 years, who were enrolled in the 'AIDA' trial conducted at the Academic Medical Center of the University of Amsterdam, the Netherlands, from April 1997 to January 1999[18]. The objective of this trial was to assess the incidence of PONV after total intravenous anesthesia (TIVA) with propofol versus inhalation anesthesia with isoflurane/nitrous oxide. The trial was approved by the institutional medical ethics committee and written informed consent was given by all patients. The design and primary results have been reported elsewhere[18].

The AIDA trial randomly assigned 1447 inpatients to receive either inhalation anesthesia or TIVA. Follow-up data at 24 hours were complete in 1403 (97%) of these patients. For the present analyses, another 15 patients were excluded from the sample because of missing values for one or more predictors included in Apfel et al.'s or Koivuranta et al.'s prediction model, leaving 1388 analyzable patients. The distribution of the other predictors

was similar among the 1388 patients with complete and the 59 patients with missing data. All types of surgery were included except cardiac surgery and intracranial procedures. The validation sample represented 1029 superficial (74%), 53 upper abdominal (4%), 81 lower abdominal (6%), 104 laparoscopic (8%), 12 strabismus (1%), 95 middle ear (7%) procedures, and 14 other procedures (1%). Other exclusion criteria were emergency surgery, (possible) pregnancy, ASA physical status 4, morbid obesity (weight > 120 kg), renal or liver disease precluding use of either anesthetic technique, use of anti-emetic or pro-emetic medication in the 2 weeks before surgery, previous enrollment in the same study and insufficient command of the Dutch language. For all study subjects, we documented information on predictors included in Apfel et al.'s and Koivuranta et al.'s prediction models, i.e. female gender, smoking behavior, history of PONV and motion sickness, the type and duration of the scheduled surgery, the duration of anesthesia and postoperative administration of opioid analgesics. All patients were interviewed (every 15 minutes in the recovery unit and hourly for 24 hours after surgery in the ward) by a thoroughly trained research nurse who was blinded to the anesthetic technique. The nurse recorded nausea, retching, vomiting and analgesic and anti-emetic medication. Patients were considered to have had PONV if at least one episode of postoperative nausea, retching or vomiting, or any combination of these, occurred in the first 24 hours.

Data analysis

To validate the original prediction models, the risk of PONV was calculated for each patient in the validation sample using equations 1 and 2. To validate the two simplified risk scores, the number of predictors present was counted for each patient in the validation sample. The predictive accuracy of the original prediction models and the simplified risk scores was quantified using measures of discrimination and calibration. Discrimination, the ability of a prediction model or risk score to distinguish between patients with and without PONV, was estimated with the area under the receiver operating characteristic curve (ROC area). The ROC area may theoretically range from 0.5 (discrimination equivalent to that of chance) to 1.0 (perfect discrimination)[19]. Calibration refers to the agreement between the observed and the predicted outcome frequencies. For each original prediction model, calibration was studied by estimating a smoothed calibration line between the observed and predicted outcomes. Such a calibration line has an intercept and a slope which are ideally 1 and 0, respectively (perfect calibration). To estimate the slope of the calibration line in our validation sample, we fitted

a binary logistic regression model with the occurrence of PONV observed in the validation sample as (binary) outcome and the linear predictor of the original prediction model as the only covariate[20]. This linear predictor was estimated for each patient in the validation sample using equations 1 and 2. The linear predictor is the sum of the intercept and the original regression coefficients multiplied by that patient's observed value of the corresponding predictor. The calibration of the original prediction models was further tested with the Hosmer-Lemeshow test statistic[21]. For each simplified risk score, calibration was studied with a histogram, comparing the incidence of PONV in the validation and development samples.

The clinical applicability of the risk scores was evaluated for several risk thresholds using conventional performance measures: positive predictive value, negative predictive value, sensitivity and specificity. For each threshold, the percentage of patients that would be treated with anti-emetic prophylaxis was also calculated.

Analyses were performed with SPSS 10.1 (SPSS Inc., Chicago, IL) and S-PLUS 2000 Professional Edition for WINDOWS.

Results

Table 1 shows the distribution of predictor and outcome values for the two development samples and the present validation sample. The three samples were largely comparable, except that the validation sample included fewer patients with a history of PONV, fewer non-smokers, and an intermediate proportion of patients who received postoperative opioids.

Table 1 Distribution of predictor and outcome values (where presented) in the development samples[4;6] and the validation sample[18]. Values are number (proportion), median (interquartile range) or median [range].

	development samples		validation sample
	Apfel et al.[4] [a]	**Koivuranta et al.[6]**	**Visser et al.[18]**
	(n=1040)	**(n=1107)**	**(n=1388)**
predictors			
female gender	593 (57)	731 (66)	791 (57)
age (years)	49 (36 - 61)	46 [4 - 86]	44 (32 - 56)
history of PONV	-	476 (43)	347 (25)
history of motion sickness	-	266 (24)	264 (19)
history of PONV or motion sickness	364 (35)	-	527 (38)
non-smoking	759 (73)	863 (78)	847 (61)
duration of surgery (minutes)	116 (73 - 152)	72 (40 - 105)	120 (80 - 173)
duration of surgery > 60 minutes	-	-	1249 (90)
use of postoperative opioids	478 (46)	786 (71)	874 (63)
outcome			
overall incidence of PONV	458 (44)	576 (52)	666 (48)

[a] Values in this combined sample were estimated by averaging the numbers presented for the two original samples (Oulu and Wuerzburg).

Apfel et al.'s prediction model

The ROC area for Apfel et al.'s original prediction model was 0.62 in the validation sample. We could not compare this with the ROC area of the prediction model found in the development sample because this was not presented in the original publication. For the simplified risk score, the ROC area was 0.63 in the validation sample, indicating that the simplified risk score and the original prediction model had a similar discriminative ability. The ROC area of the simplified risk score was larger in the development sample than in the validation sample (table 2).

Table 2 Areas under the receiver operating characteristic (ROC) curve for Apfel et al.'s[4] and Koivuranta et al.'s[6] risk scores as found in the development samples and validation sample[18]. Ranges in brackets are 95% confidence intervals (where provided in the original paper).

	development sample	validation sample
Apfel et al.[4]	0.75 (0.71 - 0.79)	0.63 (0.60 - 0.66)[a]
Koivuranta et al.[6]		
nausea	0.72	0.66 (0.63 - 0.69)
vomiting	0.70	0.65 (0.62 - 0.68)
PONV	-[b]	0.66 (0.63 - 0.69)

[a] p = 0.05

[b] The ROC area for PONV was not provided in the original paper. It may be assumed that the ROC area for PONV equals the ROC area for nausea, as the incidence of PONV commonly equals the incidence of nausea.

For Apfel et al.'s original prediction model (equation 1), the agreement between the predicted risk and the observed incidence of PONV is shown in figure 1a. The slope of the calibration line was 0.39, suggesting poor calibration. This was confirmed by a statistically significant Hosmer-Lemeshow test ($p < 0.001$). Application of the simplified risk score to the validation sample yielded observed PONV incidences as shown in figure 2a. For patients with 0, 1 or 2 predictors, the observed incidences of PONV in the validation sample were substantially higher than would have been expected based on the results from the development sample, whereas the incidences of PONV were lower for patients in whom 3 or 4 predictors were present. Hence, for patients with a low risk score, the predicted risk of PONV was too low, whereas for patients with a high risk score, the predicted risk was too high.

Table 3a shows the number of patients with and without PONV in the validation sample for each possible number of predictors according to Apfel et al.'s simplified risk score, as well as the characteristics of the risk score for several risk thresholds for administering anti-emetic prophylaxis. For example, using a threshold of three or more risk factors for defining a high risk of PONV, 39% ((389 + 152)/1388) of all patients would receive anti-emetic prophylaxis. This would result in a positive predictive value ((217+110)/(389 + 152)) and a negative predictive value ((46 + 195 + 266)/(62 + 297 + 488)) of 60% each. Of all patients who developed PONV, 49% ((217 + 110)/667) would be correctly treated (sensitivity), but 30% ((172 + 42)/721) of all patients who did not develop PONV would receive unnecessary anti-emetic treatment (1-specificity).

Koivuranta et al.'s prediction model

For Koivuranta et al.'s original prediction model and for the simplified risk score, the ROC area for PONV was 0.66 in the validation sample (table 2). Unfortunately, the ROC areas could not be compared with those of the development sample as the 95% confidence intervals and standard errors were not presented in the original paper.

For the original prediction model, the agreement between the predicted risk and the observed incidence of PONV is shown in figure 1b. The slope of the calibration line was 0.58 and the Hosmer-Lemeshow test was statistically significant ($p < 0.001$), also indicating poor calibration. Application of the simplified risk score to the validation sample yielded the incidences of PONV given in figure 2b.

Table 3b shows the number of patients with and without PONV in the validation sample for each possible number of predictors according to Koivuranta et al.'s simplified risk score. This table can be read similarly to table 3a.

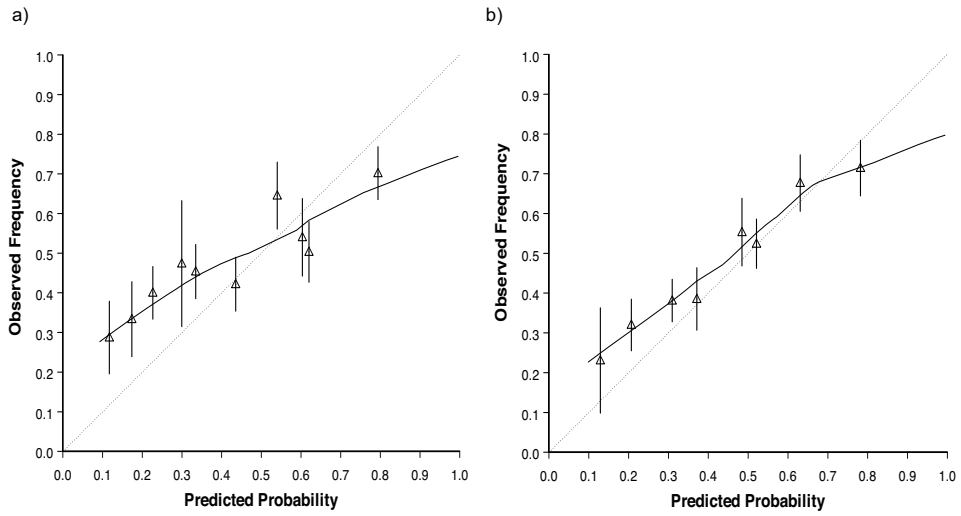

Figure 1 Calibration plots of the original prediction models developed by Apfel et al.[4] (figure 1a) and Koivuranta et al.[6] (figure 1b) in the validation sample[18]. The solid line shows the association between observed outcomes and predicted risks. Ideally, this line equals the dotted line that represents perfect calibration. Triangles indicate the observed frequency of PONV per decile of predicted risk. The vertical lines represent the corresponding 95% confidence intervals.

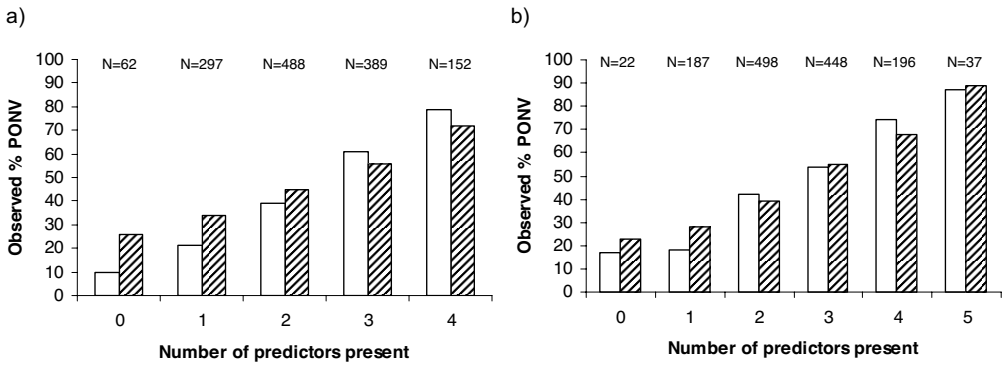

Figure 2 Percentage of patients with PONV per risk class as found in the validation sample[18] (shaded bars) compared to the development samples (white bars) of Apfel et al.[4] (figure 2a) and Koivuranta et al.[6] (figure 2b). Figures at the top indicate the number of patients per risk class in the validation sample.

Table 3 Number of patients in the validation sample[18] with and without PONV for each possible number of predictors according to risk classification with the simplified risk scores. Values are numbers or proportion. Results of possible prophylactic policies for several treatment thresholds are shown.

	Number of predictors present					
a) Apfel et al.[4]	**0**	**1**	**2**	**3**	**4**	**Total**
Number of patients with PONV	16	102	222	217	110	667
Number of patients without PONV	46	195	266	172	42	721
Total number of patients per risk class	62	297	488	389	152	1388
Threshold for administering prophylaxis		*≥ 1*	*≥ 2*	*≥ 3*	*≥ 4*	
Proportion of patients receiving prophylaxis		96	74	39	11	
Positive predictive value		49	53	60	72	
Negative predictive value		74	67	60	55	
Sensitivity		98	82	49	16	
Specificity		6	33	70	94	

	Number of predictors present						
b) Koivuranta et al.[6]	**0**	**1**	**2**	**3**	**4**	**5**	**Total**
Number of patients with PONV	5	53	194	248	134	33	667
Number of patients without PONV	17	134	304	200	62	4	721
Total number of patients per risk class	22	187	498	448	196	37	1388
Threshold for administering prophylaxis		*≥ 1*	*≥ 2*	*≥ 3*	*≥ 4*	*≥ 5*	
Proportion of patients receiving prophylaxis		98	85	49	17	3	
Positive predictive value		48	52	61	72	89	
Negative predictive value		77	72	64	57	53	
Sensitivity		99	91	62	25	5	
Specificity		2	21	63	91	99	

Discussion

In the present study, we tested the applicability of two prediction models[4;6] for predicting the risk of PONV within the first 24 hours after surgery. The predictive accuracy of the prediction models in the validation sample was substantially lower than in the samples that were used to develop the prediction models. Both prediction models provided risk predictions that were too extreme, i.e. the predicted low risks were too low and the predicted high risks too high. Also, the ability of the prediction models to discriminate between the presence and absence of PONV was poorer than expected based on results from previous studies. Furthermore, none of the possible risk thresholds for anti-emetic prophylaxis delivered satisfactory results in terms of predictive values or sensitivity and specificity. Hence, in their original forms, the prediction models do not guarantee accurate prediction of the risk of PONV in other patient populations.

Apfel et al.'s prediction model

In the validation sample, the discrimination of the original prediction model and simplified risk score was poor. For the risk score, the ROC area was 0.75 in the development sample but 0.63 in the validation sample. In previous validation studies, ROC areas were 0.68 (same institution where the prediction model was developed)[16] and 0.71 (predominantly female patients from a cancer referral center)[10]. Loss of discriminative ability is often observed when prediction models are too optimistic, which means that in new patients, high risk estimates are too high and low estimates too low[12;21;22]. A decrease in discriminative ability can sometimes also be explained by a difference in the distribution of predictor values between validation and development samples. In this case, there were indeed differences in the distribution of non-smokers (61% versus 73%) and of postoperative opioid analgesia (46% versus 67%) (table 1). However, these differences were relatively small and it is unlikely that they explain such a large decrease in ROC area. Neither is it likely that the decrease in ROC area was caused by differences in the definition of outcomes, since these were the same in the development and validation samples.

The calibration of the original prediction model and simplified risk score was poor in the validation sample. The calibration line of the original prediction model presented with a flattened slope of 0.39 instead of 1.0 (perfect calibration). This indicates that in patients for whom the original prediction model predicted a low risk of PONV, the actual incidence of PONV

was much higher than predicted, whereas in patients who were classified as having a high risk, the incidence was lower. The flattened slope confirms that the observed decrease in ROC area was caused by overoptimism of the original prediction model[12;20-22]. This means that the strength of the association of certain predictors with the outcome was probably overestimated[12]. Indeed, we found that the odds ratio for female gender and postoperative opioids were significantly higher in the development sample than in the validation sample (table 4). Moreover, postoperative opioid analgesia did not show a significant association with the outcome (OR 1.1) in the validation sample, compared with an OR of 2.2 in the development study (table 4). In the study by Koivuranta et al., use of postoperative opioids was significantly associated with nausea (OR 1.7), but it was not included in the model as only the five strongest predictors were selected.

In previous validation studies of Apfel et al.'s prediction model, the reported slopes of the calibration line were 0.82[16] and 1.08[10]. It should be noted that these values are difficult to interpret and to compare with our results. Rather than using a binary logistic model based on data of all individual patients, these slopes were estimated using linear regression analysis, based on data that were clustered within risk groups. This method has two disadvantages. First, clustering of patients' data results in loss of information. Second, as PONV is a dichotomous outcome, it is better to use a binary logistic regression model to fit the calibration line than a linear regression model, which assumes a continuous outcome[12;20-22].

Table 4 Odds ratios for the predictors of Apfel et al.'s and Koivuranta et al.'s original prediction models, as found in the development samples[4;6] and the validation sample[18]. Ranges in brackets are 95% confidence intervals.

	development sample	validation sample
Apfel et al.[4]		
female gender	3.6 (2.5 - 5.2)	1.6 (1.3 - 2.0)
history of PONV or motion sickness	1.9 (1.4 - 2.7)	2.3 (1.8 - 2.9)
non-smoking	2.1 (1.5 - 2.8)	1.7 (1.4 - 2.2)
use of postoperative opioids	2.2 (1.7 - 2.9)	1.1 (0.8 - 1.3)
Koivuranta et al.[6]		
female gender	2.5[a]	1.7 (1.3 - 2.1)
history of motion sickness	1.8[a]	1.9 (1.4 - 2.6)
history of PONV	2.3[a]	2.2 (1.7 - 2.8)
duration of surgery > 60 minutes	2.1[a]	1.7 (1.4 - 2.2)
non-smoking	1.8[a]	1.9 (1.3 - 2.8)

[a] The original paper presents regression coefficients but not the odds ratios; these were calculated based on regression coefficients presented in SCORE 1 in the original publication.

Koivuranta et al.'s prediction model

For the original prediction model, the ROC area in the development sample was 0.72 for nausea and 0.71 for vomiting. For the simplified risk score, the ROC area in the development sample was 0.72 for nausea and 0.70 for vomiting. When applied to the validation sample, the ROC area was 0.66 (nausea), 0.65 (vomiting) and 0.66 (PONV) for both the original prediction model and the simplified risk score. These results are concordant with the results from a previous validation study of Koivuranta et al.'s prediction model, which also included a large number of patients (n=1566) undergoing various types of surgery[16]. A third validation study among 1444 patients undergoing a variety of surgical procedures, however, reported ROC areas much closer to the results found in the development sample (0.71 for PONV and 0.73 for vomiting)[17].

The calibration line of Koivuranta et al.'s original prediction model showed a flattened slope of 0.58 in our data, which again indicates that the predicted risks were too extreme. As with Apfel et al.'s prediction model, this may be due to an overestimation of predictor-outcome associations, such as for female gender (OR was 2.5 in the development sample versus 1.7 in our sample) (table 4). Nevertheless, Koivuranta et al.'s prediction model was substantially less miscalibrated than Apfel et al.'s. In previous validation studies, the reported slopes of the calibration lines were 1.13[16] and 0.99[17]. However, the values of these slopes are again difficult to interpret because of the analytical method chosen. It would be interesting to see the slopes of the calibration lines in these two studies, when estimated with binary logistic regression analysis.

Clinical implications

It is statistically attractive to quantify the predictive accuracy of a clinical prediction model with measures of discriminative ability and calibration. For clinical practice, however, if only patients with a relatively high risk of PONV will be treated with prophylaxis, it is less interesting whether a predicted risk of for example 10% should actually be 20%. It is more important to define a clinically relevant threshold, for which the benefits of prophylaxis against PONV outweigh the costs and possible side-effects. Therefore, we presented the test characteristics of Apfel et al.'s and Koivuranta et al.'s risk scores (table 3) for several risk thresholds for anti-emetic prophylaxis. In the present validation sample, none of the thresholds delivered satisfying results. For example, when using Apfel et al.'s risk score with a threshold for

administering prophylactic anti-emetics of three or more predictors, 39% of all patients will be treated with a sensitivity of 49% and a specificity of 70%. This means that, at most, half of the patients who develop PONV will be treated with prophylaxis. Alternatively, when lowering the threshold to two or more predictors, the sensitivity increases to 82%, but at the expense of the specificity, which decreases to 33%. Hence, 67% of patients who will not develop PONV will receive prophylaxis unnecessarily. Since administration of prophylactic anti-emetics may expose patients to side-effects[3;23], this is not desirable.

Ideally, the predictive accuracy of a prediction model is robust across populations. In reality, however, a prediction model generally predicts less accurately in new patients than in the patients used to develop the prediction model. Therefore, clinicians considering the use of a prediction model have three options, which have recently been summarized in detail[22]. First, one can use an existing prediction model without modification, knowing that the identified predictors or the weights assigned to them (regression coefficients) may not be ideal for one's own patient population. Second, one can accept the predictors in the published prediction model, but adjust the prediction model to one's own patient population (recalibration). This is done by multiplying the intercept and/or regression coefficients of the prediction model with a correction factor (shrinkage factor) that is estimated from one's own population. Third, one can revise an existing prediction model, by re-estimating individual regression coefficients (model revision) and including new predictors (model extension).

For our population, both prediction models were clearly miscalibrated. In such an instance, recalibration (method 2) is indicated. In cases where the predictor weights of the prediction model apply to the new patient population, it may be sufficient to adjust the intercept only. This is the case when the miscalibration of the prediction model affects all patients in a similar way (e.g. due to withholding of nitrous oxide in one sample but not another). It is also the case when the developed prediction model lacks a certain predictor that plays a role in the new population. However, when not only the intercept but also the predictor weights are different, the prediction model as a whole may be recalibrated to the new population. This is done by simultaneously adjusting both the intercept and the regression coefficients with an overall correction factor, as previously described[22;24;25]. Finally, when there are differences between the development and validation samples that specifically affect patients with certain characteristics (e.g. more frequent administration of prophylactic anti-emetics to women) or when predictor definitions differ (e.g. different definitions of smoking), overall recalibration

may not be sufficient and re-estimation of individual regression coefficients may be required (method 3). If this does not improve the performance of the prediction model, the prediction model may be extended by including additional predictors[22;24-26].

The development of robust prediction models can perhaps be compared to cumulative meta-analyses for estimating the efficacy of a particular treatment. It starts with one study that develops a prediction model. This prediction model is then tested (validated) and, if needed, adjusted using newly available data to improve the accuracy of the prediction model across various populations. In our institution, we have chosen to revise the existing PONV prediction models by re-estimating individual coefficients and extending the prediction models by including additional predictors. The results of this analysis have been published recently[27]. It should be noted, however, that adjusting a prediction model to local circumstances and populations (methods 2 and 3) requires individual patient data on both the predictors and the occurrence of PONV. If such empirical data are not available, the prediction model developed by Koivuranta et al. may provide the most accurate predictions of PONV risk, as this prediction model appears to be more robust across populations than the prediction model developed by Apfel et al., according to the present and previous validation studies. To prevent unexpected outcomes, one might first test the prediction model(s) in one's own patient population before using it in daily clinical practice to guide anti-emetic management.

References

1. Kovac AL. Prevention and treatment of postoperative nausea and vomiting. *Drugs* 2000;**59**:213-43.
2. Watcha MF, White PF. Postoperative nausea and vomiting. Its etiology, treatment and prevention. *Anesthesiology* 1992;**77**:162-84.
3. Sung Y-F. Risks and benefits of drugs used in the management of postoperative nausea and vomiting. *Drug Safety* 1996;**14**:181-97.
4. Apfel CC, Läärä E, Koivuranta M, Greim CA, Roewer N. A simplified risk score for predicting postoperative nausea and vomiting: conclusions from cross-validations between two centers. *Anesthesiology* 1999;**91**:693-700.
5. Gan TJ. Postoperative nausea and vomiting - can it be eliminated? *JAMA: the journal of the American Medical Association* 2002;**287**:1233-6.
6. Koivuranta M, Läärä E, Snåre L, Alahuhta S. A survey of postoperative nausea and vomiting. *Anaesthesia* 1997;**52**:443-9.
7. Palazzo M, Evans R. Logistic regression analysis of fixed patient factors for postoperative sickness: a model for risk assessment. *British Journal of Anaesthesia* 1993;**70**:135-40.
8. Apfel CC, Korttila K, Abdalla M, Kerger H, Turan A, Vedder I et al. A factorial trial of six interventions for the prevention of postoperative nausea and vomiting. *The New England Journal of Medicine* 2004;**350**:2441-51.
9. Biedler A, Wermelt J, Kunitz O, Müller A, Wilhelm W, Dethling et al. A risk adapted approach reduces the overall institutional incidence of postoperative nausea and vomiting. *Canadian Journal of Anesthesia* 2004;**51**:13-9.
10. Pierre S, Benais H, Pouymayou J. Apfel's simplified score may favourably predict the risk of postoperative nausea and vomiting. *Canadian Journal of Anesthesia* 2002;**49**:237-42.
11. Pierre S, Corno G, Benais H, Apfel CC. A risk score-dependent anti-emetic approach effectively reduces postoperative nausea and vomiting - a continuous quality improvement initiative. *Canadian Journal of Anesthesia* 2004;**51**:320-5.
12. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* 1996;**15**:361-87.
13. Laupacis A, Sekar N, Stiell IG. Clinical prediction rules. A review and suggested modifications of methodological standards. *JAMA: the journal of the American Medical Association* 1997;**277**:488-94.
14. Mc Ginn TG, Guyatt GH, Wyer PC, Naylor CD, Stiell IG, Richardson WS. Users' guides to the medical literature XXII: how to use articles about clinical decision rules. *JAMA: the journal of the American Medical Association* 2000;**284**:79-84.
15. Wasson JH, Sox HC, Neff RK, Goldman L. Clinical prediction rules. Applications and methodological standards. *The New England Journal of Medicine* 1985;**313**:793-9.

16. Apfel CC, Kranke P, Eberhart LHJ, Roos A, Roewer N. Comparison of predictive models for postoperative nausea and vomiting. *British Journal of Anaesthesia* 2002;**88**:234-40.
17. Eberhart LH, Högel J, Seeling W, Staack AM, Geldner G, Georgieff M et al. Evaluation of three risk scores to predict postoperative nausea and vomiting. *Acta Anaesthesiologica Scandinavica* 2000;**44**:480-8.
18. Visser K, Hassink EA, Bonsel GJ, Moen J, Kalkman CJ. Randomized controlled trial of total intravenous anesthesia with propofol versus inhalation anesthesia with isoflurane-nitrous oxide: postoperative nausea and vomiting and economic analysis. *Anesthesiology* 2001;**95**:616-26.
19. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;**143**:29-36.
20. Spiegelhalter DJ. Probabilistic prediction in patient management and clinical trials. *Statistics in Medicine* 1986;**5**:421-33.
21. Hosmer D, Lemeshow S. Applied logistic regression. New York: John Wiley & Sons, Inc., 1989.
22. Steyerberg EW, Borsboom GJJM, Van Houwelingen HC, Eijkemans MJC, Habbema JDF. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Statistics in Medicine* 2004;**23**:2567-86.
23. Domino KB, Anderson EA, Polissar NL, Posner KL. Comparative efficacy and safety of ondansetron, droperidol, and metoclopramide for preventing postoperative nausea and vomiting: a meta-analysis. *Anesthesia and Analgesia* 1999;**88**:1370-9.
24. Van Houwelingen HC, Thorogood J. Construction, validation and updating of a prognostic model for kidney graft survival. *Statistics in Medicine* 1995;**14**:1999-2008.
25. Van Houwelingen HC. Validation, calibration, revision and combination of prognostic survival models. *Statistics in Medicine* 2000;**19**:3401-15.
26. Ivanov J, Tu JV, Naylor CD. Ready-made, recalibrated, or remodeled? *Circulation* 1999;**99**:2098-104.
27. Van den Bosch JE, Moons KGM, Bonsel GJ, Kalkman CJ. Does measurement of preoperative anxiety have added value for predicting postoperative nausea and vomiting? *Anesthesia and Analgesia* 2005;**100**:1525-32.

Chapter 4

Does measurement of preoperative anxiety have added value for predicting postoperative nausea and vomiting?

JE van den Bosch, KGM Moons, GJ Bonsel, CJ Kalkman

Anesthesia and Analgesia 2005; 100: 1525-32

Abstract

Preoperative anxiety has been suggested as a predictor of postoperative nausea and vomiting (PONV), but supporting data are lacking. We quantified the added predictive value of preoperative anxiety to established predictors of PONV in 1389 surgical inpatients undergoing various procedures, by using multivariable logistic regression analysis. Investigated predictors were a history of PONV or motion sickness, smoking, gender, age, ethnicity, body mass index, ASA physical status, surgery type, duration of anesthesia, anesthetic technique, and postoperative opioid analgesia. Anxiety was measured by Spielberger's State-Trait Anxiety Inventory (STAI) and the Amsterdam Preoperative Anxiety and Information Scale (APAIS). The outcome was the occurrence of PONV within the first 24 hours after surgery. The area under the receiver operating characteristic (ROC) curve of a multivariable (logistic regression) model including gender, age, smoking, history of PONV or motion sickness, surgery type, and anesthetic technique was 0.70 (95% confidence interval: 0.68-0.72; after correction for overoptimism). There was a weak but significant association of anxiety with PONV, but the addition of anxiety to the model did not further increase the area under the ROC curve. Therefore, routine preoperative measurement of anxiety does not seem warranted, provided that the other predictors are already considered.

Introduction

To decrease the incidence of postoperative nausea and vomiting (PONV), several anti-emetic strategies are available, e.g. use of propofol anesthesia or prophylactic administration of anti-emetic drugs. Routine administration of anti-emetics, however, increases costs and subjects all patients to potential side effects[1]. Reliable preoperative prediction of PONV risk would enable anesthesiologists to selectively apply anti-emetic strategies.

Prediction rests on knowledge of the etiology of PONV, which is multifactorial and includes patient-related, anesthetic, and surgical factors[1]. Gender, smoking, history of PONV or motion sickness, inhaled anesthesia, surgery type, and postoperative opioid analgesia are reported predictors of PONV[2-8]. Several authors have suggested that preoperative anxiety might also be predictive for PONV[1;9]. However, data supporting this hypothesis are still lacking.

To assess the added value of anxiety in preoperative prediction of PONV, we first analyzed the (univariable) association of anxiety and PONV. Subsequently, we assessed the predictive accuracy of a model with previously reported predictors only, and finally, we assessed whether preoperative anxiety improves the predictive accuracy of these predictors.

Methods

We analyzed data from patients who participated in a randomized trial investigating differences in PONV incidence after intravenous or inhaled general anesthesia. The design and results of this trial have been reported previously[8]. In brief, after approval by the institutional medical ethics committee and written informed consent, inpatients and outpatients scheduled for surgery and aged 18-80 years were randomized to receive surgery with intravenous anesthesia with propofol or with inhaled anesthesia using isoflurane and nitrous oxide. All types of surgery were included except for cardiac surgery and intracranial surgery. Exclusion criteria were emergency surgery, (possible) pregnancy, ASA class IV, morbid obesity (weight > 120 kg), renal or liver disease precluding use of either anesthetic technique, and use of anti-emetic or pro-emetic medication in the 2 weeks before surgery. Patients receiving regional anesthetic techniques were not included in the study, except in the case of upper abdominal surgery, for which epidural analgesic supplementation was permitted. Blinding was maintained at all stages of the trial. This analysis was based on data from the inpatients.

To quantify preoperative anxiety, patients filled in two previously validated questionnaires for assessing preoperative anxiety: the Dutch version of Spielberger's State-Trait Anxiety Inventory (STAI)[10] and the Amsterdam Preoperative Anxiety and Information Scale (APAIS)[11]. The STAI questionnaire consists of 40 items on a 4-point rating scale measuring anxiety state (20 items) and anxiety disposition (20 items). Minimum and maximum scores range from 20 to 80 points. The APAIS questionnaire consists of six questions, each scored on a 5-point scale from 1 (not at all) to 5 (extremely). The APAIS scale was specifically designed to assess patients' preoperative anxiety (four questions; range 4-20 points) and information-seeking behavior (two questions; range 2-10 points) regarding the scheduled surgery and anesthesia.

We investigated 12 additional potential predictors of PONV[1-4;6;8]. These included age, gender, ethnicity (Caucasian versus non-Caucasian), body mass index (weight [in kilograms] divided by squared height [in meters]), ASA physical status, smoking, history of PONV (PONV after previous anesthesia: yes or no), history of motion sickness, anesthetic technique (intravenous anesthesia with propofol versus inhaled anesthesia with isoflurane/nitrous oxide), and type of scheduled surgery. Type of surgery was categorized as superficial, laparoscopy, upper abdomen, lower abdomen, strabismus, middle ear, and other (seven categories). In addition, postoperative opioid analgesia and the duration of anesthesia were recorded for each patient[2;4].

The induction of anesthesia was achieved with thiopental in patients randomized to isoflurane/nitrous oxide, or with propofol in patients randomized to anesthesia with propofol/air. Anesthesia was maintained with propofol or isoflurane in nitrous oxide according to the randomization. The treatment allocation was concealed until immediately before the induction of anesthesia. Prophylactic anti-emetics were not permitted. Intra-operatively, the anesthesiologist was free to choose the types and doses of muscle relaxants and (opioid) analgesics, as well as other drugs for the supplementation of anesthesia ('pragmatic study design'). Postoperative administration of anti-emetic therapy was the responsibility of postanesthesia care unit (PACU) nurses. The protocol prescribed intravenous metoclopramide (0.15 mg/kg) as the first choice anti-emetic therapy, followed by intravenous droperidol (15 μg/kg) if necessary.

The outcome of this study was PONV within the first 24 hours after surgery. The presence of PONV was defined as at least one episode of nausea (any degree, including mild nausea) or vomiting or retching, or any combination of these emetic symptoms within the first 24 hours after surgery. Each patient with PONV was counted only once, irrespective of the frequency of emetic symptoms.

A trained, blinded research nurse, who was not responsible for the postoperative care, recorded nausea, retching, and vomiting (scored separately). During the patient's stay in the PACU, nausea, retching, and vomiting were recorded at four time points: on arrival, 1 hour after arrival, at the time of permission for discharge, and at the time of actual discharge. At the ward, symptoms were recorded each hour up to 24 hours after surgery. We also documented, at each time point, whether the patient was sufficiently alert to complete a verbal rating scale. Patients who were too drowsy or sleepy received a score of 0 for that time point.

First, we quantified the univariable association between each anxiety variable and the incidence of PONV. Subsequently, the predictive accuracy of combinations of previously reported predictors was quantified by using multivariable logistic regression modeling. The resulting (basic) model, which included the independent predictors of PONV only, was then extended by adding the anxiety (APAIS and STAI) variables, separately and in combination, to estimate their added predictive value. Of particular interest was the difference in incremental value, if any, of the brief APAIS questionnaire compared with the much longer STAI. Differences between models were estimated with the log likelihood ratio test, and a model's ability to discriminate between patients with and without PONV was quantified by using the area under the receiver operating characteristic curve (ROC area).

A simple prediction model based on these results was developed for clinical practice. The predictive accuracy of this model was estimated by its calibration (i.e. agreement between observed and predicted probabilities of PONV), discrimination, and the absolute number of correctly predicted patients with and without PONV across various risk categories. Calibration was tested by using the Hosmer and Lemeshow method[12].

The accuracy of a prediction model is always too optimistic in the development sample. Hence, we estimated this overoptimism by using bootstrapping techniques[13] (rather than split-sample or cross-validation methods, because the former has proven to be superior)[14], adjusted the model for this overoptimism, and calculated the ROC area that can be expected in future patients[15].

Results

In total, 667 patients (48%) experienced PONV within 24 hours after surgery, of which 414 (62%) developed PONV in the PACU and 253 (38%) developed PONV 2-24 hours after surgery. Vomiting occurred in 402 patients (29%).

All anxiety measurements, except the APAIS need for information score, were significantly associated with PONV (table 1). A higher anxiety level was associated with a higher incidence of PONV. Also, most investigated predictors were associated with PONV, except for ASA physical status, premedication with benzodiazepines, and peroperative and postoperative opioid analgesia (table 1). The predictors 'history of PONV' and 'history of motion sickness' showed similar associations with PONV (table 1). Therefore, they were combined into a single predictor 'history of PONV or motion sickness', which yielded an odds ratio (OR) of 2.41. Likewise, 'lower abdominal surgery' and 'middle ear surgery' were combined to the index category 'high-risk surgery' versus all other types of surgery as the reference category (table 1). A model including the four established predictors gender, history of PONV or motion sickness, smoking, and postoperative opioid analgesia yielded an ROC area of 0.67 (95% confidence interval [CI]: 0.65-0.69). All predictors were significantly associated with PONV ($p < 0.001$), except for postoperative opioid analgesia ($p = 0.20$). Excluding postoperative opioid analgesia and adding age, anesthetic technique, and type of surgery (included either as six indicator variables or as a dichotomous predictor) significantly increased the ROC area to 0.72 (95% CI: 0.70-0.74). All six predictors in this basic model were significantly associated with PONV ($p < 0.001$). None of the other previously reported predictors further improved the predictive accuracy of this basic model. For example, the p-value for premedication was 0.55 by multivariable analysis.

When adding the two STAI anxiety predictors to the basic model, the STAI anxiety state was significant (OR: 1.01 per unit change; 95% CI: 1.00-1.02; $p = 0.04$), but the STAI anxiety trait was not (OR: 1.00 per unit change; 95% CI: 0.99-1.01; $p = 0.8$). The ROC area remained 0.72. When the 2 APAIS anxiety predictors were added to the basic model, the APAIS anxiety score was significant (OR: 1.04 per unit change; 95% CI: 1.02-1.05; $p = 0.02$), but the APAIS need for information score was not (OR: 0.97 per unit change; 95% CI: 0.94-1.01; $p = 0.22$). Again, the ROC area remained 0.72. The combined addition of the four anxiety predictors did not improve the predictive accuracy of the basic model. When nausea at 24 hours or vomiting at 24 hours was used as an end-point, the same predictors were found, and the addition of preoperative anxiety as a variable again did not improve the predictive accuracy of the model.

Table 1 Univariable association of each potential predictor and the presence or absence of postoperative nausea and vomiting (PONV) within the first 24 hours after surgery. Values are percentages unless stated otherwise.

variable	total N=1389 (100%)	no PONV N=722 (52%)	PONV N=667 (48%)	odds ratio	p-value
anxiety predictors					
Spielberger state anxiety level[a,b]	40 (33-48)	38 (32-47)	42 (34-50)	1.02	< 0.001
Spielberger trait anxiety level[a,b]	37 (31-44)	36 (31-43)	37 (31-45)	1.01	0.04
APAIS[c] anxiety score[a,b]	9 (6-12)	8 (6-11)	9 (7-12)	1.06	< 0.001
APAIS[c] need for information score[a,b]	7 (5-8)	7 (5-8)	7 (5-8)	1.00	0.84
reported potential predictors					
age (years)[b,d]	45 (15)	47 (15)	43 (15)	0.98	< 0.001
female gender	57	50	65	1.89	< 0.001
ASA physical status					0.16
I	60	58	62	-	
II	36	38	35	0.89	
III	3	4	3	0.63	
current smoking	39	45	33	0.59	< 0.001
history of PONV	25	17	33	2.37	< 0.001
history of motion sickness	19	14	25	2.05	< 0.001
history of PONV or motion sickness	38	28	49	2.41	< 0.001
type of surgery					< 0.001
superficial	74	77	72	-	
upper abdomen	4	4	3	0.82	
lower abdomen	6	4	8	2.08	
laparoscopy	8	8	7	1.07	
strabismus	1	1	1	0.83	
middle ear	7	5	8	1.75	
other	1	1	1	0.87	
lower abdominal or middle ear surgery	13	9	16	1.86	< 0.001
isoflurane anesthesia	50	41	59	1.99	< 0.001
duration of anesthesia (minutes)[a,b]	119 (88-173)	110 (75-168)	127 (89-180)	1.01	< 0.001
gastric tube	19	18	21	0.80	0.09
premedication with benzodiazepines	82	83	81	1.17	0.27
perioperative opioid analgesia	99.6	99.6	99.6	1.01	0.55
opioid analgesia up to 24 hours[e]	63	63	64	1.07	0.55

[a] median (25th-75th percentile)
[b] odds ratio estimated per unit increase
[c] Amsterdam preoperative anxiety and information scale
[d] mean (standard deviation)
[e] morphine intravenous or per epidural
- reference category

Hence, the final prediction model included gender, age, smoking, history of PONV or motion sickness, type of surgery, and anesthetic technique. Table 2 shows the association of each of these predictors with the outcome after adjustment for overoptimism. Validation of the model with bootstrapping techniques yielded an estimated overoptimism in the ROC area of 0.02, such that the ROC area of the final model that might be expected in future patients is 0.70 (95% CI: 0.68-0.72). The calibration of the final model was good (the p-value of the Hosmer and Lemeshow test was 0.45).

With the nomogram (figure 1), one can preoperatively estimate for each future patient the probability of developing PONV within 24 hours after surgery. Alternatively, one can estimate this probability by using the score table in the appendix. Table 3 shows the observed incidence of PONV per risk category of the final model, which ranged from 23% to 76%. Horizontal reading of the table provides estimates of the sensitivity and specificity of the model at different thresholds. For example, when introducing a threshold at a score of 34 (estimated probability of 0.55), a score ≥ 34 (probability ≥ 0.55) should be considered as 'test positive'. When administering pre-emptive anti-emetics to these and only these patients, 72% [(124+166+190)/667] of the patients who indeed developed PONV would be treated correctly (sensitivity or true positive rate), whereas anti-emetics would be correctly withheld (specificity or true negative rate) in 57% [(270+141)/722] of the patients who did not develop PONV. Using a lower threshold at a score of 29 (probability ≥ 0.45), the sensitivity would be 88%, and the specificity would be 37% (figure 2).

Table 2 Final model for preoperative prediction of postoperative nausea and vomiting (PONV) within the first 24 hours after surgery.

variable	regression coefficient (standard error)[a]	odds ratio (95% confidence interval)[a,b]	p-value[a,b]
female gender	0.46 (0.12)	1.58 (1.26 – 2.00)	< 0.001
age (per year)	-0.022 (0.004)	0.98 (0.97 – 0.99)	< 0.001
current smoking	-0.63 (0.12)	0.53 (0.42 – 0.67)	< 0.001
history of PONV or motion sickness	0.76 (0.12)	2.13 (1.69 – 2.70)	< 0.001
lower abdominal or middle ear surgery	0.61 (0.18)	1.84 (1.30 – 2.60)	< 0.001
isoflurane anesthesia	0.72 (0.12)	2.05 (1.63 – 2.57)	< 0.001
intercept (constant)	0.15 (0.22)	-	-

Probability of PONV as estimated by the final model = 1 / (1 + exp (- (0.15 + 0.46*female gender - 0.022*age - 0.63*current smoking + 0.76*history of PONV or motion sickness + 0.61*lower abdominal or middle ear surgery + 0.72*isoflurane anesthesia)))

The area under the receiver operating characteristic curve of the final model was 0.70 (95% confidence interval: 0.68 – 0.72)[a].

[a] after bootstrapping (i.e. adjusted for overfitting)
[b] not relevant for the intercept

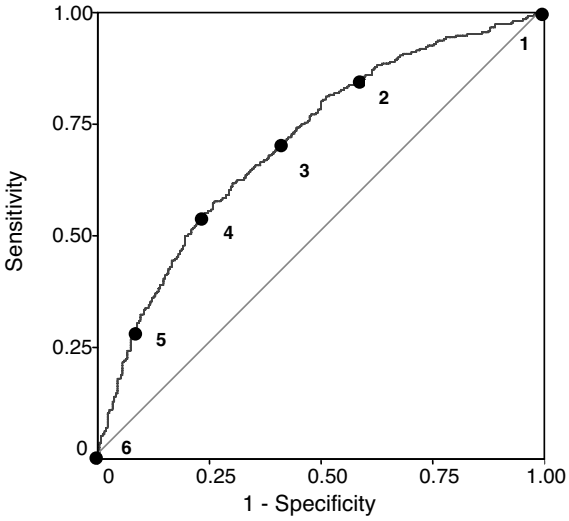

Figure 2 This receiver operating characteristic (ROC) curve graphically presents the prediction model's sensitivity and specificity for all possible cutoff values. The possible cutoff values of table 3 are marked with numbers: 1 indicates probability ≥ 0 or score ≥ 0, sensitivity of 100%, and specificity of 0%; 2 indicates probability ≥ 0.45 or score ≥ 29, sensitivity of 88%, and specificity of 37%; 3 indicates probability ≥ 0.55 or score ≥ 34, sensitivity of 72%, and specificity of 57%; 4 indicates probability ≥ 0.65 or score ≥ 39, sensitivity of 53%, and specificity of 77%; 5 indicates probability ≥ 0.75 or score ≥ 46, sensitivity of 28%, and specificity of 92%; 6 indicates probability = 1, sensitivity of 0%, and specificity of 100%.

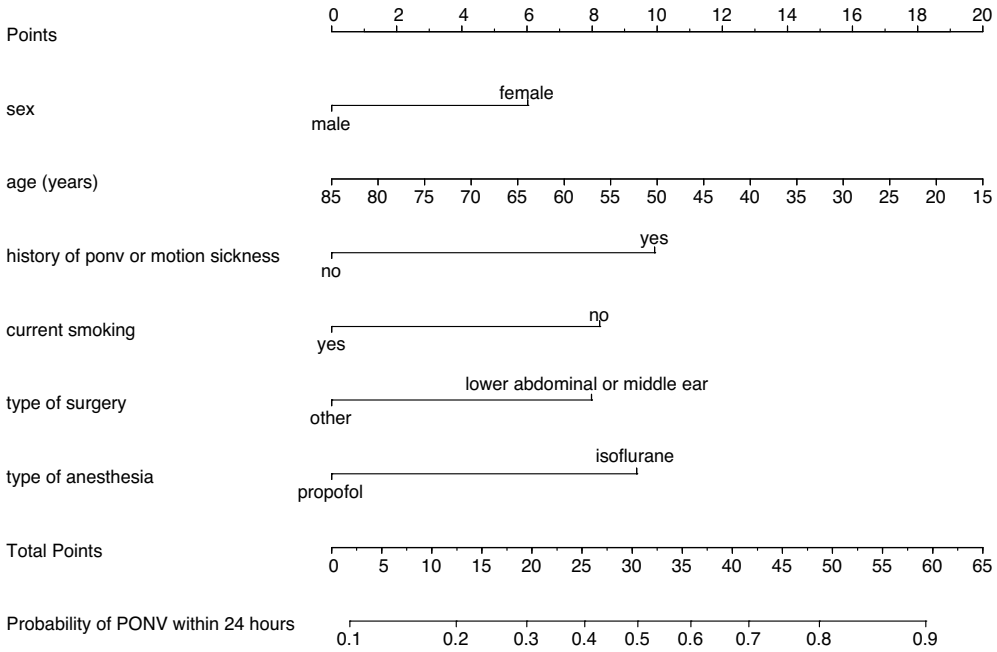

Figure 1 Nomogram relating the predictors of the final model (table 2) to the probability of postoperative nausea and vomiting (PONV) within the first 24 hours after surgery. As an example of using this nomogram, a woman (which corresponds to 6 points as determined from the 'points' scale on the top of the figure), 55 years of age (9 points), who is not currently smoking (8 points), who has had PONV after a previous anesthetic and does not experience motion sickness (10 points), and who is scheduled for lower abdominal surgery (8 points) with total intravenous anesthesia with propofol (0 points), receives a 'total points' score of 41. As shows from the lower two scales of the figure, this score corresponds to a probability of PONV in the first 24 hours after surgery of 0.69. The length of the line of each predictor in the nomogram indicates the relative contribution of the predictor (with the left boundary as anchor) to the probability of PONV.

Table 3 Empirical distribution of patients with and without postoperative nausea and vomiting (PONV) over the score and corresponding probability categories, as estimated with the final prediction model (table 2).

Estimated score[a]	**0 – 28**	**29 – 33**	**34 – 38**	**39 – 45**	**≥ 46**	
Estimated probability[b]	**< 0.45**	**0.45 – 0.54**	**0.55 – 0.64**	**0.65 – 0.74**	**≥ 0.75**	**Total**
Observed % PONV	23	43	46	62	76	48
PONV[c]	80 (12)	107 (16)	124 (19)	166 (25)	190 (29)	667 (100)
No PONV[c]	270 (37)	141 (20)	148 (21)	102 (14)	61 (8)	722 (100)
N[d]	350	248	272	268	251	1389

[a] score categories, as estimated with the nomogram (figure 1) or score table (appendix).
[b] probability categories that correspond to the score categories, as estimated with the final prediction model (table 2), nomogram (figure 1) or score table (appendix).
[c] data are presented as: absolute number of patients (percentage of the 'total' column).
[d] number of patients per score or probability category

Discussion

Although several authors have suggested that preoperative anxiety is a predictor of PONV[1;9], this is the first study that systematically quantified this association while accounting for established predictors of PONV. The results of this study among adult surgical inpatients show that high levels of preoperative anxiety are associated with PONV. However, because this association was relatively weak, the prediction of PONV based on previously reported preoperative predictors was not improved by the addition of preoperative anxiety measures. Although, for several reasons, it may be useful to be informed about the patient's preoperative anxiety status, routine measurement of anxiety to predict the occurrence of PONV does not seem warranted if the other predictors are already considered.

There is evidence from other disciplines that anxiety disorders are associated with nausea in the general community[16]. For a relationship between anxiety and nausea after anesthesia, however, no evidence was available. In a study of 51 unpremedicated children undergoing outpatient surgery and standardized general anesthesia, anxiety was shown to have no predictive value for the occurrence of PONV within 24 hours after surgery[17]. Also, Shafer et al.[18] showed that midazolam premedication significantly reduced preoperative anxiety levels in outpatients, without a parallel decrease in the incidence of PONV. The only data that support the association between anxiety and PONV come from a survey conducted by Quinn et al.[19]. In this survey, patients were questioned about their experiences during the first 24 hours after anesthesia and surgery. However, this study did not use a validated instrument to measure anxiety, and the anesthetic technique and surgical procedure were not standardized. Therefore, the evidence for a link between preoperative anxiety and PONV from this study and previous reports remains weak.

The results of this study confirm that gender, smoking status, and history of PONV or motion sickness independently contribute to the prediction of PONV. Furthermore, the predictive power of these established predictors can significantly be improved by adding the predictors age, anesthetic technique, and type of surgery. In this data set, the ROC area of a model with the predictors gender, history of PONV or motion sickness, smoking, and postoperative opioid analgesia was 0.67. The addition of age, type of surgery, and anesthetic technique, while at the same time eliminating postoperative opioid use, resulted in a significantly higher ROC area of 0.72. Multivariable analyses on our data did not confirm the independent predictive value of postoperative opioid analgesia and duration of anesthesia found in previous

studies[2;4;5;7]. One possible explanation is that the predictive information of these two predictors was already provided for by the type of surgery.

Evidence on type of surgery as an independent predictor is somewhat controversial. We found a significant association between lower abdominal (and middle ear) surgery and PONV, whereas this was not found by Cohen et al.[3]. We did not find an independent predictive effect of laparoscopic surgery, which was in accordance with some previous studies[3;4], but conflicted with another study[20]. For other types of surgery, no independent predictive ability for PONV has been found, and this is supported by our results. It should be noted that the number of subjects in some of our surgical categories was relatively small. A center effect can therefore not be excluded. Hence, the predictive effect of abdominal and middle ear surgery for the occurrence of PONV needs validation in future studies.

Risk scores seem to be the only objective method for a risk-based anti-emetic strategy[21;22], because single predictors such as the type of surgery or the patient's history of PONV alone do not sufficiently predict PONV. Several prediction models for PONV have therefore been proposed. It should be noted that any prediction model, including ours, performs better on the data from which it is developed than in another institution (overoptimism). For this reason, we estimated the amount of overoptimism of our model by using bootstrapping techniques. We adjusted the model for this overoptimism and calculated the ROC area that can be expected in future patients. After bootstrapping, our model had a predictive accuracy of 0.70. Hence, 0.70 is the ROC area that can be expected in new patients. Because this was only 0.02 less than before bootstrapping, the model seems quite robust. External validation studies in new patients may yield additional evidence for the robustness of our prediction model across clinical settings. Admittedly, our model, which includes 6 predictors, is slightly more complex than the previously developed 4-item risk score[2]. Nevertheless, using the nomogram or score table presented in this paper, the probability of PONV in individual patients can easily be estimated, particularly when electronic patient records are used.

In this study, patients received either inhaled anesthesia with isoflurane and nitrous oxide or intravenous anesthesia with propofol. There is evidence that omission of nitrous oxide decreases the incidence of PONV[23]. Furthermore, in some countries, use of isoflurane has decreased in favor of sevoflurane or desflurane, which are frequently used without nitrous oxide. If the PONV incidence would be lower with these newer inhaled anesthetic techniques, the estimated OR (table 2) for use of inhaled anesthesia might also be lower. We realize that changes in the practice of inhaled anesthesia might

compromise the generalizability of the present prediction model. However, several systematic reviews suggest that the PONV incidence with isoflurane does not substantially differ from that with other inhaled anesthetics[24;25]. Furthermore, even if the incidence of the outcome was lower in future patient populations, this might affect the intercept (constant) and, thus, the calibration of the prediction model, but it would not affect the model's discriminative ability. Therefore there is no reason to assume that the generalizability of the model is affected by local preferences for sevoflurane or desflurane over isoflurane, with or without nitrous oxide.

To appreciate the presented results, a few issues need to be addressed. First, although outpatients were included in the trial from which the data originated, for these analyses we deliberately restricted the study population to inpatients, because the PONV incidence in outpatients was substantially lower (34%) and because different types of surgery were performed (e.g. no abdominal surgery). Accordingly, our results should primarily be generalized to inpatients. It should be noted that, currently, no models are available that were developed on both inpatients and outpatients. This is still a subject for future research, particularly given the increase of ambulatory surgery. Second, the PONV incidence in this study was high (48%), but in accordance with previous studies[3;4]. The high incidence found in this study may be partly explained by the frequent assessment of the outcome (4 times at the PACU and each hour at the ward versus at 2 and 24 hours after surgery in other studies) and the wide definition of PONV. However, additional analyses with separate end-points, i.e. nausea only at 24 hours and vomiting only at 24 hours, yielded the same results for anxiety and for the other predictors and had similar predictive accuracy. Finally, we used only outcome data from the early postoperative period (first 24 hours) to conform to previous studies on prediction of PONV. However, additional analyses with the later outcomes (PONV at 72 hours or 14 days) also yielded the same predictors and identical predictive accuracy.

In conclusion, high levels of preoperative anxiety are associated with the occurrence of PONV. However, because this association is weak, routine measurement of preoperative anxiety to aid in the prediction of PONV does not seem warranted, provided that the predictors gender, age, smoking, history of PONV or motion sickness, type of surgery, and anesthetic technique are already considered.

References

1. Watcha MF, White PF. Postoperative nausea and vomiting. Its etiology, treatment, and prevention. *Anesthesiology* 1992;**77**:162-84.
2. Apfel CC, Läärä E, Koivuranta M, Greim C-A, Roewer N. A simplified risk score for predicting postoperative nausea and vomiting: conclusions from cross-validations between two centers. *Anesthesiology* 1999;**91**:693-700.
3. Cohen MM, Duncan PG, DeBoer DP, Tweed WA. The postoperative interview: assessing risk factors for nausea and vomiting. *Anesthesia and Analgesia* 1994;**78**:7-16.
4. Koivuranta M, Läärä E, Snåre L, Alahuhta S. A survey of postoperative nausea and vomiting. *Anaesthesia* 1997;**52**:443-9.
5. Palazzo M, Evans R. Logistic regression analysis of fixed patient factors for postoperative sickness: a model for risk assessment. *British Journal of Anaesthesia* 1993;**70**:135-40.
6. Pierre S, Benais H, Pouymayou J. Apfel›s simplified score may favourably predict the risk of postoperative nausea and vomiting. *Canadian Journal of Anesthesia* 2002;**49**:237-42.
7. Toner CC, Broomhead CJ, Littlejohn IH, Samra GS, Powney JG, Palazzo MGA et al. Prediction of postoperative nausea and vomiting using a logistic regression model. *British Journal of Anaesthesia* 1996;**76**:347-51.
8. Visser K, Hassink EA, Bonsel GJ, Moen J, Kalkman CJ. Randomized controlled trial of total intravenous anesthesia with propofol versus inhalation anesthesia with isoflurane-nitrous oxide: postoperative nausea and vomiting and economic analysis. *Anesthesiology* 2001;**95**:616-26.
9. Lerman J. Surgical and patient factors involved in postoperative nausea and vomiting. *British Journal of Anaesthesia* 1992;**69 (Suppl. I)**:S24-S32.
10. Spielberger CD, Gorush RL, Lushene RE. Manual for the State-Trait Anxiety Inventory . Palo Alto, California: Consulting Psychologists Press, 1970.
11. Moerman N, Van Dam FSAM, Muller MJ, Oosting H. The Amsterdam Preoperative Anxiety and Information Scale (APAIS). *Anesthesia and Analgesia* 1996;**82**:445-51.
12. Hosmer D, Lemeshow S. Applied logistic regression. New York: John Wiley & Sons, Inc., 1989.
13. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* 1996;**15**:361-87.
14. Steyerberg EW, Harrell FE, Borsboom GJJM, Eijkemans MJC, Vergouwe Y, Habbema JDF. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *Journal of Clinical Epidemiology* 2001;**54**:774-81.
15. Efron B, Tibshirani RJ. An introduction to the bootstrap. New York: Chapman and Hall, 1993.

16. Haug TT, Mykletun A, Dahl AA. The prevalence of nausea in the community: psychological, social and somatic factors. *General Hospital Psychiatry* 2002;**24**:81-6.
17. Wang S-M, Kain ZN. Preoperative anxiety and postoperative nausea and vomiting in children: is there and association? *Anesthesia and Analgesia* 2000;**90**:571-5.
18. Shafer A, White PF, Urquhart ML, Doze VA. Outpatient premedication: use of midazolam and opioid analgesics. *Anesthesiology* 1989;**71**:495-501.
19. Quinn AC, Brown JH, Wallace PG, Asbury AJ. Studies in postoperative sequelae. Nausea and vomiting - still a problem. *Anaesthesia* 1994;**49**:62-5.
20. Apfel CC, Greim C-A, Haubitz I, Goepfert C, Usadel J, Sefrin P et al. A risk score to predict the probability of postoperative vomiting in adults. *Acta Anaesthesiologica Scandinavica* 1998;**42**:495-501.
21. White PF, Watcha MF. Postoperative nausea and vomiting: prophylaxis versus treatment. *Anesthesia and Analgesia* 1999;**89**:1337-9.
22. Watcha MF. The cost-effective management of postoperative nausea and vomiting. *Anesthesiology* 2000;**92**:931-3.
23. Hartung J. Twenty-four of twenty-seven studies show a greater incidence of emesis associated with nitrous oxide than with alternative anesthetics. *Anesthesia and Analgesia* 2004;**83**:114-6.
24. Sneyd JR, Carr A, Byrom WD, Bilski AJT. A meta-analysis of nausea and vomiting following maintenance of anaesthesia with propofol or inhalational agents. *European Journal of Anaesthesiology* 1998;**15**:433-45.
25. Tramèr MR, Moore RA, McQuay HJ. Propofol anaesthesia and postoperative nausea and vomiting: quantitative systematic review of randomized controlled studies. *British Journal of Anaesthesia* 1997;**78**:247-55.

Appendix

The assigned points per value of each predictor included in the final prediction model for postoperative nausea and vomiting (PONV) (table 2).

Gender	**Points[a]**
male	0
female	6
Age (years)	
15 – 19	20
20 – 24	19
25 – 29	17
30 – 34	16
35 – 39	14
40 – 44	13
45 – 49	11
50 – 54	10
55 – 59	9
60 – 64	7
65 – 69	6
70 – 74	4
75 – 79	3
80 – 84	1
85	0

Current smoking	**Points[a]**
no	8
yes	0
History of PONV or motion sickness	
no	0
yes	10
Surgery type	
lower abdominal or middle ear	8
other	0
Anesthetic technique	
propofol	0
isoflurane	9

Total Points	**Probability of PONV**
2	0.10
12	0.20
19	0.30
25	0.40
31	0.50
36	0.60
42	0.70
49	0.80
59	0.90

[a] The points are assigned by multiplying the adjusted regression coefficient (table 2, second column) of the predictor times 14 and rounded to the nearest integer. The model's intercept (constant) is already taken into account in the score assigned to the predictor age.

Chapter 5

Prediction of PONV in surgical inpatients:

which model to use in clinical practice?

JE van den Bosch, Y Vergouwe, WA van Klei, L van Wolfswinkel,
CJ Kalkman, KGM Moons

(submitted)

Abstract

We previously validated two prediction models for postoperative nausea and vomiting (PONV) in a large Dutch validation sample of surgical inpatients and found limited predictive accuracy. Prompted by these disappointing validation results, we developed a new model using data from the validation sample. The new model contains mostly the same predictors as the two previously developed models. The present study tested the predictive accuracy of the new model in patients from another Dutch hospital (geographical validity) and a later time period (temporal validity). The two previously developed prediction models for PONV were also tested in this sample. For the new model, the area under the receiver operating characteristic (ROC) curve was 0.68 in the validation sample compared to 0.70 in the development sample. The ROC areas of the other two models were 0.69 and 0.70. The new model showed better calibration than the two previously developed models. None of the tested models outperformed the other models both in terms of calibration and discrimination. Nonetheless, all models showed much higher predictive accuracy than a single predictor approach in which the type of surgery or previous occurrence of PONV is used to guide anti-emetic prophylaxis. Thus, it is more important that the clinician actually uses one of the available prediction models for PONV than that the clinician selects a particular model.

Introduction

The incidence of postoperative nausea and vomiting (PONV) remains as high as 25-30% in surgical inpatients[1;2] and may be reduced by risk-based prophylaxis. To identify patients at high risk of PONV, several models for predicting PONV in inpatients have been proposed[3-6]. The two best-known models are those developed by Apfel et al. (Germany)[3] and by Koivuranta et al. (Finland)[5]. Both models showed good predictive accuracy in the development samples, with areas under the receiver operating characteristic (ROC) curve of 0.75[3] and 0.72[5], respectively. However, when validated in a large Dutch sample of surgical inpatients[7], the calibration and discrimination of these models were less than expected based on results from previous validation studies[8-10]. The two models provided risk predictions that were too extreme, which means that low risk predictions were estimated too low and high risk predictions were estimated too high[7]. Therefore, a new model was developed, using data from the Dutch validation sample. The new model contained six predictors (gender, age, smoking, history of PONV or motion sickness, type of surgery, and type of anesthesia)[11]. The area under the ROC curve was 0.70 (95% confidence interval [CI]: 0.68 – 0.72) after correction for overoptimism.

The present study tested the predictive accuracy of the new model in a large validation sample of surgical inpatients from an other Dutch hospital (geographical validity)[12] and a later time period (temporal validity)[12]. For comparison, the models developed by Apfel et al. and Koivuranta et al. were also validated in this sample.

Methods

Prediction models under study

In the present study, the predictive accuracy of the model for predicting PONV within 24 hours after surgery, developed by Van den Bosch et al.[11], was tested in a large validation sample of Dutch surgical inpatients. For comparison, the models developed by Apfel et al.[3] and Koivuranta et al.[5] were also validated in this sample. The predictors and formulas of the tested prediction models are shown in the appendix.

The development sample of the Van den Bosch model

The development sample of the Van den Bosch model[11] included 1389 surgical inpatients, aged 18-85 years, who participated in a randomized trial (AIDA trial)[13]. This trial was designed to investigate differences in PONV incidence after inhaled general anesthesia (with isoflurane/nitrous oxide) or intravenous anesthesia (with propofol). All types of surgery were included except for cardiac surgery and intracranial surgery. Exclusion criteria were emergency surgery, (possible) pregnancy, ASA class IV, morbid obesity (weight > 120 kg), renal or liver disease precluding use of either anesthetic technique, and use of anti-emetic or pro-emetic medication in the 2 weeks before surgery. Patients receiving regional anesthetic techniques were not included in the study, except in the case of upper abdominal surgery, for which epidural analgesic supplementation was permitted.

PONV was defined as at least one episode of nausea (any degree, including mild nausea) or retching or vomiting within the first 24 hours after surgery. In the postanesthesia care unit (PACU), nausea (mild/moderate/severe), retching and vomiting were recorded at four time points: on arrival, one hour after arrival, at the time of permission for discharge and at the time of actual discharge. At the ward, symptoms were recorded each hour up to 24 hours after surgery. The incidence of PONV was 48%.

Validation sample

The validation sample included 1203 surgical inpatients aged 18-89 years, from whom data were collected during a 7 months period from June 2004 until January 2005 at the University Medical Center of Utrecht (UMCU), The Netherlands. All types of surgery were included, except cardiac surgery and intracranial surgery. Other exclusion criteria were emergency surgery, ASA

physical status IV, insufficient command of the Dutch language and inability to communicate information on predictors or outcome.

For each patient, data were collected on the predictors included in the Van den Bosch model, namely gender, age (years), smoking (yes/no), history of PONV or motion sickness (yes/no), the type of surgery (lower abdominal or middle ear versus other types) and the type of anesthesia (propofol versus isoflurane anesthesia). In addition, information was obtained on use of nitrous oxide, administration of prophylactic or therapeutic anti-emetics, and postoperative opioid use within 24 hours after surgery. In the development sample, inhalation anesthesia was always performed with isoflurane, whereas, in the validation sample, sevoflurane was also used.

Patients were interviewed twice at the postanesthesia care unit (PACU) at 15 minutes and 60 minutes after surgery by PACU nurses. The day after surgery, patients were interviewed once at the ward by trained research nurses. The nurses recorded nausea (on a 3-point scale: no nausea / mild nausea / severe nausea), vomiting, and administration of anti-emetic medication. Patients were considered to have had PONV if at least one episode of postoperative nausea or vomiting had occurred in the first 24 hours after surgery.

The study protocol was approved by the medical ethics committee of the UMCU. The institutional medical ethics committee waived the need for written informed consent, because the Dutch law on Medical Research Involving Human Subjects (WMO) was not applicable, since patients did not receive a particular treatment and were not asked to behave in a certain way.

Statistical analysis

The prediction model for PONV, developed by Van den Bosch et al., was tested by calculating the risk of PONV occurrence within 24 hours after surgery for each patient in the validation sample with the logistic regression formula presented in the appendix. For comparison, the risk was also calculated with the models developed by Apfel et al. and Koivuranta et al. (see appendix). The predictive accuracy of each model was quantified with measures of discrimination and calibration.

Discrimination refers to the ability of a prediction model to distinguish between patients with and without PONV. The discrimination of each model in the validation sample was studied with the ROC area, which may theoretically range from 0.5 (discrimination no better than chance) to 1.0 (perfect discrimination)[14].

Calibration of a model refers to the agreement between the predicted risks and observed outcome frequencies. The calibration of each model was tested using the Hosmer-Lemeshow test[15], for which a significant test result indicates poor calibration, and studied with a calibration plot. Each calibration plot shows a smoothed calibration line that represents the agreement between the predicted risks (x-axis) and observed outcomes (y-axis) in the validation sample. This calibration line can be described with an intercept and a slope, which are ideally 0 and 1, respectively (perfect calibration). These two parameters are estimated with a logistic regression model that is fitted with PONV occurrence (yes/no) as the outcome and the linear predictor as the only covariate[16]. The linear predictor is the sum of the intercept and the regression coefficients multiplied by the predictor values observed for each patient in the validation sample (see appendix). The intercept of the calibration line was estimated with the slope of the line fixed at 1[17]. An intercept substantially different from 0 commonly reflects a difference in outcome incidence between the validation and development sample, that can not be explained by different distributions of the predictors in the model[18]. A slope substantially different from 1 (e.g. < 0.8 or > 1.2) indicates that the associations between the predictors and outcome (represented by the regression coefficients) in the validation sample are different from the associations estimated in the development sample.

Of all predictor and outcome values, 5% were missing, mainly of the predictors ‘history of PONV’ and ‘history of motion sickness’. Deleting subjects with missing values leads to bias and a loss of statistical power[19;20]. Therefore, missing values were imputed with imputation techniques (standard available in S-plus software version 6.2). Analyses were performed with S-PLUS 6.2 Professional Edition for Windows.

Results

Table 1 shows the distribution of predictor and outcome values for the validation sample compared to the three development samples. The samples were fairly comparable, apart from disparity in the percentage of females (relatively high in Koivuranta et al.'s development sample), smokers (varies between 22% and 39%), patients with a history of motion sickness (only 9% in the validation sample) and inhalation anesthesia (varies between 43% and 100%). The duration of the procedure in the validation sample and in Koivuranta et al.'s development sample was shorter than in the two other development samples.

Figure 1 shows the calibration lines of the prediction models developed by Van den Bosch et al. (figure 1a), Apfel et al. (figure 1b) and Koivuranta et al. (figure 1c) when applied to patients of the validation sample. The model developed by Van den Bosch et al. was better calibrated in the validation sample (intercept: - 0.03 and slope: 0.82) than the models developed by Apfel et al. and Koivuranta et al. (intercepts of 0.25 and 0.24 and slopes of 0.74 and 0.81, respectively), whereas the discriminative ability of the model was marginally lower (ROC areas: 0.68 versus 0.69 and 0.70) (table 2). The difference in calibration was confirmed by the Hosmer-Lemeshow test (p-value of 0.25 versus p-values < 0.001, respectively) (table 2) and by an overall test with the null hypothesis that the intercept equals 0 and the slope equals 1 (Chi-square: 5.27 and p-value: 0.07 versus Chi-squares of 29.01 and 21.46 and p-values < 0.001, respectively).

The ROC curves of the three models are shown in figure 2. The curve of the model developed by Koivuranta et al. lies above the curves of the two other models in the probability range that corresponds to the sensitivity / 1-specificity pairs of 0.87 / (1 - 0.32) and 0.49 / (1 - 0.77). This means that the model developed by Koivuranta et al. has a slightly higher discriminative ability for patients with a predicted risk between 0.20 – 0.50.

Table 1 Distribution of predictor and outcome values in the development samples of the three inpatient models (Van den Bosch et al.[11], Apfel et al.[3] and Koivuranta et al.[5]) and the inpatient validation sample. Values are proportion, median (interquartile range) or median [range].

	development samples			validation sample
	Van den Bosch et al.[11] (N = 1389)	**Apfel et al.[3 a] (N = 1040)**	**Koivuranta et al.[5] (N = 1107)**	**(N = 1203)**
predictors of PONV[b]				
female gender	57	57	66	52
age (years)	44 (32 - 56)	49 (36 - 61)	46 [4 - 86]	50 (37 - 62)
current smoking	39	27	22	33
history of PONV[b]	25	-	43[c]	29
history of motion sickness	19	-	24	9
history of PONV[b] or motion sickness	38	35	-	35
lower abdominal or middle ear surgery	13	-	-	13
duration of surgery (minutes)	120 (80 - 173)	116 (73 - 152)	72 (40 - 105)	85 (53 - 132)
duration of surgery > 60 minutes	90	-	-	67
inhalation anesthesia	50[d]	100	74	43
use of nitrous oxide	50[d]	-	-	14
administration of prophylactic anti-emetics	0	0	-	7
postoperative anti-emetics up to 24 hours	12	-	31	18
opioid analgesics up to 24 hours	63	46	71	63
outcome				
PONV[b] within 24 hours after surgery	48	44	52	44

[a] the sample contained 520 patients from a medical center in Oulu and 520 patients from a medical center in Wuerzburg. The original paper presents the data for both patient groups separately. We estimated the values for the combined sample by averaging the numbers presented for the two original data sets

[b] PONV = postoperative nausea and vomiting

[c] 43% of patients experienced PONV after previous general anesthesia (73% of patients), whereas 11 % experienced PONV after previous regional anesthesia (91% of patients)

[d] the Van den Bosch model was developed with data from a trial in which the anesthetic technique (intravenous anesthesia with propofol versus inhalation anesthesia with use of nitrous oxide) was randomly assigned

- not presented in the original publication

Table 2 Performance measures for the prediction model developed by Van den Bosch et al.[11] when applied to patients of the validation sample. Performance measures for the models developed by Apfel et al.[3] and Koivuranta et al.[5] are shown for comparison.

model developed by	discrimination	calibration		
	AUC[a] (95% CI[b])	intercept with slope fixed at 1	slope	p-value[c]
Van den Bosch et al.[11]	0.68 (0.65 – 0.71)	- 0.03	0.82	0.25
Apfel et al.[3]	0.69 (0.66 – 0.72)	0.25	0.74	< 0.001
Koivuranta et al.[5]	0.70 (0.67 – 0.73)	0.24	0.81	< 0.001

[a] area under the receiver operating characteristic curve
[b] 95% confidence interval
[c] p-value of Hosmer-Lemeshow test; p-values < 0.05 indicate poor calibration

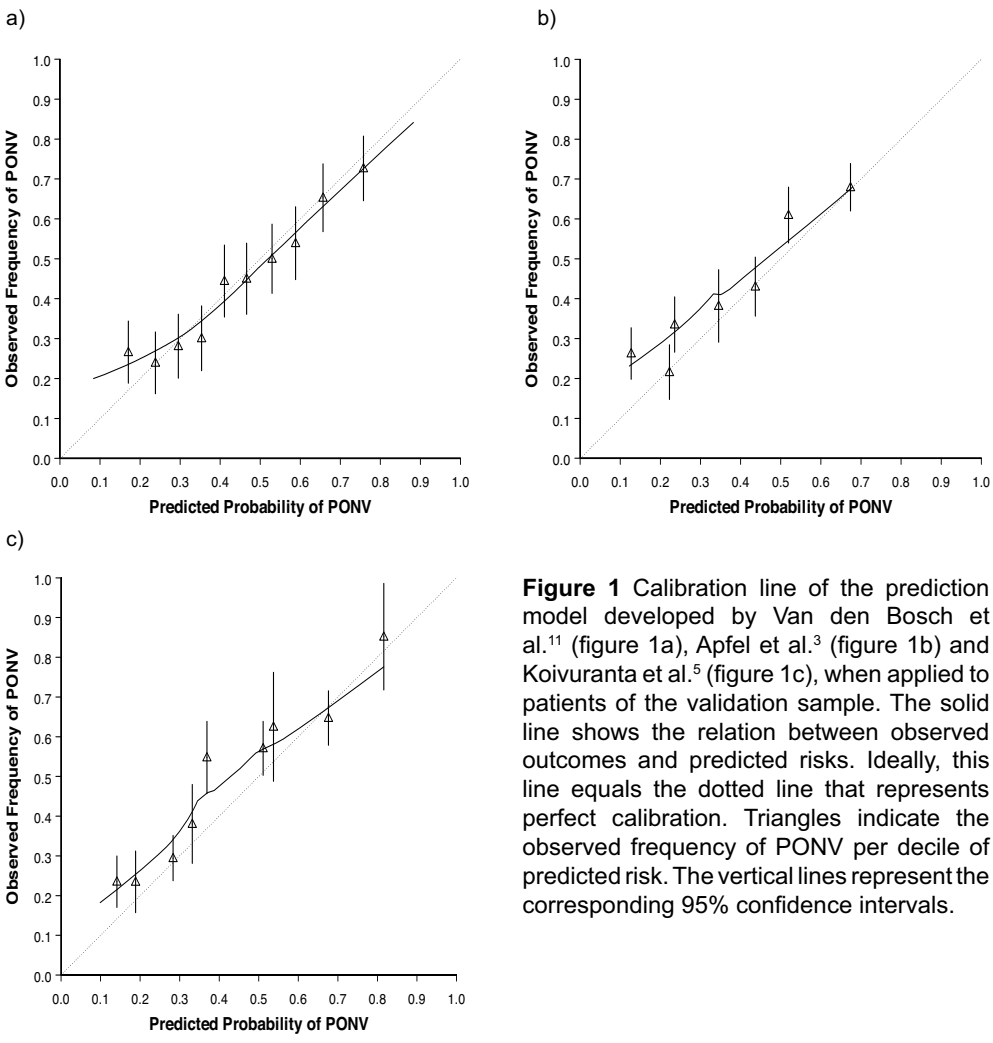

Figure 1 Calibration line of the prediction model developed by Van den Bosch et al.[11] (figure 1a), Apfel et al.[3] (figure 1b) and Koivuranta et al.[5] (figure 1c), when applied to patients of the validation sample. The solid line shows the relation between observed outcomes and predicted risks. Ideally, this line equals the dotted line that represents perfect calibration. Triangles indicate the observed frequency of PONV per decile of predicted risk. The vertical lines represent the corresponding 95% confidence intervals.

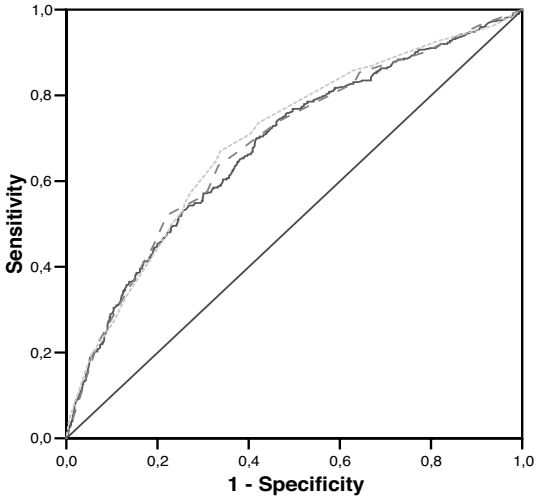

Figure 2 Receiver operating characteristic (ROC) curve of the prediction model developed by Van den Bosch et al.[11] (solid line), Apfel et al.[3] (dashed line) and Koivuranta et al.[5] (dotted line), when applied to patients of the validation sample. The diagonal line represents a ROC area of 0.5 (discrimination no better than chance).

Discussion

The present study tested the predictive accuracy of a prediction model for PONV, developed in a large sample of Dutch surgical inpatients, in a validation sample of surgical inpatients from another Dutch hospital and a later time period. In the validation sample, the ROC area was somewhat decreased (0.68 instead of 0.70). Compared to two previously developed models, which were also tested in the validation sample, the discriminative ability was slightly lower (ROC area: 0.68 compared to 0.69 and 0.70), but the model showed better calibration (intercept of -0.03 compared to 0.25 and 0.24; slope of 0.82 compared to 0.74 and 0.81).

One may ask whether one of the three tested prediction models for PONV is preferable for use in clinical practice. In this decision several factors play a role, such as the intended use of the model, anticipation on possible changes in anesthetic practice and the user-friendliness of the model.

When the intended use of the model concerns risk stratification of patients in clinical trials, model selection may occur based on discriminative ability. In such cases, the model developed by Koivuranta et al. may be the preferred model. The discriminative ability of this model has shown to be most robust across different populations, both in the present and previous validation studies[7-9]. When the model is intended for use by clinicians to estimate a patient's absolute risk of PONV, model calibration might play an important role in choosing a model. Models that are well calibrated yield the most accurate predictions of absolute PONV risks. In this validation study, the model developed by Van den Bosch et al. showed better calibration than the models developed by Apfel et al. and Koivuranta et al.

A patient's risk of PONV may vary depending on the choice of the anesthetic technique, as intravenous anesthesia with propofol reduces the incidence of PONV compared to inhalation anesthesia[13;21]. Intravenous anesthesia with propofol used to be considerably more expensive than inhalation anesthesia with isoflurane[13]. However, the acquisition costs of propofol have decreased and the price difference between generic propofol emulsions and sevoflurane has become negligible. As a result, intravenous anesthesia with propofol may become more popular in future. A possible resulting decrease in PONV incidence may influence the calibration of the models developed by Apfel et al. and Koivuranta et al., but will likely not affect the calibration of the model developed by Van den Bosch et al., since anesthetic technique (defined as inhalation anesthesia versus intravenous anesthesia) is a predictor in the latter model.

All models are available both as a logistic regression formula and in a simplified form, such as a simplified risk score (models developed by Apfel

et al. and Koivuranta et al.), score table or nomogram (model developed by Van den Bosch et al). Therefore, none of the models can be preferred over the others based on user-friendliness.

What is more important than which of the available models a clinician prefers, is that a clinician at least uses one of the available prediction models for PONV in clinical practice. Many clinicians still use a single predictor such as the type of surgery or previous occurrence of severe PONV to decide whether or not to administer prophylactic anti-emetics. The predictive accuracy of all three tested inpatient prediction models for PONV has proven to be superior to this approach, both in the present validation study and in previously conducted validation studies[8-10]. The ROC area for the predictor type of surgery was found to be 0.53 (95% CI: 0.50 – 0.56) and that for previous occurrence of PONV 0.58 (95% CI: 0.56-0.61)[22]. In contrast, several validation studies reported that the available multivariable models[3;5;11] showed ROC areas of at least 0.63[7] and often 0.65 or higher[7-10]. Therefore, objective risk estimation with multivariable prediction models for PONV should be preferred over a single factor approach that is still used by many clinicians, even considering that the predictive accuracy of available multivariable models remains limited.

This raises the question whether the predictive accuracy of the available prediction models for PONV can be further improved. Apfel et al. suggested that the discriminative power of PONV prediction models for inpatients will remain imperfect unless new predictors with a strong impact (i.e. odds ratios in the range 2 to 3) are discovered and confirmed in several centers[23]. Our previous study showed that preoperative anxiety is not a promising candidate[11]. A possible new predictor may be a genetic factor that influences the activity of liver enzymes. It is known that genetically determined differences in the activity of liver enzymes may cause individual differences in the metabolic breakdown of anesthetic agents[24]. This may influence patients' liability to experience PONV, since volatile anesthetics are an important cause of PONV within the first two hours after surgery[25]. So far, it has not been possible to model this genetic factor, but in future, knowledge on the genetic risk factors of PONV may possibly be used for PONV prediction. Until then, the discriminative ability of available models will likely remain in the range between 0.63 and 0.70. That the discriminative ability of these models is limited does not mean that these models are useless, however, since these models predict patients' absolute risks of PONV fairly accurately.

In conclusion, none of the tested models outperformed the other models both in terms of calibration and discrimination. Nonetheless, all models showed much higher predictive accuracy than a single predictor approach in which the type of surgery or previous occurrence of PONV is used to guide anti-emetic prophylaxis. Thus, it is more important that the clinician actually uses one of the available prediction models for PONV than that the clinician selects a particular model.

References

1. Kovac AL. Prevention and treatment of postoperative nausea and vomiting. *Drugs* 2000;**59**:213-43.
2. Watcha MF, White PF. Postoperative nausea and vomiting. Its etiology, treatment and prevention. *Anesthesiology* 1992;**77**:162-84.
3. Apfel CC, Läärä E, Koivuranta M, Greim C-A, Roewer N. A simplified risk score for predicting postoperative nausea and vomiting: conclusions from cross-validations between two centers. *Anesthesiology* 1999;**91**:693-700.
4. Gan TJ. Postoperative nausea and vomiting - can it be eliminated ? *JAMA: the journal of the American Medical Association* 2002;**287**:1233-6.
5. Koivuranta M, Läärä E, Snåre L, Alahuhta S. A survey of postoperative nausea and vomiting. *Anaesthesia* 1997;**52**:443-9.
6. Palazzo M, Evans R. Logistic regression analysis of fixed patient factors for postoperative sickness: a model for risk assessment. *British Journal of Anaesthesia* 1993;**70**:135-40.
7. Van den Bosch JE, Kalkman CJ, Vergouwe Y, Van Klei WA, Bonsel GJ, Grobbee DE et al. Assessing the applicability of scoring systems for predicting postoperative nausea and vomiting. *Anaesthesia* 2005;**60**:323-31.
8. Apfel CC, Kranke P, Eberhart LHJ, Roos A, Roewer N. Comparison of predictive models for postoperative nausea and vomiting. *British Journal of Anaesthesia* 2002;**88**:234-40.
9. Eberhart LHJ, Högel J, Seeling W, Staack AM, Geldner G, Georgieff M. Evaluation of three risk scores to predict postoperative nausea and vomiting. *Acta Anaesthesiologica Scandinavica* 2000;**44**:480-8.
10. Pierre S, Benais H, Pouymayou J. Apfel's simplified score may favourably predict the risk of postoperative nausea and vomiting. *Canadian Journal of Anesthesia* 2002;**49**:237-42.
11. Van den Bosch JE, Moons KG, Bonsel GJ, Kalkman CJ. Does measurement of preoperative anxiety have added value for predicting postoperative nausea and vomiting? *Anesthesia and Analgesia* 2005;**100**:1525-32.
12. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Annals of Internal Medicine* 1999;**130**:524.
13. Visser K, Hassink EA, Bonsel GJ, Moen J, Kalkman CJ. Randomized controlled trial of total intravenous anesthesia with propofol versus inhalation anesthesia with isoflurane-nitrous oxide: postoperative nausea and vomiting and economic analysis. *Anesthesiology* 2001;**95**:616-26.
14. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;**143**:29-36.
15. Hosmer D, Lemeshow S. Applied logistic regression. New York: John Wiley & Sons, Inc., 1989.

16. Spiegelhalter DJ. Probabilistic prediction in patient management and clinical trials. *Statistics in Medicine* 1986;**5**:421-33.
17. Miller ME, Hui SL, Tierney W. Validation techniques for logistic regression models. *Statistics in Medicine* 1991;**10**:1213-26.
18. Vergouwe Y, Steyerberg EW, Foster RS, Habbema JD, Donohue JP. Validation of a prediction model and its predictors for the histology of residual masses in nonseminomatous testicular cancer. *Journal of Urology* 2001;**165**:84-8.
19. Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. *American Journal of Epidemiology* 1995;**142**:1255-64.
20. Little RJA. Regression with missing X's: a review. *Journal of the American Statistical Association* 1992;**87**:1227-37.
21. Apfel CC, Kortilla K, Abdallah M, Kerger H, Turan A, Vedder I et al. A factorial trial of six interventions for the prevention of postoperative nausea and vomiting. *The New England Journal of Medicine* 2004;**350**:2441-51.
22. Apfel CC, Kranke P, Eberhart LHJ. Comparison of surgical site and patient's history with a simplified risk score for the prediction of postoperative nausea and vomiting. *Anaesthesia* 2004;**59**:1078-82.
23. Apfel CC, Kranke P, Greim C-A, Roewer N. What can be expected from risk scores for predicting postoperative nausea and vomiting? *British Journal of Anaesthesia* 2001;**86**:822-7.
24. Sweeney BP. Why does smoking protect against PONV? *British Journal of Anaesthesia* 2002;**89**:810-3.
25. Apfel CC, Kranke P, Katz MH, Goepfert C, Papenfuss T, Rauch S et al. Volatile anaesthetics may be the main cause of early but not delayed postoperative vomiting: a randomized controlled trial of factorial design. *British Journal of Anaesthesia* 2002;**88**:659-68.

Appendix

The formulas of the three tested inpatient prediction models for postoperative nausea and vomiting (PONV) within 24 hours after surgery are:

Van den Bosch et al.[11]
Risk of PONV = 1 / (1 + exp (- (0.15 + 0.46*female gender – 0.022*age – 0.63*current smoking + 0.76*history of PONV or motion sickness + 0.61*lower abdominal or middle ear surgery + 0.72*isoflurane anesthesia))) (1)

Apfel et al.[3]
Risk of PONV = 1 / (1 + exp (- (-2.28 + 1.27*female gender + 0.65*history of PONV or motion sickness + 0.72* non-smoking + 0.78*postoperative opioid use))) (2)

Koivuranta et al.[5]
Risk of PONV = 1 / (1 + exp (- (-2.21 + 0.93*female gender + 0.82*history of PONV + 0.59* history of motion sickness + 0.61*non-smoking + 0.75*duration of surgery over 60 minutes))) (3)

All three models are logistic regression models. Logistic regression modeling is used to model the relation between predictor variables and binary outcomes (e.g. PONV occurrence yes/no). The general formula of such models is:

$$\text{risk of binary outcome} = 1 / (1 + e^{-\text{linear predictor}})$$

The linear predictor is the sum of the intercept and the regression coefficients (weights) multiplied by the predictor values observed for each patient. The linear predictor of a model can be written as:

$$\text{linear predictor} = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \ldots + \beta_z * X_z$$

where β_0 represents the intercept, X_1 to X_z represent the predictor values, and β_1 to β_z represent the regression coefficients of each corresponding predictor.

Chapter 6

Can prediction models for postoperative nausea and vomiting be generalized from surgical inpatients to outpatients?

JE van den Bosch, WA van Klei, Y Vergouwe, L van Wolfswinkel, KGM Moons, CJ Kalkman

(submitted)

Abstract

As the number of outpatient surgeries has increased in the last decade, risk-based prophylaxis of postoperative nausea and vomiting (PONV) should not be limited to inpatients but be applied to outpatients as well. Therefore, we tested whether three prediction models for PONV, developed in inpatients, were generalizable to outpatients. A single prediction model developed in outpatients was tested for comparison. We hypothesized that the inpatient models would have similar predictive accuracy in outpatients, under the provision that small adjustments might be needed to account for the lower absolute PONV incidence in outpatients. The predictive accuracy of the models was evaluated in terms of the agreement between observed and predicted outcomes (i.e. calibration, which is studied with a calibration plot) and the ability to discriminate between patients with and without postoperative nausea and vomiting (i.e. discrimination, which is measured with the area under the receiver operating characteristic [ROC] curve). In outpatients, two of the inpatient models showed calibration slopes substantially lower than 1 (0.53 and 0.55 respectively), which indicates poor calibration. More than small adjustments would be needed before these models are generalizable to outpatients. The third inpatient model showed a relatively high slope value (0.85). After intercept adjustment, this inpatient model showed reasonable calibration in the outpatient validation sample. The ROC area of the third model was 0.64 compared to 0.62 and 0.64 with the other two inpatient models. The predictive accuracy of the tested outpatient model was surprisingly low (calibration slope: 0.32, ROC area: 0.60). In conclusion, although all models showed limited discriminative ability in outpatients, one model yielded reasonably calibrated predictions of PONV risk in outpatients. This model may be generalized to outpatients (after a small adjustment).

Introduction

Postoperative nausea and vomiting (PONV) are two of the most important concerns for patients after ambulatory surgery[1]. Persistent PONV increases the risk of unanticipated hospital admission[2-4] and may delay a patient's discharge[4;5] and return to work. As the number of outpatient surgeries has increased in the last decade, it is important that risk-based PONV prophylaxis is not limited to inpatients but is also applied to outpatients.

Several prediction models for PONV[5-10] have been proposed to guide risk-based PONV prophylaxis. Unfortunately, all these models -except one- were developed and validated in inpatient populations[6-10]. The single PONV prediction model that has specifically been developed in outpatients[5] has gained limited attention in the literature and has never been validated in another outpatient population.

Whether inpatient models predict PONV occurrence accurately in outpatients, is yet unknown. What is known, is that the incidence of PONV in outpatients is lower than in inpatients[5]. Consequently, the risk predictions obtained with inpatient models may be systematically too high for outpatients. Therefore, the generalizability of prediction models for PONV, developed in inpatients, needs to be tested in outpatient populations before use in ambulatory surgery can be recommended.

Our hypothesis was that prediction models for PONV, developed in inpatients, would have similar predictive accuracy in outpatients, under the provision that the models might need small adjustments to account for the lower absolute PONV incidence in outpatients. Therefore, we tested the generalizability of available inpatient models[6;8;10] to a sample of outpatients. The single prediction model developed in outpatients[5] was tested for comparison.

Methods

Prediction models under study

The predictive accuracy of three inpatient models for predicting PONV within 24 hours after surgery, developed by Koivuranta et al.[8], Apfel et al.[6] and Van den Bosch et al.[10], and the outpatient model for predicting PONV within 24 hours after surgery, developed by Sinclair et al.[5], was tested in a validation sample of outpatients. The predictors and formulas of the tested prediction models are shown in the appendix.

Validation sample

The validation sample included 623 consecutive outpatients from routine care, aged 18-84 years, who were scheduled for ambulatory surgery with general anesthesia in the period June 2004 to January 2005 in the University Medical Center of Utrecht (UMCU), The Netherlands. Patients were excluded from the study if they received regional anesthesia only or if outcome assessment was not possible due to deafness, mental retardation, psychiatric disorders or insufficient command of the Dutch language.

Data on patient characteristics, on the predictor variables included in the four tested prediction models, and on intra-operative variables (nitrous oxide, prophylactic anti-emetics and duration of surgical procedure) were systematically documented with an electronic patient record. Data on the occurrence of PONV were systematically documented using a standardized record form.

PONV was defined as at least one episode of nausea (mild/severe) or vomiting occurring within the first 24 hours after surgery. The nursing staff of the postanesthesia care unit (PACU) was instructed on the definition of PONV in a separate meeting before the start of the study. The PACU nursing staff recorded nausea, vomiting and administration of anti-emetic medication in the PACU on arrival, at 15 minutes, 30 minutes and 45 minutes after arrival and at time of discharge from the PACU. The day after surgery, standardized telephone interviews were conducted by trained PACU nurses. Patients were not interviewed if they could not be reached on two attempts.

The study protocol was approved by the medical ethics committee of the UMCU. The institutional medical ethics committee waived the need for written informed consent, because the Dutch law on Medical Research Involving Human Subjects (WMO) was not applicable, since patients did not receive a particular treatment and were not asked to behave in a certain way.

Statistical analysis

All four prediction models (formulas presented in the appendix) were applied to each patient in the validation sample to calculate the predicted risks of PONV occurrence within 24 hours after surgery. The predictive accuracy of each model was quantified with measures of calibration and discrimination.

Calibration of a model refers to the agreement between the predicted risks and observed outcome frequencies. The calibration of each model was studied with a calibration plot[11]. A calibration plot shows a smoothed calibration line that represents the agreement between the predicted risks (x-axis) and observed outcomes (y-axis) in the validation sample. This calibration line can be described with a slope and an intercept. The slope is estimated with a logistic regression model that is fitted with PONV occurrence (yes/no) as the outcome and the linear predictor (see appendix) as the only covariate[12]. The regression coefficient of the linear predictor equals the slope. The intercept of the calibration line was estimated with the same logistic regression model with the slope of the line fixed at 1[13]. The calibration of each model was also studied with the Hosmer-Lemeshow test[11], for which a significant test result indicates poor calibration.

The slope and intercept of a calibration line are ideally 1 and 0, respectively (perfect calibration). When the slope and intercept are unequal to 1 and 0, the risk predictions of the model do not correspond well to the observed outcome frequencies in the validation sample (miscalibration). A slope substantially different from 1 (e.g. < 0.8 or > 1.2) indicates that the associations between the predictors and outcome (represented by the regression coefficients) in the validation sample are different from the associations estimated in the development sample. These differences cannot easily be accounted for and require more rigorous adjustments of the original model. An intercept substantially different from 0 commonly reflects a difference in outcome incidence between the validation and development sample, that can not be explained by different values of the predictors in the model[14]. To account for this difference in outcome incidence, the intercept of the original model can be adjusted. Such an adjustment of the intercept does not influence the discrimination of the model as the ordering of the predicted risks is not changed after adjustment of the intercept.

Discrimination refers to the ability of a prediction model to distinguish between patients with and without PONV. The discrimination of each model in the validation sample was studied with the area under the receiver operating characteristic curve (ROC area). The ROC area may theoretically range from 0.5 (discrimination no better than chance) to 1.0 (perfect discrimination)[15].

Of all predictor and outcome values, 8% were missing, mainly for the predictors 'history of PONV' and 'history of motion sickness'. Deleting subjects with missing values leads to bias and a loss of statistical power[16;17]. Therefore, missing values were imputed with imputation techniques (using the aregImpute function standard available in S-plus software version 6.2). Analyses were performed with S-PLUS 6.2 Professional Edition for Windows.

Table 1 Distribution of outcome and predictor values in the development samples of the three inpatient models (Koivuranta et al.[8], Apfel et al.[6] and Van den Bosch et al.[10]) and the outpatient model (Sinclair et al.[5]). Distributions in the outpatient validation sample are shown in the last column. Values are proportion, mean (standard deviation), median (interquartile range) or median [range].

	development samples				validation sample
	Koivuranta et al.[8] (n=1107)	**Apfel et al.[6] (n=1040)**	**Van den Bosch et al.[10] (n=1389)**	**Sinclair et al.[5] (n=17,638)**	**(n=623)**
outcome					
incidence of PONV	52[a]	44	48	9	25
predictors					
female gender	66	57	57	67	62
age (years)	46 [4 – 86]	49 (36 – 61)	44 (32 – 56)	47 (21)	42 (32 – 54)
smoking	22	27	39	-	31
history of PONV or motion sickness	43[b]	35	38	-	36
lower abdominal or middle ear surgery	-	-	13	-	3
duration of surgery (minutes)	72 (40 – 105)	116 (73 – 152)	120 (80 – 173)	55 (44)	37 (26 – 53)
duration of surgery > 60 minutes	-	-	90	-	16
inhalation anesthesia	74	100	50	57	96
use of nitrous oxide	-	-	50	-	12
use of postoperative opioids	71	46	63	-	22

[a] only the incidence of nausea was reported, which was 52%

[b] The paper presents the history of PONV (43%) and the history of motion sickness (24%) separately. We estimated the incidence of the combined predictor 'history of PONV or motion sickness' using the incidence of 'history of PONV'.

- not presented in the original publication

Results

Table 1 shows the outcome incidence and distribution of predictor values in the four development samples of the tested models and in the validation sample. In the outpatient validation sample (n=623), 25% reported PONV within 24 hours after surgery, compared to 52% (Koivuranta et al.[8]), 44% (Apfel et al.[6]) and 48% (Van den Bosch et al.[10]) in the inpatient development samples. The incidence was 9% in the outpatient development sample of Sinclair et al.[5].

Figure 1 shows the calibration plots of the four tested models. All inpatient models clearly overpredicted the risk of PONV in the outpatient validation sample (in figure 1a-c the predicted risks were substantially higher than the observed risks), whereas the outpatient model underpredicted the risk of PONV in the validation sample (in figure 1d the predicted risks were lower than the observed risks). This was confirmed by the negative intercepts for the inpatient models (-0.45, -0.72 and -1.46, respectively) and the positive intercept for the outpatient model (1.22). The calibration slopes of the models developed by Koivuranta et al. and Apfel et al. were substantially different from 1 (0.53 and 0.55, respectively). The calibration slope of the model developed by Van den Bosch et al. was relatively close to 1 (0.85). All three inpatient models showed statistically significant results for the Hosmer-Lemeshow test ($p < 0.001$). The discrimination was relatively low for all inpatient models with ROC areas of 0.62 (Koivuranta et al.) and 0.64 (Apfel et al. and Van den Bosch et al.). The predictive accuracy of Sinclair et al.'s outpatient model was poor, both in terms of calibration (slope: 0.32) and discrimination (ROC area: 0.60). The poor calibration was confirmed by a statistically significant Hosmer-Lemeshow test ($p < 0.001$).

The calibration slopes of the models developed by Koivuranta et al., Apfel et al. and Sinclair et al. showed too low values to apply the models in outpatients. The (relatively) high slope value (0.85) of Van den Bosch et al.'s inpatient model indicated that this model may be applied in outpatients, provided that the intercept of this model is adjusted to account for the lower PONV incidence in outpatients. The new intercept of the model equals the intercept of the original model (0.15) plus the estimated intercept of the calibration line with the slope fixed at 1 (-1.46), i.e. 0.15 - 1.46 = -1.31. The formula of the adjusted model is therefore:

Risk of PONV = 1 / (1 + exp (- (-1.31 + 0.46*female gender – 0.022*age – 0.63*current smoking + 0.76*history of PONV or motion sickness + 0.61*lower abdominal or middle ear surgery + 0.72*isoflurane anesthesia))). (1)

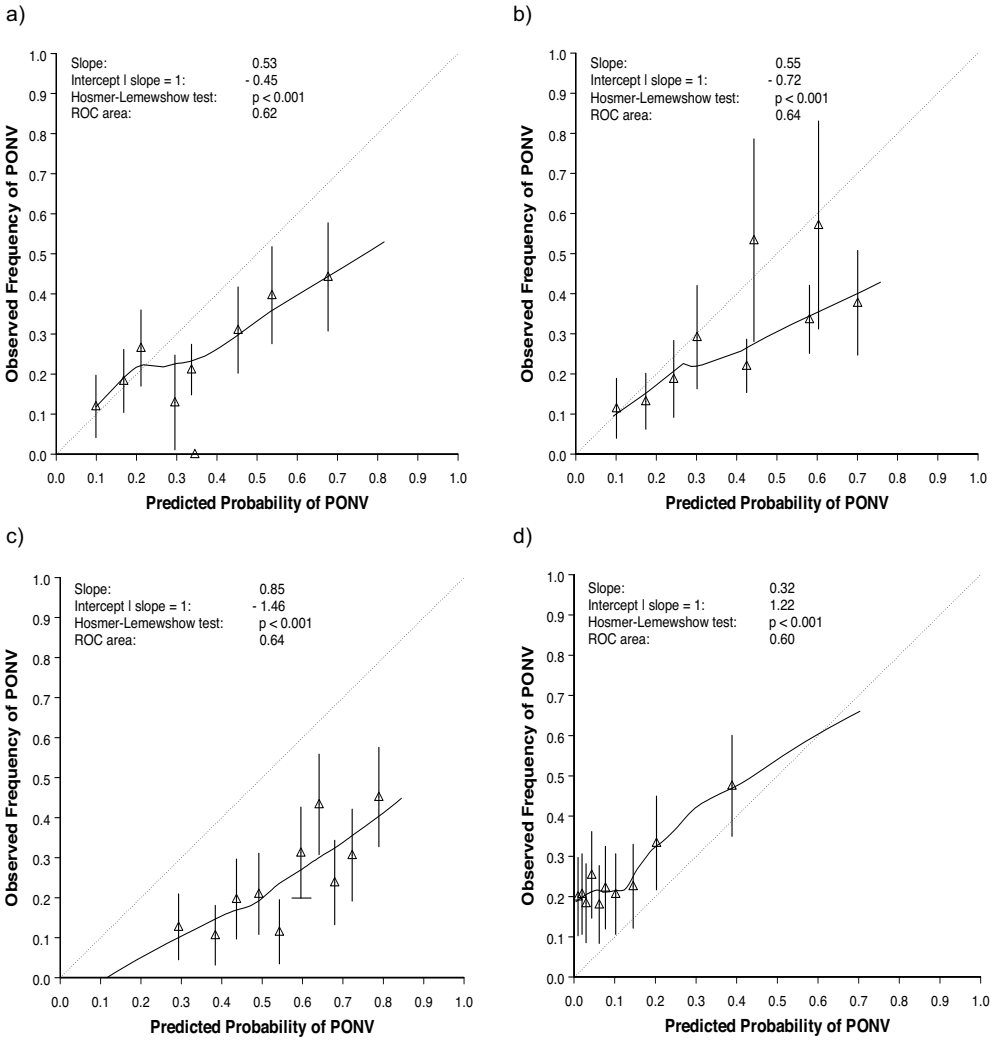

Figure 1 Calibration plots of the prediction models developed by Koivuranta et al.[8] (figure 1a), Apfel et al.[6] (figure 1b), Van den Bosch et al.[10] (figure 1c), and Sinclair et al.[5] (figure 1d), when validated in outpatients. The solid line shows the association between observed outcomes and predicted risks. Ideally, this line equals the dotted line that represents perfect calibration. Triangles indicate the observed frequency of PONV per decile of predicted risk. The vertical lines represent the corresponding 95% confidence intervals.

It should be noted that the regression coefficients of this adjusted model for outpatients are identical to the regression coefficients of the original model for inpatients. The calibration plot of the adjusted model is shown in figure 2a. The ROC area remained 0.64, as the discrimination of a model is insensitive to intercept adjustment.

For comparison, a model was developed in the outpatient validation sample, which included all predictors of the inpatient model developed by Van den Bosch et al. The regression coefficients in this outpatient model were estimated specifically for the outpatient population (table 2). The calibration plot of this outpatient model is shown in figure 2b. The ROC area was 0.66 with 95% confidence interval: 0.61 – 0.71 (after correction for optimism with bootstrapping techniques[18]).

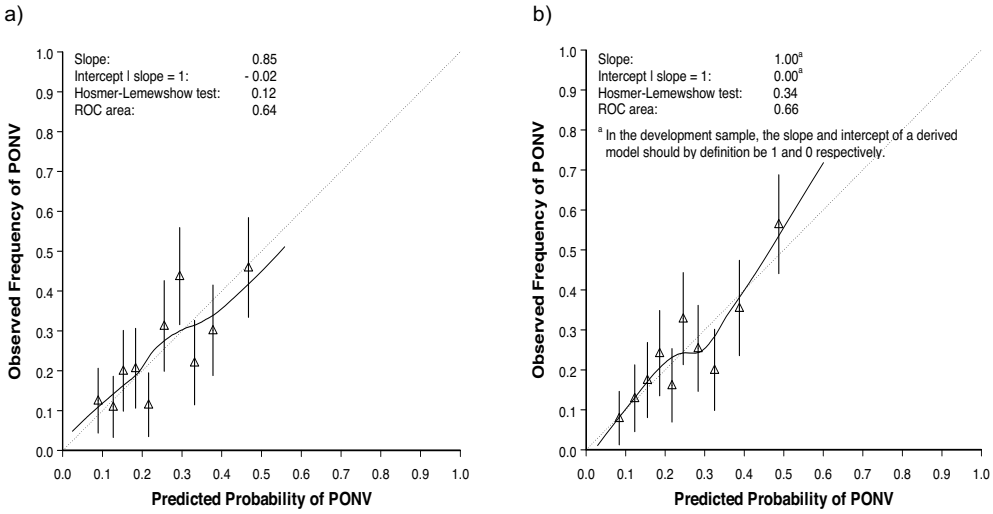

Figure 2 Calibration plots of the adjusted Van den Bosch model (figure 2a) and of the prediction model with the newly estimated regression coefficients for outpatients (figure 2b). The solid line shows the association between observed outcomes and predicted risks. Ideally, this line equals the dotted line that represents perfect calibration. Triangles indicate the observed frequency of PONV per decile of predicted risk. The vertical lines represent the corresponding 95% confidence intervals.

Table 2 Regression coefficients (standard errors) and odds ratios (95% confidence intervals [95% CI]) of the predictors in the Van den Bosch model[10], when estimated in the outpatient validation sample compared to the inpatient development sample.

	outpatient development sample		**inpatient validation sample**	
	regression coefficient (standard error)[a]	**odds ratio (95% CI)**	**regression coefficient (standard error)[a]**	**odds ratio (95% CI)**
female gender	0.45 (0.22)	1.57 (1.02 – 2.41)	0.46 (0.12)	1.58 (1.26 – 2.00)
age (per year)	- 0.023 (0.007)	0.98 (0.96 – 0.99)	- 0.022 (0.004)	0.98 (0.97 – 0.99)
current smoking	0.17 (0.21)	1.19 (0.79 – 1.79)	- 0.63 (0.12)	0.53 (0.42 – 0.67)
history of PONV or motion sickness	0.90 (0.21)	2.46 (1.63 – 3.71)	0.76 (0.12)	2.13 (1.69 – 2.70)
inhalation anesthesia[b]	0.78 (0.52)	2.18 (0.79 – 6.04)	0.72 (0.12)	2.05 (1.63 – 2.57)
lower abdominal or middle ear surgery	0.43 (0.52)	1.54 (0.55 – 4.26)	0.61 (0.18)	1.84 (1.30 – 2.60)
intercept	- 1.60 (0.62)	-[c]	0.15 (0.22)	-[c]

[a] after bootstrapping (i.e. adjusted for overfitting)

[b] in the outpatient validation sample, sevoflurane was mainly used as inhalation anesthetic, whereas in the development sample, inhalation anesthesia was always performed with isoflurane/nitrous oxide

[c] not relevant for the intercept

Discussion

This study tested the generalizability of three inpatient prediction models for PONV within 24 hours after surgery to an outpatient population[6;8;10]. Our hypothesis was that the inpatient models would have similar predictive accuracy in outpatients, under the provision that these models may need small adjustments to account for the lower absolute PONV incidence in outpatients. The predictive accuracy of the single prediction model developed in outpatients[5] was tested for comparison, but was surprisingly low (calibration slope: 0.32, ROC area: 0.60). Two of the inpatient models also showed poor predictive accuracy (calibration slopes much lower than 1: 0.53 and 0.55). More than small adjustments would be needed before these models are generalizable to outpatients. The third inpatient model, had a relatively high slope value (0.85). After intercept adjustment, this model yielded reasonably calibrated predictions of PONV risk in the outpatient validation sample. The discrimination of the adjusted model was moderate (ROC area = 0.64 compared to 0.68 in a validation sample of 1203 inpatients from the same hospital [data not shown]). For comparison, a model was developed in the outpatient validation sample, which included the predictors of the adjusted inpatient model and regression coefficients that were estimated for the outpatient population. This model performed only slightly better than the adjusted inpatient model (ROC area = 0.66).

In the outpatient validation sample, the slope value of the adjusted inpatient model was relatively close to 1 (0.85). This suggests that the predictor-outcome associations that were estimated in the inpatients were similar for the outpatients. As shown in table 2, this was indeed the case for the predictors ‘female gender’, ‘age’ and ‘history of PONV or motion sickness’. However, smoking behavior was not associated with PONV in the outpatient validation sample (table 2). It has been shown that cigarette smoke enhances the activity of liver enzymes that play an important role in the breakdown of a number of anesthetic agents. Consequently, smokers may have higher metabolic rates and thus a faster clearance of anesthetics[19], which likely influences patients’ liability to experience PONV, since volatile anesthetics are an important cause of PONV within the first two postoperative hours[20]. Hence, our observation that smoking is not associated with PONV in outpatients may be a chance finding. The odds ratios of the anesthetic technique and the type of surgery were comparable with the odds ratios estimated in the development sample, but the associations with PONV were not statistically significant. The 95% confidence intervals were large, which may be explained by the very high incidence of inhalation anesthesia (96%)

and very low incidence of lower abdominal or middle ear surgery (3%) in a relatively small sample size (n=623). Possibly, in a much larger outpatient sample with a larger number of patients receiving intravenous anesthesia (which implies a smaller number of patients receiving inhalation anesthesia) and a larger number of patients scheduled for lower abdominal or middle ear surgery, these predictors may show statistically significant associations with PONV.

In a previous paper[21], we suggested that the development of robust prediction models can be compared to cumulative meta-analyses for estimating the effect of a particular treatment. The process of model development starts with a development study. Subsequently, the model is tested in a validation sample and, if needed, adjusted using the data of the validation sample to improve the predictive accuracy of the model across various populations. Such a process will lead to robust predictions, because the model is based on large numbers of patients. Another advantage is that some development steps (e.g. variable selection) are not repeated, which reduces the risk of data-driven models that provide too extreme predictions for new patients. The intercept adjustment of the inpatient model should be seen in this perspective. The adjusted inpatient model was developed in a cumulative way, using data from both the development sample (for predictor selection and estimation of regression coefficients) and the validation sample (for estimating the adjustment of the intercept). Furthermore, in the outpatient validation sample, the odds ratios of the predictors in the adjusted inpatient model were similar to the odds ratios estimated in the inpatient development sample (table 2). Because of this and because of the larger size of the development sample (n=1389 versus n=623), the adjusted inpatient model may be preferred over the outpatient model with regression coefficients estimated specifically for outpatients, in spite of the higher ROC area of the latter model (0.66 versus 0.64). Of course, the robustness of the adjusted inpatient model should be tested in other outpatient samples, before this model is implemented in clinical practice.

Remarkably, the predictive accuracy of Sinclair et al.'s outpatient model[5] was poor in our outpatient validation sample. The high intercept value (1.22) of the calibration line (with slope fixed at 1) is probably related to the low PONV incidence (9%) in the development sample (table 1). The low incidence may be explained by the PONV definition used. PONV was defined as any volunteered report of nausea or observed active retching or vomiting requiring anti-emetics. This definition differs in two aspects from the PONV definition used in the validation sample. First, volunteered report may lead to underreporting of PONV, as episodes of mild nausea may not

be reported by patients. Also, the amount of observed active retching may be decreased at times when the workload of the nurses is high. Second, the definition was more strict as only treated retching and vomiting were recorded. Another explanation for the low PONV incidence is that only 57% of the patients in the development sample received general anesthesia (with volatile anesthetics). The other patients received monitored anesthesia care (36%), or regional or local anesthesia or a chronic pain block (7%). Those patients were approximately 11 times less likely to experience PONV than the patients who received general anesthesia[5].

The low value of the calibration slope (0.32) indicates that the estimated associations between the predictors and outcome (expressed as odds ratios [OR's]) were much stronger in the development sample than in the validation sample[22]. Indeed, the estimated OR's were more extreme (i.e. further from 1) in the development sample than in the validation sample for all predictors in the model, except for age (OR 0.87 versus 0.80, respectively) (data not shown). As a consequence, the predicted risks of PONV were too extreme in the validation sample. This means that high risk estimates were too high and low risk estimates were too low, which is reflected in the low slope value.

Although the discriminative ability of the outpatient model developed by Sinclair et al. was good in the development sample (ROC area = 0.79), it was low in the validation sample (ROC area = 0.60). Several explanations may be considered. First, the low incidence of PONV in the development sample (9%) makes it easier to discriminate between patients with and without PONV. Second, the development sample was more heterogeneous than the development sample with respect to the type of anesthesia patients received. In the validation sample, all patients (100%) received general anesthesia, whereas in the development sample only 57% received general anesthesia and 43% received anesthesia with other techniques. As these other techniques were associated with much lower PONV incidences, it was easier to distinguish between patients with and without PONV in the more heterogeneous development sample. Third, estimates of the associations between predictors and outcomes were too extreme (odds ratios too far from 1) in the development sample. As a result, the discriminative ability of the model was overestimated. The ROC area of the model would have been much lower after correction for overoptimism with for instance bootstrapping techniques.

In conclusion, although all models showed limited discriminative ability in outpatients, one model showed reasonably calibrated predictions of PONV risk in outpatients. This model may be generalized to outpatients (after a small adjustment).

References

1. Jenkins K, Grady D, Wong J, Correa R, Armanious S, Chung F. Post-operative recovery: day surgery patients' preferences. *British Journal of Anaesthesia* 2001;**86**:272-4.
2. Fortier J, Chung F, Su J. Unanticipated admission after ambulatory surgery - a prospective study. *Canadian Journal of Anesthesia* 1998;**45**:612-9.
3. Gold BS, Kitz DS, Lecky JH, Neuhaus JM. Unanticipated admission to the hospital following ambulatory surgery. *JAMA: the journal of the American Medical Association* 1989;**262**:3008-10.
4. Kovac AL. Prevention and treatment of postoperative nausea and vomiting. *Drugs* 2000;**59**:213-43.
5. Sinclair DR, Chung F, Mezei G. Can postoperative nausea and vomiting be predicted? *Anesthesiology* 1999;**91**:109-18.
6. Apfel C, Läärä E, Koivuranta M, Greim C-A, Roewer N. A simplified risk score for predicting postoperative nausea and vomiting: conclusions from cross-validations between two centers. *Anesthesiology* 1999;**91**:693-700.
7. Gan TJ. Postoperative nausea and vomiting - can it be eliminated? *JAMA: the journal of the American Medical Association* 2002;**287**:1233-6.
8. Koivuranta M, Läärä E, Snåre L, Alahuhta S. A survey of postoperative nausea and vomiting. *Anaesthesia* 1997;**52**:443-9.
9. Palazzo M, Evans R. Logistic regression analysis of fixed patient factors for postoperative sickness: a model for risk assessment. *British Journal of Anaesthesia* 1993;**70**:135-40.
10. Van den Bosch JE, Moons KG, Bonsel GJ, Kalkman CJ. Does measurement of preoperative anxiety have added value for predicting postoperative nausea and vomiting? *Anesthesia and Analgesia* 2005;**100**:1525-32.
11. Hosmer D, Lemeshow S. Applied logistic regression. New York: John Wiley & Sons, Inc., 1989.
12. Spiegelhalter DJ. Probabilistic prediction in patient management and clinical trials. *Statistics in Medicine* 1986;**5**:421-33.
13. Miller ME, Hui SL, Tierney W. Validation techniques for logistic regression models. *Statistics in Medicine* 1991;**10**:1213-26.
14. Vergouwe Y, Steyerberg EW, Foster RS, Habbema JD, Donohue JP. Validation of a prediction model and its predictors for the histology of residual masses in nonseminomatous testicular cancer. *Journal of Urology* 2001;**165**:84-8.
15. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;**143**:29-36.
16. Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. *American Journal of Epidemiology* 1995;**142**:1255-64.

17. Little RA. Regression with missing X's: a review. *Journal of the American Statistical Association* 1992;**87**:1227-37.
18. Efron B, Tibshirani RJ. An introduction to the bootstrap. New York: Chapman and Hall, 1993.
19. Sweeney BP. Why does smoking protect against PONV? *British Journal of Anaesthesia* 2002;**89**:810-3.
20. Apfel CC, Kranke P, Katz MH, Goepfert C, Papenfuss T, Rauch S et al. Volatile anaesthetics may be the main cause of early but not delayed postoperative vomiting: a randomized controlled trial of factorial design. *British Journal of Anaesthesia* 2002;**88**:659-68.
21. Van den Bosch JE, Kalkman CJ, Vergouwe Y, Van Klei WA, Bonsel GJ, Grobbee DE et al. Assessing the applicability of scoring systems for predicting postoperative nausea and vomiting. *Anaesthesia* 2005;**60**:323-31.
22. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* 1996;**15**:361-87.

Appendix

The formulas of the three tested inpatient prediction models for postoperative nausea and vomiting (PONV) within 24 hours after surgery are:

Koivuranta et al.[8]
Risk of PONV = 1 / (1 + exp (- (-2.21 + 0.93*female gender + 0.82*history of PONV + 0.59* history of motion sickness + 0.61*non-smoking + 0.75*duration of surgery over 60 minutes))) (1)

Apfel et al.[6]
Risk of PONV = 1 / (1 + exp (- (-2.28 + 1.27*female gender + 0.65*history of PONV or motion sickness + 0.72* non-smoking + 0.78*postoperative opioid use))) (2)

Van den Bosch et al.[10]
Risk of PONV = 1 / (1 + exp (- (0.15 + 0.46*female gender – 0.022*age – 0.63*current smoking + 0.76*history of PONV or motion sickness + 0.61*lower abdominal or middle ear surgery + 0.72*isoflurane anesthesia))) (3)

The formula of the tested outpatient prediction model for PONV occurrence within 24 hours after surgery is:

Sinclair et al.[5]
Risk of PONV = 1 / (1 + exp (- (-5.97 – 0.14*age – 1.03*male gender – 0.42*smoking + 1.14*history of PONV + 0.46*duration of surgery in 30 minutes increments + 2.36* general anesthesia + 1.48*ear-nose-throat-surgery + 1.77*opthalmologic surgery + 1.90*plastic surgery + 1.20*gynecologic surgery other than dilation & curretage + 1.04*orthopedic procedure involving the knee + 1.78*orthopedic procedure involving the shoulder + 0.94*orthopedic procedure involving neither knee nor shoulder))) (4)

All four models are logistic regression models. Logistic regression modeling is used to model the relation between predictor variables and binary outcomes (e.g. PONV occurrence yes/no). The general formula of such models is:

$$\text{risk of binary outcome} = 1 / (1 + e^{-\text{linear predictor}})$$

The linear predictor is the sum of the intercept and the regression coefficients (weights) multiplied by the predictor values observed for each patient. The linear predictor of a model can be written as:

$$\text{linear predictor} = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \ldots + \beta_z * X_z$$

where β_0 represents the intercept, X_1 to X_z represent the predictor values, and β_1 to β_z represent the regression coefficients of each corresponding predictor.

Chapter 7

Boundaries to the generalizability of clinical prediction models

JE van den Bosch, Y Vergouwe, WA van Klei, CJ Kalkman, KGM Moons

(submitted)

Abstract

Clinical prediction models are developed to assist diagnostic and prognostic assessments of physicians. These models predict the presence (diagnosis) or future occurrence (prognosis) of outcomes in relation to patient and disease characteristics, and test results. Ideally, a developed prediction model is generalizable to a wide variety of patients. In reality, the predictive accuracy of a model is often decreased when applied in new patients. Therefore, a newly developed prediction model should be tested in various patient samples (validation samples), before implementation in clinical practice. Validation samples are preferably different from but related to the development sample. Criteria for a ´different but related' sample are, however, not clearly defined. We explore this concept, illustrated with examples of prediction models that are developed and tested in different care settings (e.g. primary versus secondary care) or different age categories (e.g. adult versus pediatric patients). We discuss whether the generalizability of prediction models can be improved by using more heterogeneous development samples.

Introduction

Clinical prediction models are developed to assist diagnostic and prognostic assessments of physicians. These models predict the presence (diagnosis) or future occurrence (prognosis) of outcomes in relation to patient and disease characteristics and test results. For example, recently a prediction model was developed to diagnose deep venous thrombosis (DVT) in primary care patients who present with symptoms and signs suggestive of DVT[1]. This model predicts the probability that a patient has DVT based on gender, vein distension, the results of a rapid D-dimer test and several other predictors (table 1). Other well-known examples of prediction models are the Ottawa ankle rule to diagnose fractures in patients with ankle injuries (diagnostic model)[2], the Framingham risk score to predict the risk of cardiovascular events in the general population[3] and the APACHE score to predict the risk of hospital mortality in critically ill hospitalized adults (prognostic models)[4].

Table 1 Example of a prediction model to predict the risk of deep venous thrombosis (DVT) in primary care patients suspected of DVT[1]. This prediction model was simplified to a so called risk score with an accompanying score chart. To estimate the risk of DVT in an individual patient, a total score is calculated, based on the scores that correspond to the patient's predictor values (upper part of the table). The risk of DVT that corresponds to the total score can be read from the table (lower part of the table).

predictors	**score**
male gender	1
use of oral contraceptives	1
presence of malignancy	1
recent surgery	1
absence of leg trauma	1
vein distension	1
calf difference ≥ 3 cm	2
abnormal D-dimer test	6
total score	**risk of DVT**
0 – 3 (very low risk)	0.7 %
4 – 6 (low risk)	4.5 %
7 – 9 (moderate risk)	21.7 %
10 – 13 (high risk)	51.3 %

A prediction model yields its highest predictive accuracy in the patient sample from which it has been developed. Hence, it may sound appealing to develop a separate prediction model for each particular patient population to obtain optimal predictive accuracy for that specific population. However, this strategy is not feasible because time and resources are limited. Therefore, diagnostic and prognostic researchers aim to develop prediction models that are generalizable to a wide variety of patient populations and clinical settings. Unfortunately, a prediction model is usually overfitted on the development sample. As a result, the predictive accuracy is often decreased when the model is applied in new patient samples. Therefore, the predictive accuracy of any developed prediction model should be tested in various patient samples (so called external validation studies), before the model is widely implemented in clinical practice.

The predictive accuracy of a prediction model can be evaluated in terms of calibration and discrimination. Calibration refers to the agreement between predicted risks and observed outcome frequencies. Discrimination refers to the ability to distinguish between patients with and without the outcome. The discriminative ability of a model with a dichotomous outcome (e.g. presence of DVT [yes/no]) is commonly studied with the area under the receiver operating characteristic curve (ROC). This parameter theoretically ranges from 0.5 (discrimination no better than chance) to 1.0 (perfect discrimination)[5].

It is suggested that external validation studies should be performed with patient samples that are different from but related to the patient sample from which the model was developed (development sample)[6;7]. Criteria for a ´different but related' sample are, however, not clearly defined.

In this paper, we examine the concept of a ´different but related´ validation sample. Next, we explore whether two common scenarios can be considered as validations of developed prediction models in ´different but related´ samples. These scenarios are: 1) testing a model in a validation sample from a different care setting (e.g. primary versus secondary care) and 2) testing a model in a validation sample with a different age range (e.g. adults versus children). Finally, we discuss whether the generalizability of prediction models can be improved by using more heterogeneous development samples.

When are populations different but related?

To assess whether a validation sample and a development sample are different but related, the characteristics of the samples should be compared.

Validation samples that contain new patients are automatically different from the development sample. Relatedness, however, is more difficult to define. The first and perhaps most important criterion for relatedness in case of a diagnostic model is that the validation sample should contain patients suspected of the same disease (e.g. patients who present with symptoms and signs suggestive of DVT) as the patients in the development sample. In case of a prognostic model, the validation sample should contain patients with the same disease (e.g. colon cancer). Second, the definitions of the outcome and the predictors in the model should be similar. If, for instance, abnormal D-dimer concentrations are defined as higher than 500 ng/ml in the development sample and as higher than 1,000 ng/ml in the validation sample, the samples are less related. Third, the ranges of predictor values in the validation sample should fall within the ranges of predictor values in the development sample.

The third criterion is illustrated by an example of a hypothetical prediction model with a development sample and three different validation samples (figure 1). This hypothetical model has a dichotomous outcome (relapse within one year after treatment of disease) and two predictors (the categorical predictor disease severity [mild/moderate/severe] and the continuous predictor age [18-85 years]), which both positively related to the outcome. It is assumed that the definition of the outcome and predictors as well as the associations between the predictors and the outcome are similar for the development and validation samples. In the first validation sample (figure 1b), the range and distributions of the predictor values are similar to those in the development sample (figure 1a). Consequently, the outcome frequencies are comparable. In the second validation sample (figure 1c), the range of values for disease severity is similar, but the distribution within this range is different (more severe patients are included). As a result, the outcome frequency also differs (more relapses), since the occurrence of relapse is positively related to the severity of disease. In the third validation sample (figure 1d) only older patients are included, which narrows the age range. Here again, the outcome frequency is increased, as the occurrence of relapse is also positively related to age. In all three validation samples, the ranges of the predictor values fall within the ranges of the predictor values in the development sample, such that all samples can be considered as validation samples that are different from but related to the development sample.

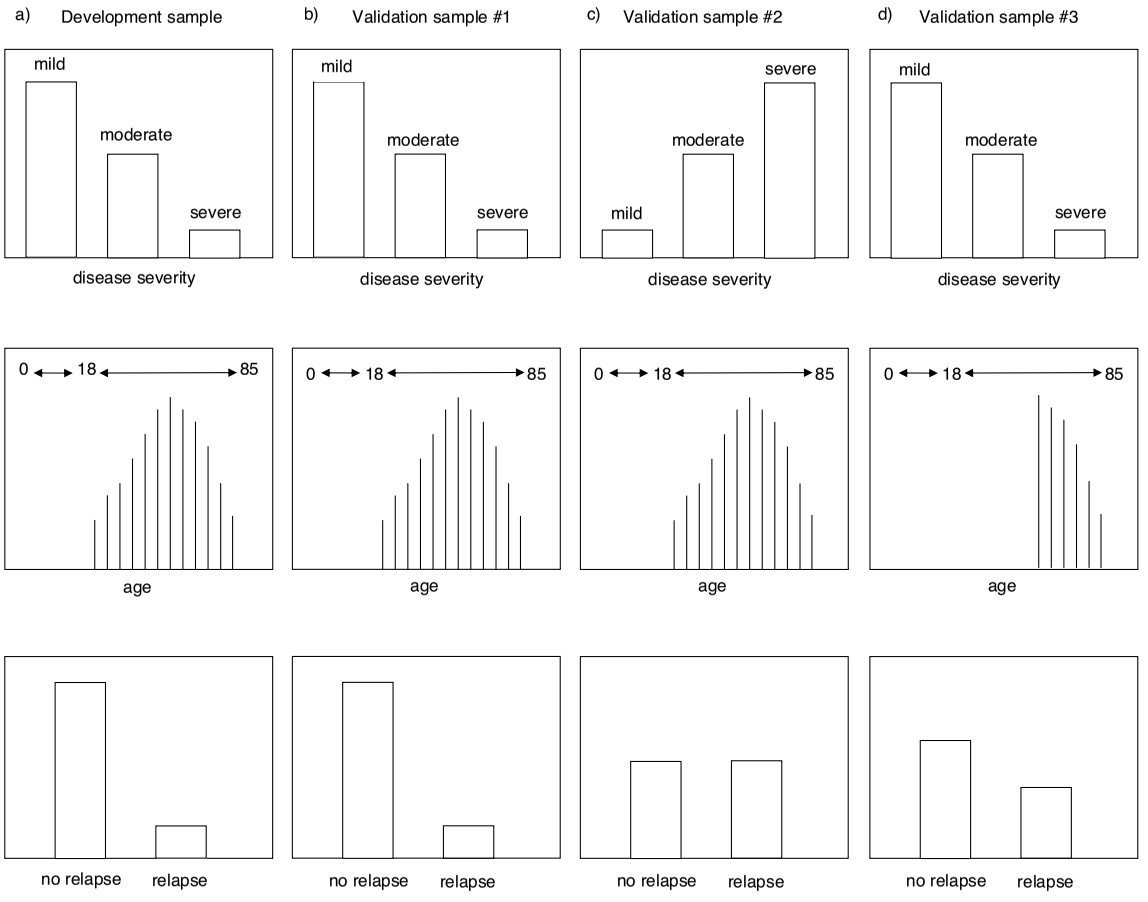

Figure 1 Example of a hypothetical prediction model with a development sample and three different validation samples. This hypothetical model has a dichotomous outcome (e.g. relapse within one year after treatment of colon cancer), and two predictors (e.g. the categorical predictor disease severity [mild/moderate/ severe] and the continuous predictor age [18-85 years]). It is assumed that the definition of the outcome and predictors as well as the associations between the predictors and the outcome are similar for the development and validation samples. Figure 1a shows the ranges and distributions of the two predictors (first two rows) and the outcome (last row) in the development sample. Figure 1b shows a validation sample in which both the ranges and distributions of the predictor values are comparable to those in the development sample. Figure 1c shows a validation sample in which the range of values for disease severity is comparable, but the distribution of disease severity within this range is different. Figure 1d shows a validation sample in which the age range is narrowed.

Although this hypothetical example sheds some light on the concept of a ´different but related´ sample, it is not yet clear whether samples with different ranges of predictor values (e.g. different age ranges such as adults versus children) or from different care settings (e.g. primary care versus secondary care) can be considered as related samples. We will explore these issues in the next section.

Two common scenarios in diagnostic and prognostic research

Generalizability of a prediction model to a different care setting

The majority of the clinical prediction models are developed in secondary care settings. When tested in primary care settings, the predictive accuracy is usually decreased[7]. For example, the well-known Wells rule to diagnose DVT[8] was developed in secondary care patients suspected of DVT (ROC area: 0.64). When this rule was tested in 1295 primary care patients, the ROC area was only 0.59[9]. A second model, developed in the same primary care population, showed a significantly higher accuracy (ROC area: 0.78 in the development sample[1] and 0.75 when validated in an other sample of primary care patients[10]). When, in turn, this primary care model was externally validated in a secondary care sample, the discriminative ability of the model was surprisingly high (ROC area: 0.87)[11]. Hence, the DVT model developed in a secondary care setting showed poor accuracy in primary care, whereas the DVT model developed in primary care showed good accuracy when tested in a secondary care setting.

These findings may be explained by the fact that secondary care patients are selected by referral from the primary care population. As a result, the disease spectrum presented to the medical specialist is narrower and thus more homogeneous than the spectrum seen by the primary care physician. Also, disorders presented to the medical specialist are often in a later, more developed stage and thus (severe) disease is more prevalent in a secondary care population than in a primary care population[12]. Secondary care populations are not only different from but also related to primary care populations. After all, the ranges of predictor values in a secondary care population fall within the ranges of the primary care population from which it was selected (by referral). The reverse, however, is not the case. Therefore, a primary care population is not necessarily related to a secondary care population. Consequently, prediction models developed in a primary care setting can often easier be generalized to a secondary care setting, than vice versa.

Generalizability of a prediction model to a different age range

In various situations, prediction models (for the same outcome) have been developed separately for adults and children. For example, the APACHE score was developed to predict mortality after intensive care unit (ICU) admission in adults[4]. Later, the PRISM score was developed specifically for children admitted to the ICU[13]. Similarly, the MELD score was developed to predict short-term mortality in adult patients with end-stage liver disease[14], which was followed by a similar prediction model (PELD) for children[15]. The question arises whether separate models for pediatric populations are always needed or whether it is possible to generalize prediction models developed in adult patients to pediatric patients.

In theory, it should be possible to generalize a prediction model developed in adult patients to pediatric patients. After all, the efficacy of medical treatments is often tested in clinical trials including adults only. If the treatment has proven to be efficacious and introduced in clinical practice, the drug is subsequently applied in children after dose adjustment for body mass. Before prediction models can be transported from adult to pediatric patients, however, two important issues need to be considered. First, if age is a predictor in the model, the predictive effect of age is extrapolated outside the range of values in which it was estimated. Second, a validation sample with patients from a different age range may differ from the development sample in other ways that limit the use of the model. One possible difference is that a predictor defined for adults may be difficult to assess in children. For example, the level of preoperative pain was a predictor in a prediction model for severe postoperative pain in adults[16]. The level of preoperative pain was assessed with a verbal rating scale, which is less suitable for pain assessment in young children. Another possible difference is that predictors defined for adults may contain less information when collected in children. For example, the history of nausea and vomiting after previous anesthesia is a known predictor of postoperative nausea and vomiting (PONV) in adults[17-19]. Obviously, fewer children have had previous anesthesia compared to adults. Consequently, the estimated association between 'history of nausea and vomiting after previous anesthesia' and PONV may be different in adults and children. Indeed, when available prediction models for PONV, developed in adults[17;18;20;21], were tested in a pediatric surgical population, their predictive accuracy was substantially decreased[22]. The authors therefore concluded that a specialized model for children was required, which they subsequently developed[23].

In summary, samples with adult patients may differ from samples with children. Not only the range of values for the predictor age will differ, but also the associations of other predictors with the outcome may be different. These samples are therefore less related.

Can the generalizability of a prediction model be improved?

From the two scenarios described above it becomes clear that validation samples are related to a development sample if the ranges of predictor values in the validation sample fall within the ranges of the predictor values in the development sample. In that case, a model will likely have similar predictive accuracy in the related validation sample. Samples are less related, if the ranges of predictor values in the validation sample are wider than the range of predictor values in the development sample (e.g. generalizing a prediction model from a secondary care to a primary care setting), or if the ranges do not overlap at all (e.g. generalizing a prediction model across age ranges). In such cases, the predictive accuracy of a model may decrease substantially in the validation sample and separate prediction models may need to be developed for each patient population.

It seems that prediction models generalize well to samples with ranges of predictor values that fall within the ranges of predictor values in the development sample. This raises the question whether the use of more heterogeneous development samples, including wider ranges of predictor values (e.g. a broader spectrum of ages or patients from various clinical settings), might improve the generalizability of prediction models. Although no publications are yet available to demonstrate that such a strategy enhances the generalizability of a developed model, data are available on the generalizability across surgical settings. Based on our own findings, we briefly illustrate in the next section how a prediction model developed from a more heterogeneous development sample can be a promising alternative to developing separate models.

Improving the generalizability of a prediction model by using a more heterogeneous development sample: an example

A prognostic model was developed for preoperative prediction of severe postoperative pain in surgical inpatients (data collected in a hospital in The Netherlands)[16]. The discriminative ability of this model was fairly good in the development sample (ROC area: 0.73). When the model was applied in surgical outpatients (data collected in the same hospital and time period),

however, the ROC area was only 0.59 and the predicted probabilities were systematically too high[24]. The investigators decided to combine the samples of in- and outpatients to a more heterogeneous development sample, and to develop a general model for pain prediction in both in- and outpatients. This model included the same predictors as the original model for inpatients, with one additional variable for surgical setting (inpatients versus outpatients). As the associations of some predictors differed for in- and outpatients, the adapted model also included interaction terms to represent these different associations. The discriminative value of this adapted model for outpatients (ROC area: 0.69) had much improved compared to the original model (ROC area increased with 17%). When this model was subsequently tested in a new sample of in- and outpatients (data collected in another Dutch hospital in a later time period), the ROC area was 0.65. Hence, the predictive accuracy of a model developed from a more heterogeneous sample was much higher than the accuracy of a model generalized across surgical settings. This example suggests that, although one may always develop a separate model for patient populations that were not represented in the development sample, developing prediction models in more heterogeneous samples may be a promising alternative to developing separate models.

Discussion

Different but related validation samples are generally recommended for testing the generalizability of prediction models[6;7]. As criteria for a different but related validation sample were not yet clearly defined, this concept was explored, illustrated with a validation example of a hypothetical prediction model and examples of prediction models that are developed and validated in different care settings (e.g. primary versus secondary care) or different age categories (e.g. adult versus pediatric patients). The latter two examples show that the performance of prediction models decreases when validation samples are less related to the development sample, i.e. if the ranges of predictor values in the validation sample are wider than the ranges of predictor values in the development sample (e.g. generalizing a prediction model from secondary care to primary care), or if the ranges do not overlap at all (e.g. generalizing a prediction model across age ranges). It seems that prediction models generalize well to samples that are closely related, i.e. samples with ranges of predictor values that fall within the ranges of predictor values in the development sample.

We hypothesized that the generalizability of prediction models may be improved by using more heterogeneous development samples, including

wider ranges of predictor values, e.g. patients from various clinical settings and age categories. Clinical setting may then be included as a separate predictor in the model (as in the surgical setting example) and age can be modeled over a wider range of values. This strategy has several limitations, however. It requires that patients from several settings and age groups are represented in the development sample in sufficient numbers. Furthermore, modeling of interaction terms may be needed to account for possibly different associations between predictors and outcome in different settings or age ranges. Consequently, as more predictors are included in the model, larger sample sizes are required. Finally, one should be aware that a model developed in a heterogeneous sample will often yield lower discriminative ability when tested or applied in a more homogeneous sample, as it is more difficult to discriminate between patients in a more homogeneous sample.

More heterogeneous development samples may also facilitate the generalizability of a prediction model across hospitals and countries. It is less appealing to model the effect of hospital or country. Nevertheless, a model based on patients from several hospitals and countries is more likely to generalize well to other centers and countries than a model developed in only one center and country.

Concluding remarks

We hypothesize that the use of more heterogeneous development samples may improve the generalizability of prediction models to various samples. The development sample should include a wide range of predictor values and, if applicable, several settings, hospitals or countries.

References

1. Oudega R, Moons KGM, Hoes AW. Ruling out deep venous thrombosis in primary care. A simple diagnostic algorithm including D-dimer testing. *Thrombosis and Haemostasis* 2005;**94**:200-5.
2. Stiell IG, Greenberg GH, McKnight RD, Nair RC, McDowell I, Worthington JR. A study to develop clinical decision rules for the use of radiography in acute ankle injuries. *Annals of Emergency Medicine* 1992;**21**:384-90.
3. Kannel WB, McGee D, Gordon T. A general cardiovascular risk profile. *The American Journal of Cardiology* 1976;**38**:46-51.
4. Knaus WA, Wagner DP, Draper EA, Zimmerman JE, Bergner M, Bastos PG et al. The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. *Chest* 1991;**100**:1619-36.
5. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;**143**:29-36.
6. Altman DG, Royston P. What do we mean by validating a prognostic model? *Statistics in Medicine* 2000;**19**:453-73.
7. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Annals of Internal Medicine* 1999;**130**:515-24.
8. Wells PS, Anderson DR, Bormanis J, Guy F, Mitchell M, Gray L et al. Value of assessment of pretest probability of deep-vein thrombosis in clinical management. *Lancet* 1997;**350**:1795-8.
9. Oudega R, Hoes AW, Moons KGM. The Wells rule does not adequately rule out deep venous thrombosis in primary care patients. *Annals of Internal Medicine* 2005;**143**: 100-7.
10. Toll DB, Oudega R, Bulten RJ, Hoes AW, Moons KGM. Validation of a diagnostic rule for deep venous thrombosis in primary care. *submitted* 2006.
11. Oudega R, Hoes AW, Moons KGM. Excluding deep venous thrombosis in secondary care using a decision tool derived in primary care. *submitted* 2006.
12. Knottnerus JA. Between iatrotropic stimulus and interiatric referral: the domain of primary care research. *Journal of Clinical Epidemiology* 2002;**55**:1201-6.
13. Pollack MM, Ruttimann UE, Getson PR. Pediatric risk of mortality (PRISM) score. *Critical Care Medicine* 1988;**16**:1110-6.
14. Malinchoc M, Kamath PS, Gordon FD, Peine CJ, Rank J, ter Borg PC. A model to predict poor survival in patients undergoing transjugular intrahepatic portosystemic shunts. *Hepatology* 2000;**31**:864-71.
15. McDiarmid SV, Anand R, Lindblad AS, principal investigators and institutions of the studies pediatric liver transplantation (SPLIT) research group. Development of a pediatric end-stage liver disease score to predict poor outcome in children awaiting liver transplantation. *Transplantation* 2002;**74**:173-81.

16. Kalkman CJ, Visser K, Moen J, Bonsel GJ, Grobbee DE, Moons KG. Preoperative prediction of severe postoperative pain. *PAIN* 2003;**105**:415-23.
17. Apfel CC, Läärä E, Koivuranta M, Greim C-A, Roewer N. A simplified risk score for predicting postoperative nausea and vomiting. Conclusions from cross-validations between two centers. *Anesthesiology* 1999;**91**:693-700.
18. Koivuranta M, Läärä E, Snåre L, Alahuhta S. A survey of postoperative nausea and vomiting. *Anaesthesia* 1997;**52**:443-9.
19. Van den Bosch JE, Moons KG, Bonsel GJ, Kalkman CJ. Does measurement of preoperative anxiety have added value for predicting postoperative nausea and vomiting? *Anesthesia and Analgesia* 2005;**100**:1525-32.
20. Palazzo M, Evans R. Logistic regression analysis of fixed patient factors for postoperative sickness: a model for risk assessment. *British Journal of Anaesthesia* 1993;**70**:135-40.
21. Sinclair DR, Chung F, Mezei G. Can postoperative nausea and vomiting be predicted? *Anesthesiology* 1999;**91**:109-18.
22. Eberhart LHJ, Morin AM, Guber D, Kretz FJ, Schäuffelen A, Treiber H et al. Applicability of risk scores for postoperative nausea and vomiting in adults to paediatric patients. *British Journal of Anaesthesia* 2004;**93**:386-92.
23. Eberhart LHJ, Geldner G, Kranke P, Morin AM, Schäuffelen A, Treibner H et al. The development and validation of a risk score to predict the probability of postoperative nausea and vomiting in pediatric patients. *Anesthesia and Analgesia* 2004;**99**:1630-7.
24. Janssen KJM, Kalkman CJ, Grobbee DE, Bonsel GJ, Moons KGM, Vergouwe Y. Effectively predicting severe acute postoperative pain in outpatients. *submitted* 2006.

Chapter 8

General Discussion

Risk-based anti-emetic prophylaxis may reduce the incidence of postoperative nausea and vomiting (PONV). In the past, patients at high risk of PONV were classified by reference to a single factor, such as previous PONV occurrence or the type of surgery. Nowadays, several multivariable prediction models for PONV are available[1-4]. These models yield more objective risk estimations and predict a patient's risk of PONV more accurately[5]. The predictive accuracy and generalizability of these models is, however, still limited. The research described in this thesis was conducted to test the generalizability of these models across populations, including outpatient populations, and to improve the predictive accuracy of currently available inpatient prediction models for PONV.

This chapter discusses several methodological issues related to prediction of PONV, explores the clinical implications of the research described in this thesis and outlines directions for future PONV prediction research.

Methodological issues

Calibration versus discrimination

Both calibration and discrimination are important aspects of a model's predictive accuracy. These aspects should therefore be analyzed concurrently. However, in specific situations one aspect may be more important than the other depending on the intended application of the prediction model[6]. When risk predictions are used by clinicians to make informed decisions, model calibration is particularly important. In such cases, clinicians are not concerned about how high a patient's risk is compared to the risk of other patients. Instead, they are concerned with the absolute risk that their patient will suffer the predicted outcome (e.g. PONV). In contrast, if risk predictions are used for risk stratification of patients in clinical trials, patients should be correctly ranked in order of risk, and thus discrimination is more important[6].

Model calibration can be influenced by the outcome incidence of the population in which the model is studied. If a model is miscalibrated in a new population due to a different outcome incidence, the model can often easily be recalibrated for the new population by adapting the model's intercept[7]. Adaptation of the intercept does not affect the discrimination of the model, because the ordering of the predicted risks remains the same after intercept adjustment.

Model discrimination is influenced by the heterogeneity (distribution of predictor values) of the population in which the model is applied. In

populations that are more homogeneous (e.g. a population in a gynecological setting that includes only female patients) it is more difficult for a model to discriminate between patients. This does not necessarily mean that a model is useless in populations that are more homogeneous than the population in which the model was developed. The model may be well calibrated and correctly indicate that the patients in the 'new' population generally have a high risk of PONV, which could justify the administration of prophylactic anti-emetics.

In contrast to calibration, the discrimination of a model can not easily be improved. This may be a reason to prefer a miscalibrated model that discriminates well over a perfectly calibrated model with only modest discrimination. A miscalibrated model may be adjusted to improve calibration as described above, but poor discrimination can not easily be improved.

Improving prediction models for PONV

The discrimination of a model may increase when more predictors are included in the model. However, the research described in this thesis showed that the discrimination of a model with established predictors of PONV did not improve when preoperative anxiety was included in the model (chapter 4). Inclusion of surgical type and patient age in a newly developed model did improve discrimination, but did so in the development sample only. In the validation sample, the discrimination was comparable to that of the risk scores developed by Apfel et al. and Koivuranta et al., both of which do not include age or type of surgery (chapter 5).

Apfel et al. have suggested that the discriminative ability of inpatient prediction models for PONV will remain imperfect (ROC areas at best in the range of 0.7 to 0.8), unless new predictors with a strong impact are discovered[8]. This 'ceiling effect' is well-known in prognostic prediction research, although in diagnostic prediction research higher ROC areas are not uncommon[9]. The time lag between risk predictions and outcome measurements in prognostic prediction research is often given as an explanation for this phenomenon. In the case of PONV, however, the time lag is very short (only 24 hours) and another explanation for the limited discriminative ability of PONV prediction models should be searched for.

The data that are used for estimating a patient's risk of PONV rely partly on subjective information obtained from the patient. Patients' reporting on outcomes and predictors may be selective depending on what a patient considers relevant. For example, nausea is a highly subjective experience which makes it vulnerable to selective reporting. Furthermore, motion

sickness is an unclearly defined predictor. Patients who get sick in the car when reading may label this as motion sickness, or consider it insignificant. Similarly, patients who suffer motion sickness in the back seat but are free of complaints in the front seat (and therefore always sit in the front seat or drive the car by themselves) may either label this as motion sickness or consider it insignificant. Reporting may also be selective due to social pressure or incomplete memory. For example, smokers generally underreport their smoking behavior. Furthermore, adults with a history of PONV in early childhood may not be informed about their medical history and elderly patients with a long history of surgical procedures may have difficulties in recalling mild episodes of nausea. The resulting misclassification may have a negative influence on the discriminative ability of currently available prediction models for PONV. Thus, to improve the discriminative ability of these models, new potential predictors should not only have a strong impact but should preferably be measurable in a more objective way.

A possible new predictor may be a genetic factor that influences the activity of liver enzymes. It is known that genetically determined differences in the activity of liver enzymes may cause individual differences in the metabolic breakdown of anesthetic agents[10]. This may influence the liability of patients to experience PONV, since volatile anesthetics are an important cause of PONV within the first two hours after surgery[11]. So far, it has not been possible to model this genetic factor. Whether knowledge on the genetic risk factors of PONV may be obtained and used for PONV prediction in future remains to be defined.

PONV prediction in outpatients

There are several possible strategies for PONV prediction in outpatients. Available inpatient models may be adjusted for application in outpatients (chapter 6). Alternatively, a separate model may be developed for outpatients or one general model may be developed for both in- and outpatients.

The development of prediction models is a cumulative process, in which the predictive accuracy of a newly developed model is tested in various validation samples and, if needed, adjusted to improve the accuracy of the model across various populations. Such a process will lead to robust predictions, because the model is based on large numbers of patients. Another advantage is that some development steps (e.g. variable selection) are not repeated, which reduces the risk of data-driven models that provide too extreme risk predictions for new patients. The proposed adjusted inpatient model for outpatients (described in chapter 6) was developed in a cumulative

way, using data from both the development sample (for predictor selection and estimation of regression coefficients) and the validation sample (for estimating the adjustment of the intercept). Furthermore, in outpatients the odds ratios of the predictors in the adjusted inpatient model were similar to the odds ratios in inpatients (chapter 6). Because of this, and because of the larger size of the development sample, the adjusted inpatient model may be preferred over the newly developed outpatient model. Of course, the robustness of this adjusted inpatient model should be tested in other outpatient samples, before this model is implemented in clinical practice.

One may also consider to combine available datasets of in- and outpatients. Combining data sets yields larger and more heterogeneous development samples. Larger datasets may enable more accurate estimations of the regression coefficients of predictors. More heterogeneous datasets may yield prediction models that can be generalized across a wider variety of populations and settings. We described how the generalizability of an inpatient prediction model for postoperative pain to outpatients improved when the model was developed in a more heterogeneous sample of in- and outpatients and the surgical setting (inpatients versus outpatients) was included as a predictor (chapter 7). These results suggest that it may be a viable strategy to develop one general prediction model for both in- and outpatients.

Clinical implications

Protocol for PONV prophylaxis

Implementation of a protocol for PONV prophylaxis requires that clinically relevant risk thresholds for PONV prophylaxis are established. Unfortunately, there is yet no consensus about the ideal thresholds for anti-emetic prophylaxis, i.e. for which thresholds the benefits of PONV prophylaxis would outweigh the costs and possible side-effects. There are arguments for a high risk threshold for PONV prophylaxis. With a high risk threshold, the specificity of the prophylactic policy increases (fewer patients will be treated unnecessarily) and the incidence of adverse effects decreases. But of course, this choice will always be made at the cost of sensitivity (PONV will be prevented in fewer patients). What would be the most ethical and economically responsible decision is still open for discussion. We suggest that anti-emetic strategies should be reserved for patients at high risk of PONV, and should be applied selectively to guard the cost-effectiveness of a prophylactic policy.

It is known that the anti-emetics ondansetron, dexamethasone, and droperidol are similarly effective in preventing PONV[12]. Each anti-emetic reduces the risk of PONV by about 26% (relative risk reduction)[12]. Propofol reduces the risk of PONV by 19% and substitution of nitrogen for nitrous oxide reduces the risk by 12% (also relative risk reductions)[12]. The relative risk reductions provided by these interventions are independent of the patients' baseline risk[12]. The absolute risk reductions, however, depend critically on the patients' baseline risk[12]. The absolute risk reduction provided by a second and third intervention will therefore be less than the absolute risk reduction provided by the initial intervention. Consider, for example, a patient whose baseline risk of PONV is estimated to be 72%. Administration of a single anti-emetic would reduce the patient's risk of PONV by 26% and the patient's risk would thus become 72%*(1-0.26) = 53% (absolute risk reduction of 19%). The risk will become 53%*[1-0.26] = 39% if the patient receives a combination of two anti-emetics (absolute risk reduction of 14%). Administration of three anti-emetics would reduce the risk to 39%*(1-0.26) = 29% (absolute risk reduction of 10%). Hence, as each anti-emetic yields the same relative risk reduction, the safest (or cheapest) anti-emetic should be used as an initial intervention. Additional interventions will further reduce the patient's risk, but to a lesser extent than the initial intervention will do (table 1), whereas they increase costs and the likelihood of side-effects. Multiple interventions should therefore be reserved for patients at high risk of PONV and for patients in whom even a single episode of postoperative emesis would pose a significant hazard of surgical complications (e.g. wiring of jaw).

The issue of suitable thresholds for administering PONV prophylaxis should be more widely discussed. Table 1 may provide a stepping stone for this discussion. One could argue that the risk of PONV for each patient should be reduced to about 30%. This would mean that patients with a baseline risk between 30% and 45% should receive one anti-emetic, patients with a baseline risk between 45% and 65% should receive two anti-emetics, and patients with a baseline risk above 65% should receive three anti-emetics. Patients with a baseline risk below 30% should not receive any intervention. Of course, other target risks can be considered, and the corresponding thresholds for anti-emetic prophylaxis can be read from the table. We believe that this stepwise policy for administering prophylaxis should make it possible to enjoy the benefits of prophylaxis at acceptable costs, whatever thresholds one prefers.

Table 1 The estimated risk of PONV in relation to the patient's baseline risk and possible interventions.

baseline risk[a]	estimated risk of PONV		
	1 anti-emetic	2 anti-emetics	3 anti-emetics
10%	7%	6%	4%
20%	15%	11%	8%
30%	22%	16%	12%
40%	30%	22%	16%
50%	37%	27%	20%
60%	44%	33%	24%
70%	52%	38%	28%
80%	59%	44%	32%
90%	67%	49%	36%

[a] The baseline risk of a patient may vary depending on the choice of the anesthetic technique, since intravenous anesthesia with propofol reduces the incidence of PONV compared to inhalation anesthesia[12;13]. The baseline risk with either anesthetic technique may be calculated with the model developed by Van den Bosch et al.[4], which includes anesthetic technique (inhalation anesthesia versus intravenous anesthesia) as a predictor, but not with the models developed by Apfel et al.[1] and Koivuranta et al.[2].

Simplified risk scores versus logistic regression formulas

Clinical prediction models can be presented in different formats. Two frequently used formats are logistic regression formulas and simplified risk scores. Currently, simplified risk scores seem to be most attractive for clinical practice because of their user-friendliness. However, one should be aware that their user-friendliness is relative. With the introduction of the electronic patient record (EPR), the more complex logistic regression formulas can easily be implemented in the computer software used for the EPR. Relevant predictor information can be retrieved automatically from the data fields in the EPR and the risk of the outcome can be directly presented on screen, even accompanied by treatment recommendations. This is of course more user-friendly than handmade calculations of PONV risk that require extra time. Implementation of logistic regression models in the EPR may have other advantages. In general, logistic regression models yield more precise risk estimates than risk scores, which may lead to better calibrated and better discriminating predictions. Furthermore, data from the EPR may be used to adjust a prediction model to keep up with (future) changes in anesthestic practice. In line with past developments in anesthetic practice, such as the introduction of prophylactic anti-emetics, new anesthetic agents (e.g. propofol and sevoflurane) and the gradual elimination of nitrous oxide, one may expect that anesthetic practice will continue to change. As a

consequence, the risks of PONV predicted with currently available models may not be correct for future populations and therefore recalibration of these models may be needed (chapters 3 and 6). In clinics where data on PONV predictors and outcomes are routinely collected, yearly model recalibration may be an interesting option. Model revisions and extensions (see discussion section chapter 3) may also be considered to keep up with (future) changes in anesthestic practice.

In summary, the relative merits of simplified risk scores (user-friendliness) may diminish over time. At present, most clinics are not yet equipped with an electronic patient record and therefore risk scores may still be preferred. Anticipating on technical progress, however, we recommend to use logistic regression formulas, if possible, as these generally provide risk estimates that are more precise. Furthermore, data collection with the EPR allows for model adjustments. This may be helpful in keeping up with (future) changes in anesthestic practice and establishing a quality cycle for continuous improvement of PONV care.

Directions for future PONV prediction research

When validated in a large sample of Dutch surgical inpatients, the predictive accuracy of the models developed by Apfel et al. and Koivuranta et al. was lower than expected based on the results from previous validation studies (chapter 3). When these models were validated in another Dutch hospital and a later time period, their predictive accuracy was more favorable (chapter 5). Although insight in the causes of compromised predictive accuracy is growing, in many cases it remains unclear which factors are responsible for a decreased predictive accuracy in a particular patient sample. Consequently, in case of poor validation results, it is not always clear how the validation process should be continued. The causes of compromised predictive accuracy should be further investigated.

One contribution of this thesis is the finding that one of the currently available prediction models for PONV, developed in an inpatient population, can be applied in outpatients, provided that the model's intercept is adjusted to account for the lower PONV incidence in outpatients (chapter 6). The presented adjusted model showed acceptable predictive accuracy in our outpatient sample. The robustness of the predictive accuracy of this adjusted model should be tested as part of future research, before the model is applied in ambulatory surgery.

In line with the discussion about the cumulative process of model development and validation, it may be interesting to develop one general

model to predict PONV in both in- and outpatients. Such a model should be developed on a combined sample of in- and outpatients and should include surgical setting (clinical versus ambulatory surgery) as a predictor. Interaction terms may need to be included to represent possibly different associations between predictors and outcomes. The combined development set should preferably include several data sets of in- and outpatients from different hospitals and countries. Hence, it would be interesting to pool the inpatient development samples of Apfel et al.[1] and Koivuranta et al.[2] and the inpatient validation sample described in chapter 5, with the outpatient validation sample of Sinclair et al.[3] and the sample described in chapter 6. Such a large pooled sample would enable more accurate estimations of the regression coefficients of the predictors and may yield prediction models that can be generalized across a wide(r) variety of populations and settings. Whether the use of more heterogeneous development samples will indeed improve the generalizability of prediction models for PONV remains to be defined. The boundaries to generalizing prediction models that are developed in heterogeneous samples therefore deserve particular attention in future research.

The directions outlined above may contribute to improved PONV prediction models. However, to improve the quality of anesthesia care, these prediction models must not only be adequately developed and validated. They must also be used by doctors, and if used, reduce PONV incidence. So far, two studies have investigated the impact of risk-based PONV prophylaxis on PONV incidence[14;15]. These studies yielded promising results, which should be further consolidated, as sample sizes were rather small [n=162[14] and n=428[15]]). It is yet uncertain whether the risk estimates obtained with prediction models for PONV do have an impact on the treatment decisions of anesthesiologists, when offered in a routine care setting instead of a study setting. The clinical impact of the available prediction models for PONV should therefore be investigated with formal impact analyses[16;17]. An impact analysis of the model developed by Van den Bosch et al.[4] is currently being performed in the University Medical Center of Utrecht, The Netherlands.

References

1. Apfel CC, Läärä E, Koivuranta M, Greim C-A, Roewer N. A simplified risk score for predicting postoperative nausea and vomiting: conclusions from cross-validations between two centers. *Anesthesiology* 1999;**91**:693-700.
2. Koivuranta M, Läärä E, Snåre L, Alahuhta S. A survey of postoperative nausea and vomiting. *Anaesthesia* 1997;**52**:443-9.
3. Sinclair DR, Chung F, Mezei G. Can postoperative nausea and vomiting be predicted? *Anesthesiology* 1999;**91**:109-18.
4. Van den Bosch JE, Moons KG, Bonsel GJ, Kalkman CJ. Does measurement of preoperative anxiety have added value for predicting postoperative nausea and vomiting? *Anesthesia and Analgesia* 2005;**100**:1525-32.
5. Apfel CC, Kranke P, Eberhart LHJ. Comparison of surgical site and patient's history with a simplified risk score for the prediction of postoperative nausea and vomiting. *Anaesthesia* 2004;**59**:1078-82.
6. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Annals of Internal Medicine* 1999;**130**:515-24.
7. Steyerberg EW, Borsboom GJJM, Van Houwelingen HC, Eijkemans MJC, Habbema JDF. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Statistics in Medicine* 2004;**23**:2567-86.
8. Apfel CC, Kranke P, Greim C-A, Roewer N. What can be expected from risk scores for predicting postoperative nausea and vomiting? *British Journal of Anaesthesia* 2001;**86**:822-7.
9. Iglesias del Sol A, Moons KGM, Hollander M, Hofman A, Koudstaal PJ, Grobbee DE et al. Is carotid intima-media thickness useful in cardiovascular risk assessment? The Rotterdam Study. *Stroke* 2001;**32**:1532-8.
10. Sweeney BP. Why does smoking protect against PONV? *British Journal of Anaesthesia* 2002;**89**:810-3.
11. Apfel CC, Kranke P, Katz MH, Goepfert C, Papenfuss T, Rauch S et al. Volatile anaesthetics may be the main cause of early but not delayed postoperative vomiting: a randomized controlled trial of factorial design. *British Journal of Anaesthesia* 2002;**88**:659-68.
12. Apfel CC, Korttila K, Abdalla M, Kerger H, Turan A, Vedder I et al. A factorial trial of six interventions for the prevention of postoperative nausea and vomiting. *The New England Journal of Medicine* 2004;**350**:2441-51.
13. Visser K, Hassink EA, Bonsel GJ, Moen J, Kalkman CJ. Randomized controlled trial of total intravenous anesthesia with propofol versus inhalation anesthesia with isoflurane-nitrous oxide: postoperative nausea and vomiting and economic analysis. *Anesthesiology* 2001;**95**:616-26.

14. Biedler A, Wermelt J, Kunitz O, Müller A, Wilhelm W, Dethling J et al. A risk adapted approach reduces the overall institutional incidence of postoperative nausea and vomiting. *Canadian Journal of Anesthesia* 2004;**51**:13-9.
15. Pierre S, Corno G, Benais H, Apfel CC. A risk score-dependent antiemetic approach effectively reduces postoperative nausea and vomiting - a continuous quality improvement initiative. *Canadian Journal of Anesthesia* 2004;**51**:320-5.
16. McGinn TG, Guyatt GH, Wyer PC, Naylor CD, Stiell IG, Richardson WS. Users' guides to the medical literature XXII: how to use articles about clinical decision rules. *JAMA: the journal of the American Medical Association* 2000;**284**:79-84.
17. Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. *Annals of Internal Medicine* 2006;**144**:201-9.

Summary

Chapter 1 – General Introduction

Postoperative nausea and vomiting (PONV) are still frequent and unpleasant side-effects of anesthesia. Therefore, cost-effective prevention of PONV is desirable. This may be achieved by selective administration of prophylactic anti-emetics or use of less emetogenic anesthetic agents with patients at high risk of PONV. Such high-risk patients can be identified preoperatively using a prediction model for PONV. The research described in this thesis was conducted to test the generalizability of available prediction models for PONV across populations, including outpatient populations, and to improve the predictive accuracy of these models.

Chapter 2 – Willingness to pay for PONV prevention

How much patients value prevention of PONV can be assessed by asking them how much they are willing to pay (WTP) out-of-pocket for an intervention that effectively prevents PONV. This chapter describes the assessment of patients' preoperative and postoperative WTP for PONV prevention as compared to prevention of postoperative pain, in 808 consecutive patients scheduled for surgery in a trial setting. Furthermore, it was determined whether postoperative experiences of nausea, vomiting and pain changed patients' WTP, when corrected for other factors that might elicit a WTP change. Preoperative WTP and individual WTP changes after surgery were analyzed in relation to baseline characteristics and postoperative experiences of PONV and pain. Although it appears from the literature that, to some patients, PONV is even more distressing than postoperative pain, PONV prevention was valued less (median preoperative WTP: US$ 17) than prevention of postoperative pain (median preoperative WTP: US$ 35). Also, experiences of severe postoperative pain increased patients' WTP for pain prevention, but experiences of PONV did not increase patients' WTP for prevention of PONV.

Chapter 3 – Validation of two prediction models for PONV

The predictive accuracy of a prediction model is generally lower in other patient samples than the sample from which the model was developed. The predictive accuracy of a model should therefore be tested in various patient samples (so called external validation studies), before the model is implemented in clinical practice. This chapter describes the validation of the two best-known prediction models for PONV in a large and heterogeneous

validation sample of Dutch inpatients. In this validation sample, the calibration and discrimination of the models were less than expected based on results from previous validation studies. Both models provided risk predictions that were too extreme, which means that low risk estimates were too low and high risk estimates were too high. The calibration slopes were 0.39 and 0.58 (ideal value is 1) and the areas under the receiver operating characteristic curve were 0.63 and 0.66 (range between 0.5 and 1; ideal value is 1), respectively. Neither of the models provided a risk threshold for anti-emetic prophylaxis that yielded satisfying results in terms of predictive values, sensitivity and specificity. This indicates that the models do not guarantee accurate prediction of PONV risk in other populations, which is important information, given the fact that one of the models was already implemented by several groups.

Chapter 4 – Development of a new prediction model for PONV

Because the two best-known prediction models for PONV showed disappointing validation results in a large Dutch validation sample of inpatients (chapter 3), a new model for PONV was developed using data from the Dutch validation sample. This model included six predictors: female gender, age, a history of PONV or motion sickness, smoking, the type of surgery and anesthetic technique. In the development sample, this model showed an area under the receiver operating characteristic (ROC) curve of 0.70 (after correction for overoptimism). Further, we quantified the added predictive value of preoperative anxiety to the predictors included in the newly developed model. Although there was a weak but significant association of anxiety with PONV, addition of anxiety to the model did not further increase the ROC area. This suggests that routine preoperative measurement of anxiety is not warranted, provided that the other predictors are already considered.

Chapter 5 – Which model to use for PONV prediction in surgical inpatients?

It was tested whether the predictive accuracy of the newly developed inpatient model (chapter 4) could be reproduced in a validation sample from another Dutch hospital and a later time period. For comparison, two previously developed models were also validated in this sample. The ROC area of the new model was 0.68 in the validation sample compared to 0.70 in the development sample. The ROC areas of the other two models were 0.69 and 0.70. The new model showed better calibration than the two previously developed models. None of the tested models outperformed the other models

both in terms of calibration and discrimination. Nonetheless, all models showed much higher predictive accuracy than a single predictor approach in which the type of surgery (ROC area 0.53) or previous occurrence of PONV (ROC area 0.58) is used to guide anti-emetic prophylaxis. It is therefore more important that the clinician actually uses one of the three available prediction models for PONV, than that the clinician selects a particular model.

Chapter 6 – Generalizability of inpatient prediction models for PONV to outpatients

Postoperative nausea and vomiting (PONV) are two of the most important concerns for patients after ambulatory surgery. Therefore, risk-based prophylaxis of PONV should not be limited to inpatients but be applied to outpatients as well. Available prediction models for PONV are, however, mainly developed in inpatients. It is yet unknown whether these models yield accurate predictions of PONV risk in outpatients. What is known is that the incidence of PONV in outpatients is lower than in inpatients, which may influence a model's predictive accuracy. We tested whether three prediction models for PONV, developed in inpatients, were generalizable to outpatients. Although all models showed limited discriminative ability in outpatients (ROC areas: 0.64, 0.62 and 0.64), one model showed reasonable calibration (calibration slope: 0.85). The intercept of this model was adjusted to account for the lower absolute PONV incidence in outpatients (25%) as compared to inpatients (48%). After adjustment of the intercept, this inpatient model yielded reasonably calibrated predictions of PONV risk in the outpatient validation sample. The ROC area of the model was 0.64 in outpatients compared to 0.68 in a sample of inpatients from the same hospital (chapter 5). For comparison, a model was developed in the outpatient validation sample. This model included the predictors of the adjusted inpatient model and regression coefficients that were estimated specifically for the outpatient population. The estimated regression coefficients in outpatients were similar to those in inpatients. Because of this, and because of the larger size of the development sample of the adjusted inpatient model (N=1389 versus N=623), the adjusted inpatient model may be preferred over the newly developed outpatient model, in spite of the higher ROC area of the latter model (0.66 versus 0.64). Of course, the robustness of the adjusted inpatient model should be tested in other outpatient samples, before this model is implemented in clinical practice.

Chapter 7 – Exploring boundaries to the generalizability of clinical prediction models

Chapter 7 explores boundaries to the generalizability of clinical prediction models. Clinical prediction models can ideally be generalized to a wide variety of patients. In reality, however, the predictive accuracy of a model is often decreased when applied to other patient samples than the sample in which the model was developed. Therefore, a newly developed prediction model should be tested in various (validation) samples, before it is implemented in clinical practice. Such validation samples are preferably different from but related to the development sample. Criteria for a ´different but related' sample are, however, not clearly defined. Chapter 7 explores this concept, illustrated with examples of prediction models that are developed and tested in different care settings (e.g. primary versus secondary care) or different age categories (e.g. adult versus pediatric patients). These examples show that the performance of prediction models decreases when validation samples are less related to the development sample, i.e. if the ranges of predictor values in the validation sample are wider than the ranges of predictor values in the development sample (e.g. generalizing a prediction model from secondary care to primary care), or if the ranges do not overlap at all (e.g. generalizing a prediction model across age ranges). It seems that prediction models generalize well to samples that are closely related, i.e. samples with ranges of predictor values that fall within the ranges of predictor values in the development sample. The chapter also discusses whether the generalizability of prediction models can be improved by using more heterogeneous development samples.

Chapter 8 – General Discussion

Chapter 8 discusses methodological issues related to PONV prediction research. The chapter comments on the relative value of calibration and discrimination for evaluating the predictive accuracy of prediction models for PONV. It also addresses the question how prediction models for PONV can be improved and what strategy should be followed for PONV prediction in outpatients. Further, chapter 8 describes clinical implications and outlines directions for future PONV prediction research.

Samenvatting

Hoofdstuk 1 – Introductie

Postoperatieve misselijkheid en braken (POMB) zijn nog steeds veel voorkomende en onplezierige bijwerkingen van de narcose. Kosteneffectieve preventie van POMB is dan ook wenselijk. Dit kan worden bewerkstelligd door selectieve toediening van profylactische anti-emetica en/of gebruik van minder emetogene anesthetica bij patiënten met een hoog risico op POMB. Dergelijke patiënten kunnen al voor de operatie worden geïdentificeerd met behulp van predictiemodellen voor POMB. Het onderzoek dat in dit proefschrift wordt beschreven is uitgevoerd om de generalizeerbaarheid van beschikbare predictiemodellen voor POMB te testen. Hierbij werd onder andere getest of de beschikbare modellen, die zijn ontwikkeld voor klinische patiënten, succesvol kunnen worden toegepast bij dagbehandelingspatiënten. Daarnaast beoogde dit onderzoek de prestaties van bestaande modellen te verbeteren.

Hoofdstuk 2 – Bereidheid om te betalen voor POMB preventie

Hoeveel waarde patiënten hechten aan preventie van POMB kan worden bepaald door hen te vragen welk geldbedrag zij uit eigen zak zouden willen betalen voor een middel dat POMB voorkomt. Dit hoofdstuk beschrijft de geldbedragen die patiënten voor en na de operatie bereid waren te betalen voor preventie van POMB. Hierbij werd een vergelijking gemaakt met de bedragen die deze patiënten wilden betalen voor preventie van postoperatieve pijn. Het in dit hoofdstuk beschreven onderzoek is uitgevoerd binnen de context van een klinische trial. Van de patiënten die deelnamen aan deze klinische trial werden achtereenvolgens 808 chirurgische patiënten voor ons onderzoek geïncludeerd. Bij deze patiënten hebben wij gemeten of postoperatieve ervaringen van misselijkheid, braken en pijn invloed hadden op de geldbedragen die patiënten wilden betalen. Bij de analyse van deze gegevens werd gecorrigeerd voor andere factoren die hierop van invloed zouden kunnen zijn. De geldbedragen die patiënten voor de operatie wilden betalen, en eventuele veranderingen hierin na de operatie, werden geanalyseerd in relatie tot patiëntkenmerken en postoperatieve ervaringen van misselijkheid, braken en pijn. Hoewel in de literatuur is beschreven dat POMB voor sommige patiënten meer ongerief veroorzaakt dan postoperatieve pijn, bleek dat patiënten minder waarde hechtten aan preventie van POMB dan aan preventie van postoperatieve pijn. Het mediane bedrag dat patiënten voor POMB preventie wilden betalen was US$ 17 versus US$ 35 voor preventie van postoperatieve pijn. Ook bleek dat het optreden van ernstige

postoperatieve pijn geassocieerd was met een toenemende bereidheid bij patiënten om voor preventie van postoperatieve pijn te betalen, terwijl het optreden van postoperatieve misselijkheid en braken geen invloed had op de waarde die patiënten hechtten aan preventie van POMB.

Hoofdstuk 3 – Validatie van twee predictiemodellen voor POMB

De prestatie van een predictiemodel neemt doorgaans af wanneer het model wordt toegepast bij andere patiënten dan de patiënten waarbij het model is ontwikkeld. De prestatie van een model moet daarom worden getest in verschillende patiëntgroepen (met behulp van zogenaamde externe validatie onderzoeken). Dit hoofdstuk beschrijft de validatie van de twee bekendste predictiemodellen voor POMB in een grote en heterogene validatiedataset met Nederlandse klinische patiënten. In deze validatiedataset was de calibratie en discriminatie van beide modellen minder goed dan te verwachten was op basis van resultaten van eerder uitgevoerde validatie onderzoeken. Voor beide modellen gold dat hoge kansschattingen te hoog waren geschat en lage kansschattingen te laag. De hellingen van de calibratiecurven waren 0.39 en 0.58 (ideale waarde is 1). De oppervlakten onder de ROC curven waren 0.63 en 0.66 (het bereik ligt tussen 0.5 en 1.0; de ideale waarde is 1). Voor geen van beide modellen kon een behandeldrempel voor profylaxe van POMB worden gedefinieerd die bevredigende resultaten opleverden in termen van voorspellende waarde, sensitiviteit en specificiteit. Dit betekent dat de voorspelde kansen op POMB mogelijk niet nauwkeurig zijn wanneer deze modellen in nieuwe patiëntgroepen worden toegepast. Dit is een belangrijke bevinding omdat één van deze modellen al door meerdere groepen in gebruik is genomen.

Hoofdstuk 4 – Ontwikkeling van een nieuw predictiemodel voor POMB

Het in hoofdstuk 3 beschreven onderzoek liet zien dat de validatieresultaten van de twee bekendste predictiemodellen voor POMB teleurstellend waren in een validatiedataset met Nederlandse klinische patiënten. Daarom werd een nieuw predictiemodel ontwikkeld met de data van de Nederlandse patiënten. Dit nieuwe model bevatte zes predictoren: geslacht, leeftijd, roken, voorgeschiedenis van misselijkheid en braken na eerdere operatie of bekendheid met wagenziekte, type geplande chirurgische ingreep en anesthesietechniek. In de dataset waarmee dit nieuwe model werd ontwikkeld was de oppervlakte onder de ROC curve 0.70 (na correctie voor overoptimisme). Verder werd in dit onderzoek bepaald of preoperatieve angst

voorspellende waarde toevoegt aan de predictoren in het ontwikkelde model. Er bleek een zwakke maar significante relatie te zijn tussen de variabele angst en POMB. Toevoeging van deze variabele aan het ontwikkelde model leidde echter niet tot een verhoging van de oppervlakte onder de ROC curve. Routinematige metingen van preoperatieve angst zijn dan ook niet geïndiceerd als de predictoren in het ontwikkelde model reeds in beschouwing zijn genomen.

Hoofdstuk 5 – Welk predictiemodel is het meest geschikt voor het voorspellen van POMB bij klinische patiënten?

Dit onderzoek werd uitgevoerd om te testen of de prestaties van het nieuw ontwikkelde model voor POMB, zoals gevonden in de Nederlandse ontwikkelingsdataset, konden worden gereproduceerd in een validatiedataset uit een ander Nederlands ziekenhuis en een andere tijdsperiode. Ter vergelijking werden de twee eerder ontwikkelde predictiemodellen ook gevalideerd in deze dataset. De oppervlakte onder de ROC curve van het door ons ontwikkelde model was 0.68 in de validatiedataset versus 0.70 in de ontwikkelingsdataset. Voor de andere modellen was de oppervlakte onder de ROC curve 0.69 en 0.70. Het nieuw ontwikkelde predictiemodel was in de validatiedataset enigszins beter gecalibreerd dan de andere twee modellen. Helaas presteerde geen van de drie predictiemodellen duidelijk beter dan de andere. Wel presteerden de modellen alledrie veel beter dan een methode waarbij slechts één predictor (bijvoorbeeld: het type geplande chirurgische ingreep [oppervlakte onder de ROC curve = 0.53] of een eventuele voorgeschiedenis van misselijkheid en braken na een eerdere operatie [oppervlakte onder de ROC curve = 0.58]) wordt gebruikt om te bepalen of een patiënt in aanmerking komt voor anti-emetische profylaxe. Het is daarom van groter belang dat een clinicus tenminste één van de drie beschikbare predictiemodellen voor POMB gebruikt dan dat een clinicus kiest voor een bepaald model.

Hoofdstuk 6 – Kunnen predictiemodellen voor klinische patiënten worden gegeneraliseerd naar dagbehandelingspatiënten?

Postoperatieve misselijkheid en braken behoren tot de meest belangrijke zorgen die patiënten hebben ten aanzien van de periode na een ingreep in dagbehandeling. Daarom zouden predictiemodellen voor POMB niet alleen moeten worden gebruikt om te bepalen welke klinische patiënten in aanmerking komen voor profylaxe van POMB. Ook bij dagbehandelingspatiënten zouden

deze modellen moeten worden toegepast. Tot nu toe zijn predictiemodellen voor POMB echter voornamelijk ontwikkeld bij klinische patiënten. Het is niet bekend hoe deze modellen presteren bij dagbehandelingspatiënten. Wat wel bekend is, is dat de incidentie van POMB bij dagbehandelingspatiënten lager is dan bij klinische patiënten, wat van invloed zou kunnen zijn op de calibratie van de kansschattingen. In dit onderzoek werd getest of drie predictiemodellen voor POMB, alle ontwikkeld in klinische patiënten, kunnen worden gegeneralizeerd naar dagbehandelingspatiënten. Hoewel de modellen alledrie een beperkt discriminerend vermogen hadden bij dagbehandelingspatiënten (oppervlakten onder de ROC curven: 0.64, 0.62 en 0.64) was één van de drie modellen redelijk gecalibreerd (helling van de calibratiecurve: 0.85). Het intercept van dit model werd aangepast om het verschil in POMB incidentie tussen klinische patiënten (25%) en dagbehandelingspatiënten (48%) te verdisconteren. Na aanpassing van het intercept, waren de kansschattingen van het aangepaste model bij dagbehandelingspatiënten vrij goed gecalibreerd. De oppervlakte onder de ROC curve bij dagbehandelingspatiënten was 0.64 versus 0.68 bij klinische patiënten uit hetzelfde ziekenhuis (hoofdstuk 5). Ter vergelijking werd een model ontwikkeld met de data van de dagbehandelingspatiënten. Dit model bevatte dezelfde predictoren als het aangepaste model (dat was ontwikkeld bij klinische patiënten) en regressiecoëfficiënten die werden geschat met de data van de dagbehandelingspatiënten. De bij de dagbehandelingspatiënten geschatte regressiecoëfficiënten kwamen overeen met de regressiecoëfficiënten zoals die bij de klinische patiënten waren geschat. Daarnaast was de ontwikkelingsdataset van het aangepaste model groter dan de dataset met dagbehandelingspatiënten (N=1389 versus N=623). Dit zouden aanleidingen kunnen zijn om de voorkeur te geven aan het aangepaste model (dat was ontwikkeld bij klinische patiënten) boven het specifiek voor dagbehandelingspatiënten ontwikkelde model, ondanks het feit dat de oppervlakte onder de ROC curve voor dit laatste model iets hoger was (0.66 versus 0.64). De robuustheid van het aangepaste model moet uiteraard nog worden getest in andere groepen dagbehandelingspatiënten voordat dit model kan worden geïmplementeerd in de klinische praktijk.

Hoofdstuk 7 – Generalizeerbaarheid van POMB predictiemodellen: grensverkenningen

Hoofdstuk 7 verkent grenzen aan de generalizeerbaarheid van klinische predictiemodellen. In het ideale geval kunnen predictiemodellen worden gegeneralizeerd naar een variëteit aan patiëntgroepen. In de praktijk blijken

de prestaties van dergelijke modellen echter vaak af te nemen wanneer deze worden toegepast bij andere patiëntgroepen dan de patiëntgroep waarmee het model ontwikkeld was. Daarom moeten ontwikkelde modellen eerst worden getest in verschillende (validatie)datasets, voordat zij worden geïmplementeerd in de klinische praktijk. Dergelijke validatiedatasets zouden enerzijds verschillend moeten zijn van de ontwikkelingsdataset, maar anderzijds ook hieraan verwant. Tot op heden zijn echter nog geen heldere criteria geformuleerd om te bepalen wanneer een validatiedataset verschillend is van, maar tegelijkertijd ook verwant is aan de ontwikkelingsdataset. Hoofdstuk 7 verkent dit concept. Hierbij wordt gebruik gemaakt van voorbeelden van predictiemodellen die werden gevalideerd in andere zorgcontexten (bijvoorbeeld eerste lijn versus tweede lijn van de gezondheidszorg) of andere leeftijdsgroepen (bijvoorbeeld volwassenen versus kinderen) dan die waarin zij waren ontwikkeld. Deze voorbeelden laten zien dat de prestaties van een predictiemodel afnemen wanneer de validatiedataset minder verwant is aan de ontwikkelingsdataset, d.w.z. wanneer het bereik van de predictorwaarden in de validatiedataset (bijvoorbeeld afkomstig uit de eerste lijn) breder is dan het bereik van deze waarden in de ontwikkelingsdataset (bijvoorbeeld afkomstig uit de tweede lijn), of wanneer de predictorwaarden in de validatiedataset en de ontwikkelingsdataset in het geheel geen overlap vertonen (bijvoorbeeld leeftijd bij volwassenen versus kinderen). Het lijkt erop dat predictiemodellen beter generalizeerbaar zijn naar patiëntgroepen die nauw verwant zijn aan de ontwikkelingsdataset, d.w.z. dat de predictorwaarden in de validatiedataset vallen binnen het bereik van de predictorwaarden in de ontwikkelingsdataset. Tot slot beschouwt dit hoofdstuk de mogelijkheid om de generalizeerbaarheid van modellen te vergroten door gebruik te maken van heterogenere ontwikkelingsdatasets.

Hoofdstuk 8 – Algemene beschouwing

In hoofdstuk 8 wordt aandacht geschonken aan enkele methodologische kwesties op het gebied van POMB predictie-onderzoek. Hierbij wordt een korte beschouwing gewijd aan de waarde van calibratie en discriminatie voor het evalueren van de prestaties van predictiemodellen voor POMB. Ook wordt aangegeven hoe bestaande predictiemodellen voor POMB zouden kunnen worden verbeterd. Tevens wordt beschreven welke strategieën mogelijk zijn om tot een predictiemodel voor POMB bij dagbehandelingspatiënten te komen. Verder gaat dit hoofdstuk in op klinische implicaties. Het hoofdstuk sluit af met een schets van perspectieven voor toekomstig onderzoek met betrekking tot predictie van POMB.

Dankwoord

Prof. dr. Kalkman, Cor: Dank dat je mij als promovenda hebt willen aanstellen op dit promotie-onderzoek. De nadere kennismaking met de medische praktijk heb ik als zeer waardevol ervaren.

Prof. dr. Moons, Carl: Dank voor je inspanningen om van ons werk en samenwerking een succes te maken. Ik hoop dat de predictieclub onder jouw handen mag uitgroeien tot het 'predictiehart' van Nederland.

Dr. Van Klei, Wilton: integer, betrokken, verantwoordelijk.

Dr. Vergouwe, Yvonne: dank voor je inhoudelijke bijdrage aan dit proefschrift en de predictieclub. Zoals Corné al in het dankwoord van zijn proefschrift schreef: je invloed is groter geweest dan je zelf denkt. Het is moeilijk te zeggen wat ik in jou het meest ben gaan waarderen: je vakbekwaamheid, je communicatieve vaardigheid, je correctheid of je oprechte en bescheiden aanwezigheid.

Prof. dr. Bonsel, Gouke: dank voor de enthousiasmerende intellectuele impulsen. Het was mij een genoegen.

Prof. dr. Grobbee, Diederick: dank voor je intellectuele input en bijdrage aan hoofdstuk 3 van dit proefschrift.

Leden van de PoPiMiB-stuurgroep, Pim Hennis en Ingrid van Dijk: dank dat ook jullie op deze wijze wilden bijdragen aan de totstandkoming van dit proefschrift.

Leo van Wolfswinkel en Edwin Hammink: dank voor de moeite en tijdsinvestering waarmee jullie de logistieke weg van de PoPiMiB-studie hebben geplaveid. Tevens gaat mijn dank uit naar Martijn Estor en Teus Kappen voor de technische (nood)hulpverlening bij IT hobbels op de weg.

Dames van de POS-poli en teamleiders, Jeroen Moen en José Schipper-Kleijn: hartelijk dank voor het faciliteren van de dataverzameling.

Mijn bijzondere dank gaat uit naar het verpleegkundige personeel van recovery F4 en F2 en naar de research verpleegkundigen van het trialbureau van de DP&S (Ada, Marcel, Margaret en Selma). Wat hebben jullie veel patiënten bevraagd. Jullie bijdrage blijft vaak onzichtbaar, maar eerlijk is eerlijk: een groot deel van het werk is door jullie verzet. Heel veel dank.

Graag wil ik hierbij ook Marjanne Lammersma, Jacqueline van Dijk, Bregt Michel, Willemijn Vos, Jeanne Beekes en Heleen Van Koeven noemen vanwege hun inspanningen bij het op gang houden van de dataverzameling en de verwerking van de gegevens.

Aan de verzameling en verwerking van de gegevens hebben ook studenten een niet geringe bijdrage geleverd: Renée van der Pal, Tobias Boerboom, Debbie Kalkman, Sonja en Linda van Bunnik, Manon Bruens, Janneke Jaspers en Annelies Bakker. Ook jullie wil ik hartelijk danken voor jullie inspanningen.

Annelies Sterkenburg, Tessa Ledderhof en Marcel Out van GlaxoSmithKline BV: met genoegen kijk ik terug op een zeer prettige samenwerking!

Prof. dr. Djolicz en secretaresse: dank voor de medewerking en gastvrijheid ten tijde van het statusonderzoek in de catacomben van het AMC.

(Voormalige) leden van de predictieclub: Corné, Kristel, Eva, Diane, Peter, Lidewij en Martijn: ik hoop dat het tempo waarin de predictieclub zich tijdens mijn promotietijd heeft ontwikkeld een hoge voorspellende waarde heeft voor het toekomstige succes van deze club.

Secretaresses van de DP&S en Paula van Rodenburg: dank voor de ondersteuning in de kleine maar zeker niet onbelangrijke dingen.

Dames en heren van het bedrijfsbureau en andere (tijdelijke) bevolkers van gang E.02: dank voor de informatie over het reilen en zeilen in het UMCU, de gezellige integratie en natuurlijk, niet te vergeten, voor jullie belangstelling voor mijn onderzoek.

Mijn bijzondere dank gaat tenslotte uit naar mijn kamergenoot Lidwien: “…grmmpff …” “… alles in het Grieks … ?“. Lidwien, wat moet ik zeggen anders dan dank, dank en nogmaals dank voor je steun en behulpzame, verdraagzame aanwezigheid. Inmiddels blijf je gelukkig niet meer alleen achter op de eens zo eenzame gang der onderzoekers. Heb je desondanks toch nog belangstelling voor de goudvis en de plant … ?

Graag wil ik ook mijn dank betuigen aan hen die in het volbrengen van dit promotie-onderzoek een bijzondere rol hebben gespeeld:

Lieve ouders, allerliefst zusje: in rampspoed en gevaar staan jullie altijd voor mij klaar. Mijn dank daarvoor.

Maatje: dank dat jij altijd in mij bent blijven geloven …

Curriculum Vitae

Jolanda van den Bosch was born on September 25, 1978 in Giessenburg, The Netherlands. In 1996, she graduated grammar school with distinction at the Gymnasium Camphusianum in Gorinchem. In 2001, she obtained a M.Sc. degree in biomedical sciences with distinction at the University of Utrecht. Parallel to her studies in biomedical sciences she studied philosophy of biomedical sciences at the Faculty of Philosophy of the University of Utrecht. The curriculum included several courses in biomedical ethics, philosophy of science, history and philosophy of biology and philosophy of neuroscience.

As part of her training as a biomedical scientist she conducted three research projects. First, she worked with a rat model of spinal cord injury at the Rudolf Magnus Institute for Neurosciences in Utrecht. Second, she worked for the animal experiments committee of Solvay Pharmaceuticals BV in Weesp, on a project about the interpretation of legally obliged animal experiments. This project was conducted in collaboration with the Center for Bioethics and Health Law (CBG) in Utrecht. Third, she conducted research at the Department of Molecular Immunopathology of the University Medical Center of Utrecht. This research project focussed on cardiac matrix remodelling after myocardial infarction and was completed with a thesis on the promise of stem cell transplantation for post-infarction cardiac repair.

In 2002, she started the research work described in this dissertation at the Department of Perioperative care and Emergency Medicine of the University Medical Center of Utrecht (promotor: prof. dr. C.J. Kalkman, anesthesiologist), and the Julius Center for Health Sciences and Primary Care in Utrecht (promotor: prof. dr. K.G.M. Moons, clinical epidemiologist). During this period she received training as a clinical epidemiologist at the Netherlands Institute for Health Sciences in Rotterdam and obtained a M.Sc. degree in clinical epidemiology in 2003. Since July 2005, she conducts benchmark research in the field of disaster medicine at the NIVEL in Utrecht.